



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 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 visualization modules of VIQoR 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. Furthermore, filtering according to the factor analysis signal-to-noise ratios (S/N) have been shown to reduce the false protein quantification rates (FQR) [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 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 the 'Input' tab the user can import the data required for the analysis, inspect them in interactive tables and manually validate them.

The image shows a dark-themed sidebar for the 'Input' tab. It is divided into two main sections: 'CSV FILE' and 'FASTA FILE'. The 'CSV FILE' section includes a 'Choose CSV file:' label, a 'Browse...' button, and a 'No file selected' status with a red circle containing the number 1. Below this is an 'Example dataset' link with a red circle containing the number 3. The 'Choose CSV file settings:' section includes a 'Header:' label with a red circle containing the number 4, a checked 'Header?' checkbox, a 'Separator:' label with a red circle containing the number 5, and radio button options for ';', '.', ',', and 'tab'. The 'Decimal separator:' label has a red circle containing the number 6 and radio button options for '.', ',', and ' '. The 'Data transformation:' label has a red circle containing the number 7 and a 'Data log transformed?' checkbox. The 'FASTA FILE' section includes a 'Choose FASTA file:' label and a 'Browse...' button, with a 'No file selected' status and a red circle containing the number 2.

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 imported data can be of either labeled or label-free quantification experiments and acquired by DDA (data-dependent acquisition) or DIA (data-independent acquisition) mode. 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 residue 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 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. To utilize all the visualization and optimization modules in VIQoR's interface, the

conditions of replicated experiments should have the same number of replicates. Nonetheless, any experimental design is accepted. 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. An example .csv file can be found [here](#).

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

Hint: the imported peptide/PSM reports should be already filtered according to the user's preferences and based on the experiment type (FDR, number of missed cleavages etc.) but should not be preprocessed regarding protein inference. Some of the software that can identify spectra, extract/assign intensities and export peptide/PSM reports are: [SearchGUI \(Tutorial\)](#), [MaxQuant \(Documentation\)](#) and [Proteome Discoverer \(Manual\)](#).

2 Click 'Browse' to import the protein sequence database. Preferably the one that have been used for the peptide identification 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 for protein inference. An example .fasta file can be found [here](#).

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 corresponding to E. coli protein [P0A6Y8](#) 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 of the example dataset are provided in all the 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.

- 7 Check the box if the peptide intensities are already log2-transformed.

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).

The figure shows two interactive tables within the 'Input' tab. The top table is titled 'Peptide abundance table' and the bottom table is titled 'Fasta file table'. Both tables have a search bar and a 'Show 10 entries' dropdown.

Peptide abundance table:

	sequence	intensity.in.HYE110_B.1	intensity.in.HYE110_B.2	intensity.in.HYE110_B.3	intensity.in.HYE110_A.1	intensity.in.HYE110_A.2	intensity.in.HYE110_A.3
1	VELLGTSIAECTLYLDNGVVFVGSR					31399	
2	ALQLLDEVLTMPADPQPLD	45709	48502	53271			
3	QLTEMLPSILNQLGADSLTSLR						132512

Fasta file table:

	Accession	Annotation	Sequence
1	ADA024R1R8	ADA024R1R8_HUMAN HCG2014768, isoform CRA_a	MSSHEGGKKKALKQPKKQAKEMDEEEKAFKQKQKEEQKKLEVLKAKVVGKPLATGGIKKSGKK
2	ADA024RBG1	NUD4B_HUMAN Diphosphoinositol polyphosphate phosphohydrolase NUDT4B	MMKFKPNQTRTYDREGFKKRAACLCFRSEQEDEVLLVSSRYPDQWIPGGGMEPEEPPGGAHREYEEAGVKGKLRLLGIFEQHQQRKHRTYYVLTVEILEDWEDSVNIG
3	ADA075B6H5	ADA075B6H5_HUMAN T cell receptor beta variable 20/ORB-2 (non-functional) (Fragment)	NETWITLIPREGGVGPSRKMLLLLLLPQSGLSAWSQHPSPRVCKSGTSVNECRSLDFQATTNFWROLRKQSLHLMATSNESSEVTYEQGWKKFPIHPNLTFSALTVP

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

To continue to 'Conditions and Samples' tab click [Next](#)

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 the '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.

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.

The image shows a dark-themed sidebar titled "CONDITIONS AND SAMPLES". It contains five numbered settings, each with a label, a help icon, a control, and a number in a blue circle. 1. "Number of quantitative columns: ?" with a slider from 4 to 7, set at 6. 2. "Number of conditions: ?" with a slider from 1 to 6, set at 2. 3. "Reference condition: ?" with a checked checkbox "Reference condition?" and a slider from 1 to 2, set at 1. 4. "Column arrangement: ?" with an unchecked checkbox "Run based arrangement?". 5. "Maximum NA values per peptide: ?" with a slider from 1 to 4, set at 1.

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

Note: in case of experiments with unequal number of replicates per condition, set the number of conditions to be equal to the total number of quantitative columns. The protein inference and protein summarization processes in VIQoR are independent of the experimental design, however some of the visualization and optimization modules (GCI optimization module and per condition summary in 'Protein Quantification' tab) are not available for such experiments.

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 selected condition's samples. 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.

Conditions and samples table							
Show 10 entries		Search:					
sequence	intensity.in.HYE110_B.1	intensity.in.HYE110_B.2	intensity.in.HYE110_B.3	intensity.in.HYE110_A.1	intensity.in.HYE110_A.2	intensity.in.HYE110_A.3	
1 AAAAAAAAAAPAAATAPTTAATTAATAAQ	620427	871877	689264	480466	456745	433096	
2 AAADALSDLELKDSK	272953	104880	289138	1271460	1329684	1659389	
A AAASSSLQWK	191936	144295	128754	18269	16872	14567	

Conditions and samples table							
Show 10 entries		Search:					
sequence	intensity.in.HYE110_B.1	intensity.in.HYE110_B.2	intensity.in.HYE110_B.3	intensity.in.HYE110_A.1	intensity.in.HYE110_A.2	intensity.in.HYE110_A.3	
1 AAAAAAAAAAPAAATAPTTAATTAATAAQ	620427	871877	689264	480466	456745	433096	
2 AAADALSDLELKDSK	272953	104880	289138	1271460	1329684	1659389	
B AAASSSLQWK	191936	144295	128754	18269	16872	14567	

Conditions and samples table							
Show 10 entries		Search:					
sequence	intensity.in.HYE110_B.1	intensity.in.HYE110_B.2	intensity.in.HYE110_B.3	intensity.in.HYE110_A.1	intensity.in.HYE110_A.2	intensity.in.HYE110_A.3	
1 AAAAAAAAAAPAAATAPTTAATTAATAAQ	620427	871877	689264	480466	456745	433096	
2 AAADALSDLELKDSK	272953	104880	289138	1271460	1329684	1659389	
C AAASSSLQWK	191936	144295	128754	18269	16872	14567	

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).

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.

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 the 'Modifications and Normalization' tab, the user can select the types of modifications to be separated for PTMs visualization analysis, perform log2-transformation and per sample zero-center normalization of the peptide abundances. Additionally, due to identical molecular mass, the Isoleucine residues (I) are substituted by Leucine residues (L) in all the unmodified sequences to ease the protein inference process. Modified peptide sequences remain intact, and the residue substitution is applied only for visualization purposes in VIQoR plot.

1 Check the box(es) of the modification type(s) that should proceed for 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.

MODIFICATIONS

Choose modifications to quantify: 1

☒ Acetylation

☐ Phosphorylation

Choose!

NORMALIZATION

Normalization method: 2

☐ average ☒ median ☐ quantile

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.


Sequence	Counterpart	Modification	Position	Intensity.in.HYE110_B.1	Intensity.in.HYE110_B.2	Intensity.in.HYE110_B.3	Intensity.in.HYE110_A.1	Intensity.in.HYE110_A.2	Intensity.in.HYE110_A.3
1 LLNVLVEEF(Acetylation)K	LLNVLVEEFK	Acetylation	10	24373	24227	21151	359321		
2 NTTLPT(Acetylation)K	NTTLPTK	Acetylation	7	55955	67227	53024	429923		
A QVEEAGDI(Acetylation)K	QVEEAGDIK	Acetylation	8	48427	45302	47640	448278		

sequence	Intensity.in.HYE110_B.1	Intensity.in.HYE110_B.2	Intensity.in.HYE110_B.3	Intensity.in.HYE110_A.1	Intensity.in.HYE110_A.2	Intensity.in.HYE110_A.3
1 AAAAAAAAAAPAAATPTTAATTAATAAQ	3.0226747887186	3.51119657484729	3.19552326154967	2.58730049439188	2.52256181361706	2.43360645599063
2 AADALSDLELKIDSK	1.83806587086915	0.455811544286227	1.94222482334911	3.99128032215804	4.06418440212271	4.3714998306495
B AAEESSIQIK	1.53003864969807	0.915093277348413	0.775083848879547	-2.13037517799883	-2.33813421674322	-2.4853044946467354

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

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.

2 Select the normalization method. The log2-transformed peptide abundances can be zero-center normalized by the average, median or quantile of each sample. Alternation of the method will update the normalization 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 PTM 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  Download button 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 the 'Protein Inference' tab the user can assemble proteins from the peptides of the imported dataset, or in other words, to perform protein identification. First, 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 proteins having overlapping peptide sets. Therefore, a protein group can contain more than one protein accession. The single-peptide protein identifications are excluded to reduce the false discovery rate (FDR) [5,6].

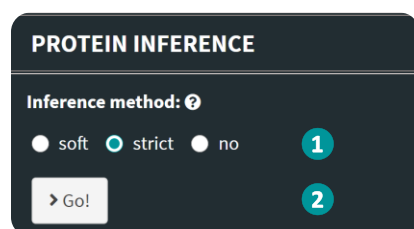


Figure 7: Sidebar of the 'Protein Inference' tab.

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 largest set of peptides, until the presence of all peptides in the component 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.

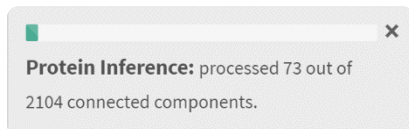
The *Soft* parsimony, groups together two proteins if the peptides related to the latter is a subset of the peptide set related to the former. In addition, the degenerate peptides assigned already to a protein group can be shared with all the following protein groups that are related to these peptides. 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.



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

Figure 8: Progress bar during protein inference analysis.

Once the protein inference is completed the list with the inferred protein groups 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 (/).

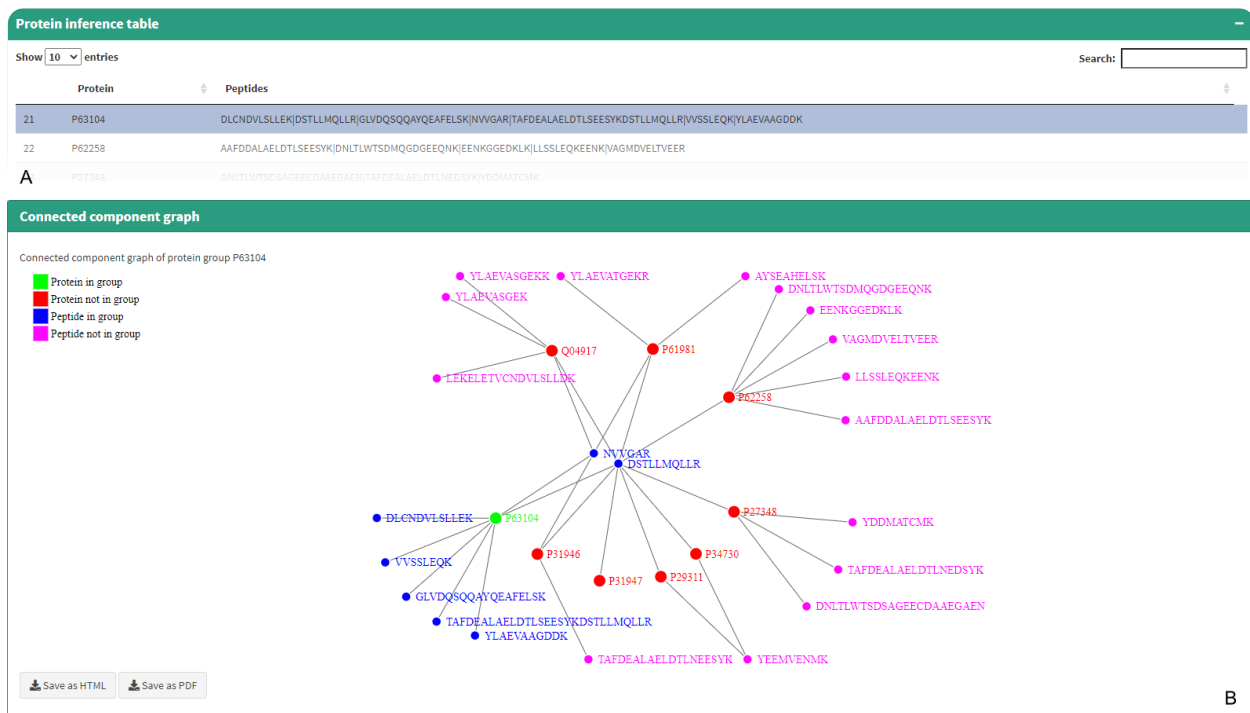





Figure 9: The 'Protein inference table' and the 'Connected component graph' in the body of the 'Protein Inference' tab. Example for the protein group P63104.

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 alphabetical 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 to *'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 assess their quality [1,2]. The analysis assigns individual weights to the peptides, which denote their coherence to the protein group 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 to protein expressions.

- 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.

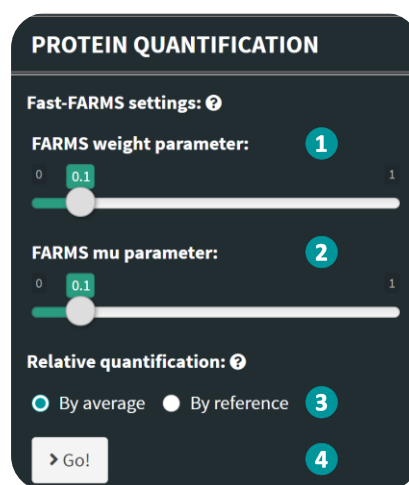


Figure 10: Sidebar of the 'Protein Quantification' tab.

- 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 of all samples.

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 condition (average of the condition 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 sub-tabs provide the same data inspection and visualization modules. The 'Per sample' tab additionally provides a slider to enable the alternation of the peptide weight threshold and the Global Correlation Index (GCI) module to optimize the value of the threshold (*Figure 11B*). This manual for reasons of brevity, presents and explains the modules provided only in the 'Per sample' tab.

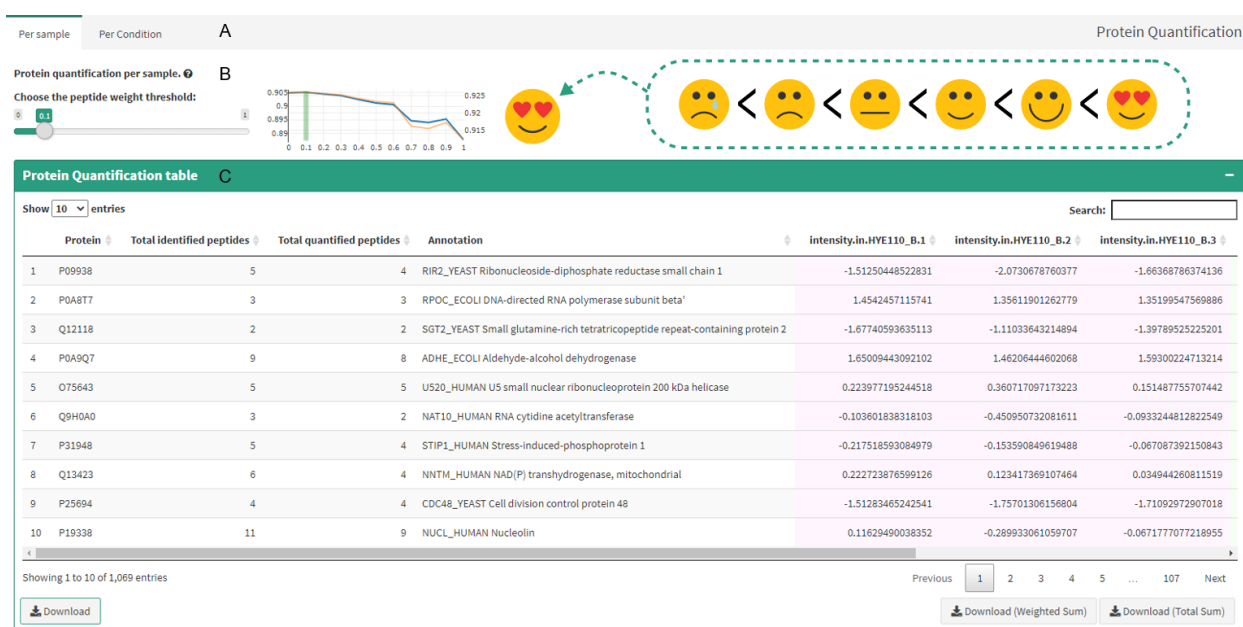


Figure 11: Body of the 'Protein Quantification' tab. GCI (B) and 'Protein Quantification table' (C).

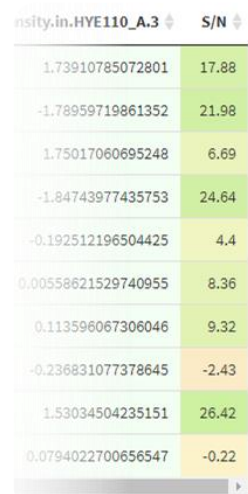
Global Correlation Index (GCI)

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) (*Figure 11B*), 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 (Blue trace) in the CGI module, results in the optimal protein summarization. Additionally, a series of smiley icons are provided to visually ease the parameter tuning. 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' sub-tabs according to the selected value.

Protein Expressions


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, 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*), the protein expressions for all the samples/conditions (*Figure 11C*) and the signal-to-noise ratio (S/N) of the factor analysis in dB (*Figure 12*). High S/N values correspond to informative protein groups, while very low values to noisy groups. The authors of fast-FARMS demonstrated that a minimum threshold of -20 dB improves the total protein FQR [2].



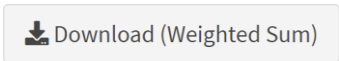
Density.in.HYE110_A.3	S/N
1.73910785072801	17.88
-1.78959719861352	21.98
1.75017060695248	6.69
-1.84743977435753	24.64
-0.192512196504425	4.4
0.00558621529740955	8.36
0.113596067306046	9.32
-0.236831077378645	-2.43
1.53034504235151	26.42
0.0794022700656547	-0.22

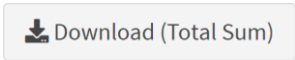
Figure 12: Signal-to-noise ratios column at the 'Protein Quantification table'

To download the 'Protein Quantification table' in .csv format click the button

 Download

As an addition to the relative protein expressions displayed in *'Protein Quantification table'*, VIQoR provides two more protein summarization approaches. The first one (referred to as *'Weighted Sum'*) utilizes the peptide weights of the factor analysis and summarizes the protein abundances as a weighted sum of the corresponding peptide intensities. The second one (referred to as *'Total Sum'*) summarizes the protein expressions as the total summation of the corresponding peptide intensities. The protein expressions for both the additional methods are then log2-transformed and normalized by the normalization method selected in *'Modifications and Normalization'* tab. The interactive modules in *'Protein Quantification'* tab however display and visualize only the relative protein expressions.

To download the protein expressions summarized by the *'Weighted Sum'* approach in .csv format click the button 

To download the protein expressions summarized by the *'Total Sum'* approach in .csv format click the button 

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.

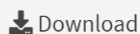
Hint: the data can be sorted by a specific column. Click the identifier of a column of interest once or twice to sort in increasing or decreasing order.

Protein expression correlation heatmap

The *'Protein expression correlation heatmap'* is located under the *'Protein Quantification table'* (Figure 13). 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 (from minimum of 1 cluster to maximum number of clusters 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



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 13: 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 14). 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, while their opacity depends on the assigned peptide weights. The eliminated peptides are colored grey and illustrated with dashed lines. For instance, in *Figure 14* the line plot of the selected protein group POA6Y8 is presented for a peptide weight threshold of 0.1. The peptide with sequence AVLTVPAYFNDAQR that is following an opposite expression pattern 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

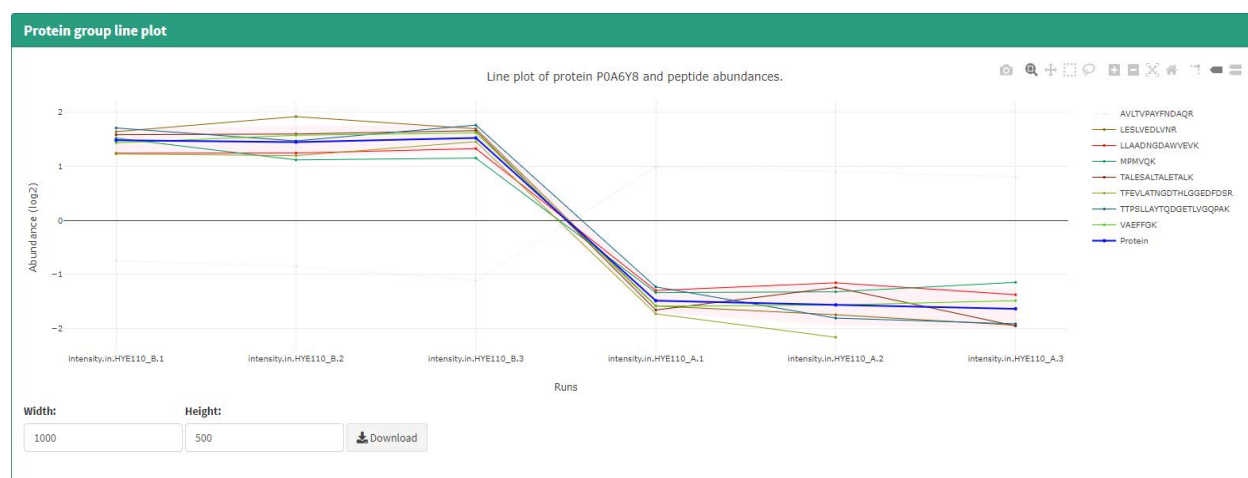


Figure 14: The 'Protein group line plot' in the 'Protein Quantification' tab.

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 (Reference vs Testing). 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 protein group (blue segment). Additionally, the individual PTMs are annotated on the peptide segments according to their position 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 can also identify missed cleaved peptides according to specific enzyme digestion rules and highlight them with different color (orange). The supported enzymes are: Trypsin, Trypsin strict (P at C-term to K or R is not an exception for cleavage), Chymotrypsin (High or Low specificity), Pepsin (pH 1.3 or pH > 2), Lys-C and Arg-C. 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 15* 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 AVLTPAYFNDAQR, 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. Additionally, by testing the digestion efficiency on the tryptic peptides of protein group P0A6Y8, no missed cleavages are

identified, unlike the protein group P0A9Q7 in *Figure 17* where two peptides contain missed cleavage sites (orange segments).

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

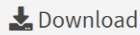


Figure 15: The ‘Protein group VIQoR plot’ in the ‘Protein Quantification’ tab.

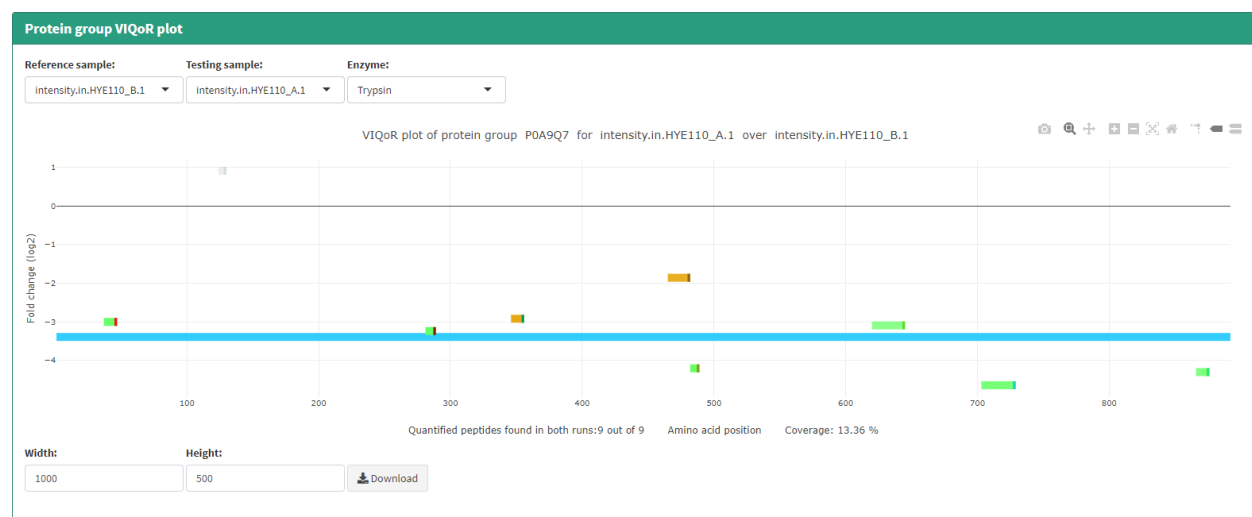


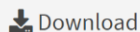
Figure 16: The ‘Protein group VIQoR plot’ – demonstration of missed cleaved peptide identification

Peptide abundance correlation heatmap

The ‘*Peptide abundance correlation heatmap*’ for the selected protein group unfolds under the ‘Protein group VIQoR plot’ (*Figure 17*). 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 17 illustrates the heatmap of the peptide profile for the protein group POA6Y8 when peptide weight threshold is set to 0. The peptide with sequence AVLTPAYFNDAQR 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



Note: the selected POA6Y8 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 17: 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.
5. Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A; Working Group on Publication Guidelines for Peptide and Protein Identification Data. The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol Cell Proteomics*. 2004 Jun;3(6):531-3.
6. Bradshaw RA, Burlingame AL, Carr S, Aebersold R. Reporting protein identification data: the next generation of guidelines. *Mol Cell Proteomics*. 2006 May;5(5):787-8.