User Guide for LFAQ

Version 1.0.0

1 Introduction

LFAQ is an efficient tool for label-free absolute protein quantification based on a novel concept *peptide quantification efficiency*, which is learned by a semi-supervised machine learning method. Additionally, LFAQ provides user-friendly interfaces for parameter setting, quantification analysis and result visualization. The core of LFAQ was written in standard C++ language on the platform of Microsoft Visual Studio ultimate 2013 in Windows System. The graphical user interface (GUI) of LFAQ was implemented in C# language on the same platform. LFAQ is freely available at https://LFAQ.github.io/LFAQ/.

2 Running LFAQ

Now LFAQ can be run on Windows system smoothly, requiring the .net framework version 4.5 or above and R (version 3.2.5 and above) software to be installed in advance. The .net framework can be downloaded from: https://www.microsoft.com/en-us/download/details.aspx?id=30653. R software can be downloaded from https://www.r-project.org/. The directory which includes RScript.exe is required to be added into the system environment variable. The method for setting system environment variable can be found at http://www.computerhope.com/issues/ch000549.htm. Since the R package “ggplot2” is used in the visualization module of LFAQ, the “ggplot2” package should be installed in advance.

After completing the preparation work, the user can double-click “LFAQ.exe” and set parameters to run LFAQ (Fig. 1). A parameter file including input data file path and quantification parameters is required before starting LFAQ.

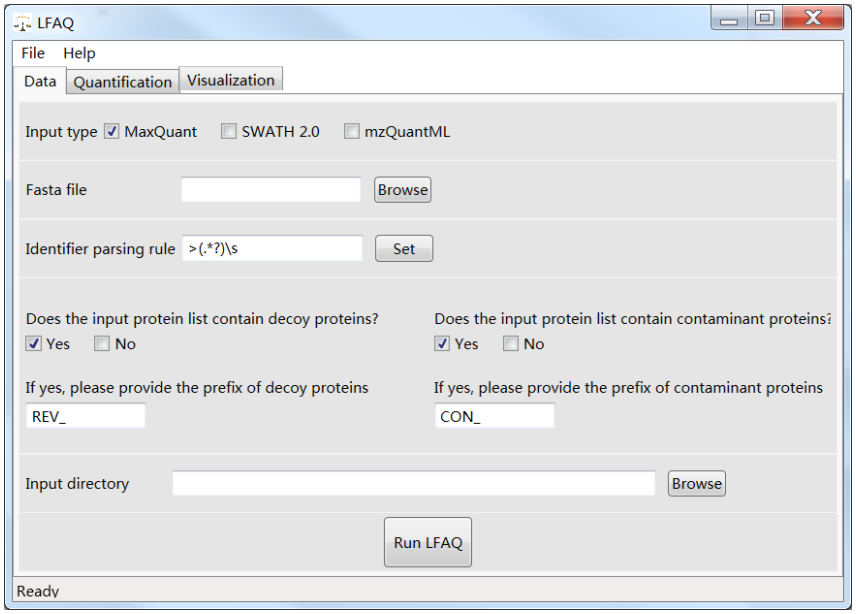

Figure 1. The interface for parameter setting.

2.1 Setting parameters and generating a new parameter file in GUI

In GUI, the LFAQ parameters are separated into two groups: the parameters about input data (Fig.2&3) and the parameters about quantification (Fig. 4&6).

2.1.1 Setting parameters about input data

2.1.1.1 Choose the input of LFAQ. LFAQ quantifies proteins based on the peptide quantification results from other software tools. Currently, LFAQ is compatible with the results of MaxQuant[1], SWATH 2.0 or the public standard quantification data format mzQuantML[2] proposed by HUPO/PSI (Fig. 2a). For MaxQuant results, the following files are required: (1) proteinGroups.txt, (2) peptides.txt, (3) experimentalDesignTemplate.txt. For SWATH 2.0 results, the second sheet “Area - peptides” of the result excel file should be saved as a separate CSV file at first. For mzQuantML format, just one file suffixed with .mzq is required.

2.1.1.2 Load the protein sequence file (in fasta format) used in identification. Click “Browse” button to select the fasta file. Note: the rows of protein headers must begin with the symbol “>” (Fig. 2b).

2.1.1.3 Provide a regular expression to properly extract the protein identifiers

from the protein sequence file. Click the “Set” button (Fig. 2c) to open the “Set Regular Expression” dialog (Fig. 3). Note that the “Input directory” should be set in advance when the input type is MaxQuant.

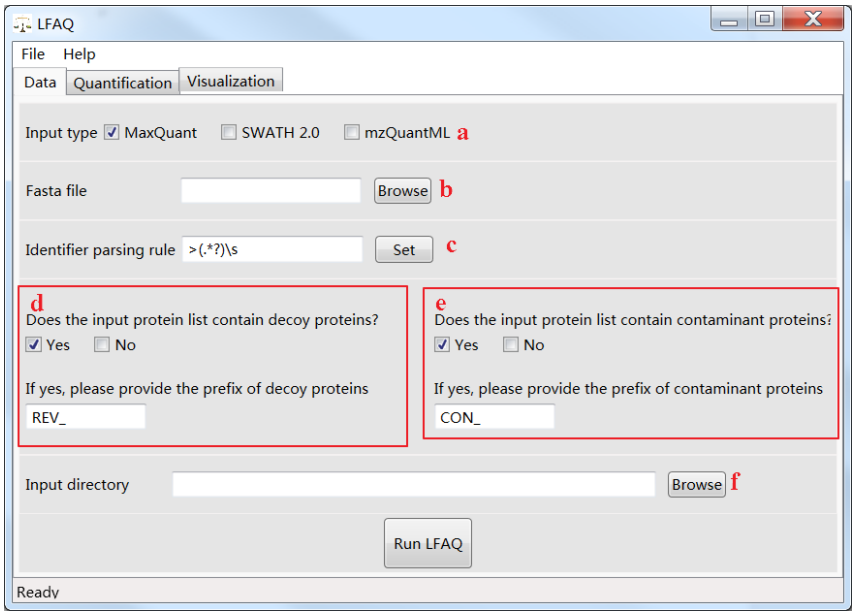

Figure 2. Parameter setting interface for input data.

In the “Set Regular Expression” dialog (Fig. 3), user can specify the regular expression used to extract the protein identifiers from the protein sequence.

(1) Click “Browse” button to choose the fasta file (Fig. 3a).

(2) Fill in the regular expression (Fig. 3b). There are some examples in the below for reference.

(3) Click “Test” button to check the regular expression (Fig. 3c). Note that the identifiers extracted from the fasta file by the regular expression should be the same as the protein IDs in input files.

(4) Click “OK” button if the regular expression is confirmed (Fig. 3d).

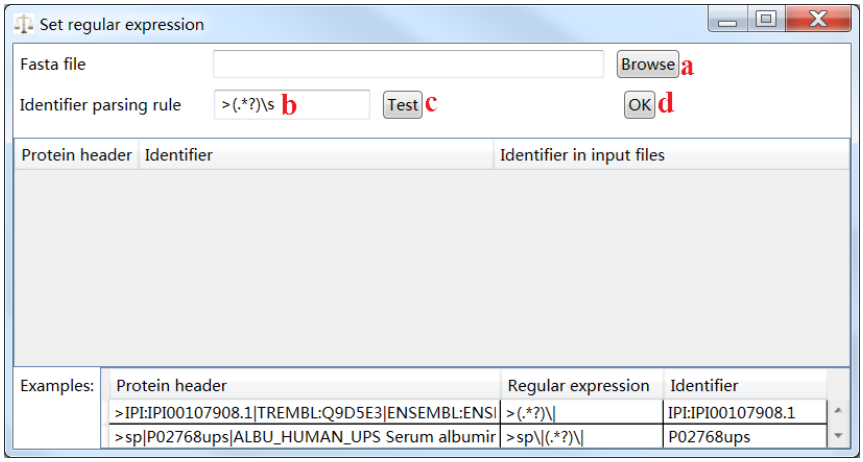

Figure 3. Parameter setting interface for the regular expression.

2.1.1.4 Choose if the input protein list contains decoy proteins or not. If yes, please provide the prefix of decoy proteins, otherwise leave it blank (Fig. 2d).

2.1.1.5 Choose if the input protein list contains contaminant proteins or not. If yes, please provide the prefix of contaminant proteins, otherwise leave it blank (Fig. 2e).

2.1.1.6 Choose the location (directory) where the input files were located or the absolute path of input file by clicking the “Browse” button (Fig. 2f). The absolute path of input file is required for SWATH 2.0 and mzQuantML. And the input directory is required for MaxQuant.

2.1.2 Setting parameters about quantification

2.1.2.1 Choose a regression method. To predict peptide quantification efficiency, we have implemented two regression methods in LFAQ: BART (Bayesian Addictive Regression Trees) and stepwise regression (Fig. 4a). When one of the regression methods is chosen, the corresponding parameters will appear (Fig. 5). Based on our experience, we suggest users choose the BART as the preferred selection.

2.1.2.2 Choose an enzyme (Fig. 4b). Until now, LFAQ only supports the “trypsin” enzyme.

2.1.2.3 Choose if the sample contains standard proteins or not (Fig. 4c). If yes, please provide the identifiers which can be used to distinguish standard proteins from other proteins. For example, if UPS2 proteins were used as standard proteins, just fill in “ups” for the textbox because all UPS2 protein names contain “ups”

(case-insensitive here). Then click the “Set standard proteins” button to open a dialog to specify the injection amounts of standard proteins.

2.1.2.4 Calculating iBAQ is optional (Fig. 4d).

2.1.2.5 Calculating Top3 is optional (Fig. 4e).

2.1.2.6 Choose the directory where the quantification results will be saved by clicking the “Browse” button (Fig. 4f).

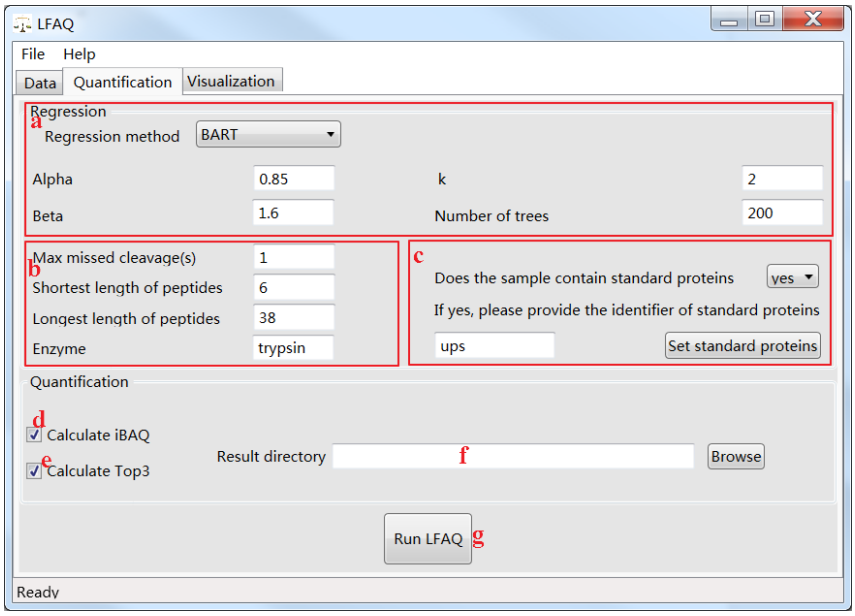

Figure 4. Parameter setting interface for quantification using BART regression method.

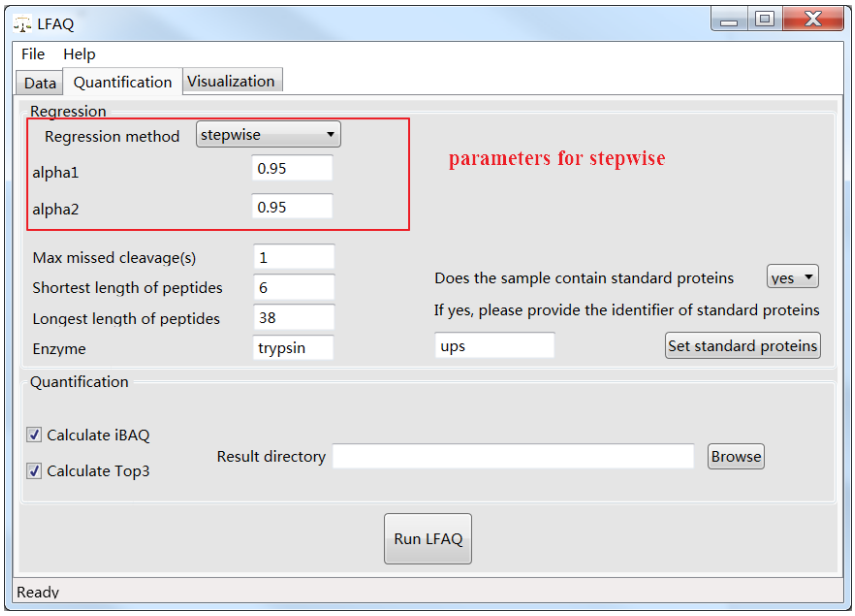

Figure 5. Parameter setting interface for quantification using stepwise regression method.

In the dialog of “Set standard proteins”, the user can edit the table of standard proteins (Fig. 6). The first column represents protein IDs, and the second column represents the injection amounts of the proteins. Fill in the amounts of standard proteins, then click “Save” button to save the information to the “StandardProteins.txt” file. Note that the corresponding standard protein IDs in the input file should contain the standard protein IDs in the “StandardProteins.txt” file.

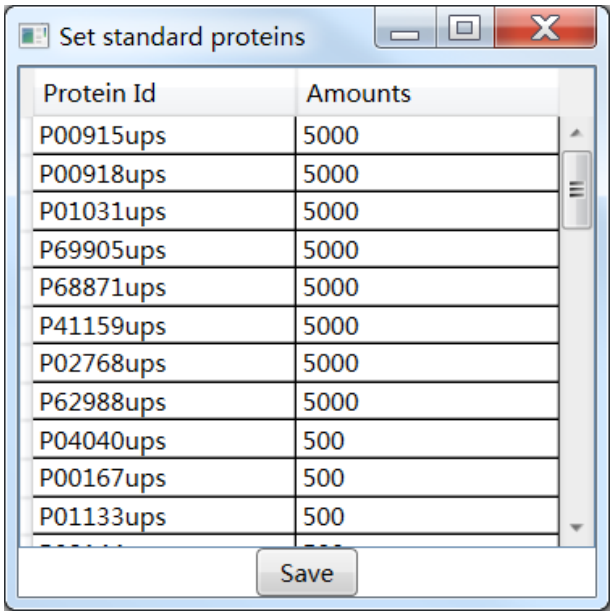

Figure 6. Standard protein setting dialog.

2.2 Loading an existing parameter file

If an existing parameter file (*.params) is available, a convenient way to set parameters is loading this parameter file directly (Fig. 7).

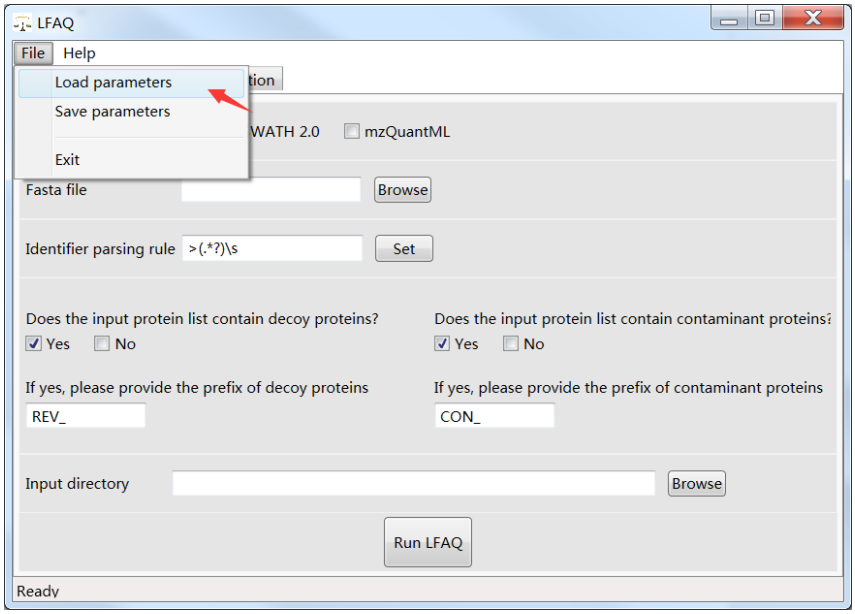

Figure 7. Interface of manually loading a parameter file.

2.3 Running label-free quantification

Make sure that all the parameters are appropriate, and then click the "Run LFAQ" button to start LFAQ (Fig. 4g). Then a parameter file named as the local time (e.g. parameters20170713-09-37-07.params) containing all the parameters set in the GUI is generated in the result file folder. When the background program is running, the status bar at the bottom of the window would show "LFAQ is running!" (Fig. 8). If the background program is done successfully, a dialog showing "LFAQ quantification finished successfully!" would appear (Fig. 9) and the status bar would show "Finished!".

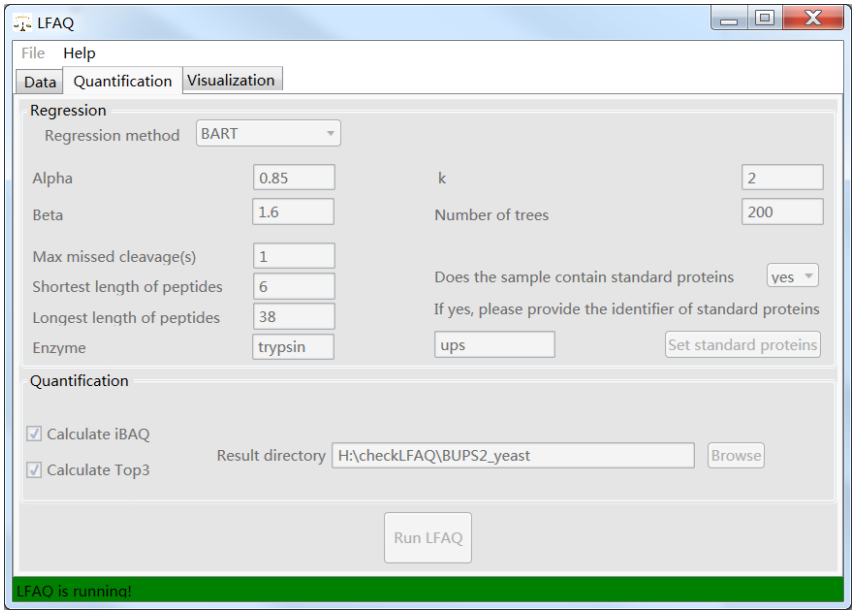

Figure 8. Interface when the background program is running.

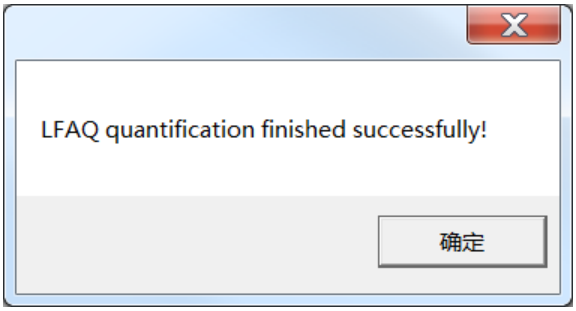

Figure 9. Dialog when the background program finished successfully.

3 Output File Formats

Once the calculation is done, LFAQ generates several files in the result directory (Table 1).

Table 1. Annotations of LFAQ quantification results.

File name	Annotations
ProteinsInfo.txt	The information of peptides and proteins extracted from input files.
ProteinResults*ExperimentName*.txt	The LFAQ protein quantification results of experiment *ExperimentName*. Note: CVOriginal and CVAfterCorrected represent the CV of peptide original intensities and the Q-factor corrected peptide intensities within a protein, respectively.
ProteinMergedResults.txt	The merged protein quantification results of different experiments.
RegressionResults*Experi*	The regression results to calculate the peptide Q-factor based on

mentName.txt	the protein input list of experiment *ExperimentName*.
MergedRegressionResults.txt	The merged regression results of different experiments.
SelectedFeatures*ExperimentName*.txt	The selected features by regression based on the protein input list of experiment *ExperimentName*.
PeptideFeatures*ExperimentName*.txt	All the 587 features of the peptides used in LFAQ based on the protein input list of experiment *ExperimentName*.
LFAQResultsForStandardProteins*ExperimentName*.txt	The actual and calculated abundances of standard proteins if the amounts of spiked-in standard proteins is provided based on the protein input list of experiment *ExperimentName*.
log.txt	All the running records in LFAQ.

Note that the *ExperimentName* represents the name of the corresponding experiment.

3.1 Annotation of ProteinResults*ExperimentName*.txt

This file contains the information of the quantification results of proteins from the experiment *ExperimentName*. The detailed description of every column in this file is given in Table 2.

Table 2. Descriptions of headers in ProteinResultsExperimentName.txt.

Column Name	Description
Protein IDs	All the protein IDs in a protein group.
Majority protein IDs	Representative protein IDs of this protein group.
NumberOfUniquePeptides	The number of unique peptides in the protein group.
LFAQ	The LFAQ value of the protein group, in which the Qfactor was used to correct the MS intensities at peptide level.
iBAQ	The iBAQ value is provided when the user chooses to calculate iBAQ. If MaxQuant is chosen as input, LFAQ directly outputs the iBAQ value calculated by MaxQuant，otherwise, the iBAQ value is calculated by LFAQ.
Top3	The Top3 value is provided when the user chooses to calculate Top3.
CVOriginal	The CV of native intensities of peptides in this protein group.
CVAfterCorrected	The CV of Q-factor corrected peptide intensities within this protein group.
NumberOfTheoreticEnzyme	The number of theoretically digested peptides.
PredictedMol(LFAQ)	The predicted amount of the protein group is calculated using LFAQ, when the user provides the amount of standard proteins.
PredictedMol(iBAQ)	The predicted amount of the protein group is calculated using iBAQ if checkbox “Calculate iBAQ” is chosen, and the user provides the amount of standard proteins.
PredictedMol(Top3)	The predicted amount of the protein group is calculated using

	Top3 if checkbox “Calculate Top3” is chosen, and the user provides the amount of standard proteins.
Coverage	The represent protein sequence coverage.
PeptideSequences	The sequences of the quantified peptides in this protein group.
PeptideOriginalIntensities	The native intensities of peptides in this protein group.
PeptideQfactors	The peptide Q-factor used to correct the original intensities at peptide level.
PeptideCorrectedIntensities	The corrected intensities of peptides in this protein group.
PeptideMWs	The molecule weight of peptides in this protein group.

3.2 Annotation of ProteinMergedResults.txt

This file contains the merged protein quantification results of LFAQ from different experiments. Note that for now, LFAQ only supports the SWATH 2.0 file containing a single experiment with one or more replicates.

Table 3. Descriptions of headers in ProteinMergedResults.txt.

Name	Description
Protein IDs	All the protein IDs in a protein group.
Majority protein IDs	Representative protein IDs of this protein group.
Experiments	The names of experiments where the representative proteins are identified.
LFAQ	The LFAQ values separated by semicolon of the protein group of all experiments.
iBAQ	The iBAQ values separated by semicolon of this protein group of all experiments are provided when the user chooses to calculate iBAQ. If MaxQuant is chosen as input, LFAQ directly outputs the iBAQ value calculated by MaxQuant，otherwise, the iBAQ value is calculated by LFAQ.
Top3	The Top3 values separated by semicolon of this protein group of all experiments are provided when the user chooses to calculate Top3.
PredictedMol(LFAQ)	The predicted amounts separated by semicolon of the protein group of all experiments are provided based on LFAQ and the amounts of standard proteins.
PredictedMol(iBAQ)	The predicted amounts separated by semicolon of the protein group of all experiments are provided based on iBAQ and the amounts of standard proteins.
PredictedMol(Top3)	The predicted amounts separated by semicolon of the protein group of all experiments are provided based on Top3 and the amounts of standard proteins.
LFAQCV	The protein CV of LFAQ values of all the experiments.
iBAQCV	The protein CV of iBAQ values of all the experiments.
Top3CV	The protein CV of Top3 values of all the experiments.

PeptidesIntensities	The corrected peptide intensities from all the experiments. They are separated by comma in one experiment and by semicolon in different experiments.

3.3 Annotation of LFAQResultsForStandardProteins*ExperimentName*.txt

This file contains the real abundances and the predicted abundances of standard proteins if the amounts of spiked-in standard proteins are provided based on the protein input list of experiment *ExperimentName*.

Table 4. Descriptions of headers in LFAQResultsForStandardProteinsExperimentName.txt.

Name	Description
Protein ID	The standard protein ID.
NumberOfUniquePeptides	The number of unique peptides of this protein.
SpikedInMols	The spiked-in amount of this standard protein.
LFAQ	The LFAQ value of this standard protein.
iBAQ	The iBAQ value of this standard protein is provided if the user chooses to calculate iBAQ.
Top3	The Top3 value of this standard protein is provided if the user chooses to calculate Top3.
PredictedMol(LFAQ)	The predicted amount of this standard protein is calculated using LFAQ
PredictedMol(iBAQ)	The predicted amount of this standard protein is calculated using iBAQ if checkbox “Calculate iBAQ” is chosen, and the user provides the amount of standard proteins.
PredictedMol(Top3)	The predicted amount of this standard protein is calculated using Top3 if checkbox “Calculate Top3” is chosen, and the user provides the amount of standard proteins.
Coverage	The represent protein sequence coverage of the used unique peptides.
ProteinSequenceLength	The length of represent protein sequence.

4 Result visualization

All LFAQ results are saved in text files. The visualization panel only shows two important files: ProteinMergedResults.txt and LFAQResultsForStandardProteins *ExperimentName*.txt (Fig. 8).

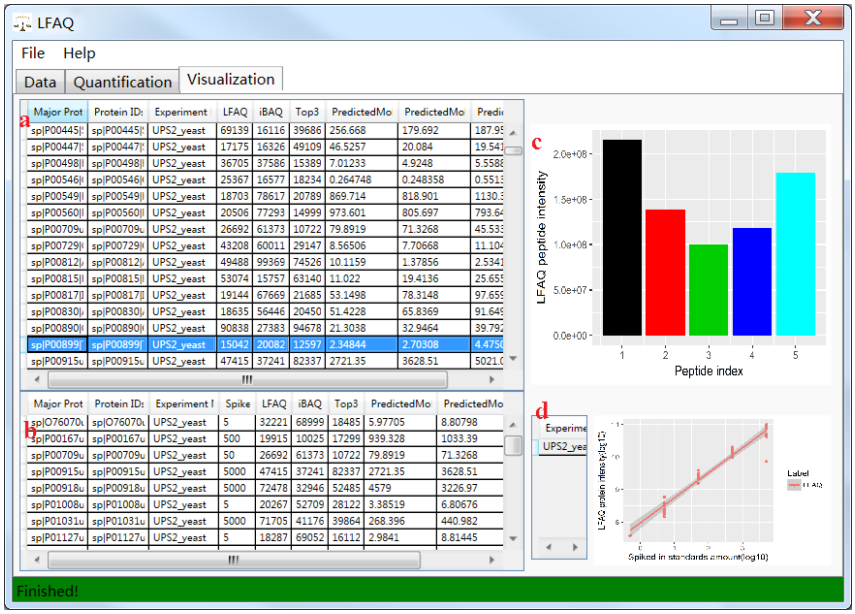

Figure 10. Visualization interfaces of LFAQ. (a) Table of the protein quantification results of different experiments. In this table, the results can be sorted by any attribute in column. (b) Bar plot of all the peptide intensities of the protein chosen in the left table. (c) Table of the standard protein quantification results of different experiments. (d) Scatterplot for the standard protein prediction results. The user can choose the experiment from which the prediction results were derived.

5 References

1. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372 (2008).

2. Walzer, M. et al. The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. *Mol Cell Proteomics* **12**, 2332-2340 (2013).