

**Online Quantitative Proteomics p-value Calculator for
Permutation-Based Statistical Testing of Peptide Ratios**

Author

Chen, David, Shah, Anup, Hien, Nguyen, Loo, Dorothy, Inder, Kerry L, Hill, Michelle M

Published

2014

Journal Title

Journal of Proteome Research

Version

Accepted Manuscript (AM)

DOI

[10.1021/pr500525e](https://doi.org/10.1021/pr500525e)

Rights statement

This document is the unedited Author's version of a Submitted Work that was subsequently accepted for publication in Journal of Proteome Research, copyright 2014 American Chemical Society after peer review. To access the final edited and published work see <http://dx.doi.org/10.1021/pr500525e>.

Downloaded from

<http://hdl.handle.net/10072/65208>

Griffith Research Online

<https://research-repository.griffith.edu.au>

Technical Note

Online Quantitative Proteomics p-value Calculator for Permutation-Based Statistical Testing of Peptide Ratios

David Chen¹, Anup Shah², Hien Nguyen^{2,3}, Dorothy Loo², Kerry L. Inder², Michelle M. Hill²

¹School of Information and Communication Technology, Griffith University, Brisbane, Australia

²The University of Queensland Diamantina Institute, The University of Queensland, Brisbane, Queensland, 4102, Australia

³School of Mathematics and Physics, The University of Queensland, Brisbane, Queensland, 4072, Australia

* Corresponding author:

Michelle M Hill

The University of Queensland Diamantina Institute

Level 5, Translational Research Institute, 37 Kent Street, Woolloongabba, QLD AUSTRALIA 4102

Tel: +61 (0)7 3443 7456

Fax: +61 (0)7 3443 5946

E-mail: m.hill2@uq.edu.au

Running title: Quantitative Proteomics p-value Calculator

Key words: quantitative proteomics, bioinformatics, SILAC, statistics

Abstract

The utility of high throughput quantitative proteomics to identify differentially abundant proteins en-masse relies on suitable and accessible statistical methodology, which remains mostly an unmet need. We present a free web-based tool, called Quantitative Proteomics p-value Calculator (QPPC), designed for accessibility and usability by proteomics scientists and biologists. Being an online tool, there is no requirement for software installation. Furthermore, QPPC accepts generic peptide ratio data generated by any mass spectrometer and database search engine. Importantly, QPPC utilizes the permutation test which we recently found to be superior to other methods for analysis of peptide ratios as it does not assume normal distributions¹. QPPC assists the user in selecting significantly altered proteins based on numerical fold-change, or standard deviation from the mean or median, together with the permutation p-value. Output is in the form of comma separated values files, along with graphical visualization using volcano plots and histograms. In this Technical Note, we evaluate the optimal parameters for use of QPPC, including the permutation level and the effect of outlier and contaminant peptides on p-value variability. The optimal parameters defined are deployed as default for the web-tool at <http://qppc.di.uq.edu.au/>.

INTRODUCTION

Comparative profiling (also called shotgun) proteomics experiments are now widely employed across biological and biomedical research. With the increased sensitivity and capability of modern mass spectrometers, a new challenge in comparative proteomics is the statistical assessment of quantitatively altered peptides and by inference, proteins, in these large datasets². Various labeling or label-free methods have been established to facilitate relative proteome quantitation (reviewed in ³). Chemical or metabolic labeling methods allow sample multiplexing during mass spectrometry. Relative quantitation is obtained by generating peptide ratios from the intensities of the precursor ions for each sample. A biological dataset with replicates commonly consists of thousands of peptide ratios, averaged to generate a mean protein ratio after database searching.

Currently an arbitrary protein ratio fold-change cut-off is often used as the sole criteria for determining the list of altered proteins in quantitative proteomics, possibly due to the lack of easily accessible statistical tools. While fold-change can be used as a simple indication of quantitatively differential proteins, statistical tests are needed to account for random errors and multiple hypothesis testing. Standard student's t-tests are used frequently in proteomics but are unsuitable if fewer than three biological replicates are performed and in any case, resulting p-values need to be adjusted for multiple hypothesis testing. Other statistical tests that have been applied to distinguish significantly altered proteins based on the peptide ratios include L-statistics⁴, linear mixed effects model⁵, linear regression model⁶, parametric heteroskedasticity model⁷, re-sampling based non-parametric statistical test⁸, regularized Bayesian T-test⁹ and a combined use of multiple tools¹⁰. Several free software packages have also been developed for

the research community, for example MaxQuant¹¹, Diffprot⁸ and Cyber-T Quantitative Mass Spec module^{9a}. Some current limitations of these tools include the requirement to install the software package locally, vendor-specific data formats. Furthermore, most of these methods assume that peptide ratios in such quantitative proteomics experiments have a normal distribution. This normality assumption has shown to be incorrect for a range of quantitative ratios, including SILAC (Stable Isotope Labeling with Amino acids in Culture) data^{1, 4, 12}. Through a comparison with four widely used statistical methods, we demonstrated the superiority of the permutation test for the statistical assessment of significant protein ratio changes. Orthogonal validation of the permutation p-values has been performed by western blotting of selected candidates¹³.

Here we report the development and implementation of a web-based tool called Quantitative Proteomics Permutation Calculator, QPPC. This tool is offered free of charge to the research community, with goals of providing Accessibility, Computation, and Usability:

- **Accessibility:** QPPC should be widely accessible by proteomics scientists without any technical setup prior to use. In addition, it should accept input quantitative proteomics data in a generic format regardless of the mass spectrometer used.
- **Computation:** QPPC should accurately compute the p-values for the input data and the user specified parameters using the permutation test described in Nguyen et al¹. Then it should allow users to examine the output p-values before determining the parameters to generate the list of significantly altered proteins. Based on these inputs, it should accurately select significantly altered proteins and generate relevant graphs for visualizing the results.

- **Usability:** Needless to say, QPPC should be easy to use. However, the main challenge in making QPPC user friendly is that the time it takes to upload the data and to compute p-values varies a lot depending on the network bandwidth, amount of data, and the number of permutations performed. QPPC should cater for usage scenarios ranging from where the result can be obtained within a few seconds, to where the result may take tens of minutes to produce. Furthermore, the results generated should be viewable online and downloadable for storing and printing.

EXPERIMENTAL PROCEDURES

QPPC is the result of a collaboration between biologists, a statistician, and a software engineer. The biologists provided the requirements and tested various versions of QPPC. The statistician designed and implemented the algorithm for computing p-values and identifying significantly altered proteins. The software engineer designed and implemented QPPC including the integration of the statistical computation code into the application.

Design and implementation

Accessibility

We chose a web application as a solution to meet the accessibility goal, because it is not always simple for users to install software on their local machines. In addition, the client side of this web application utilized HTML, CSS, and JavaScript, which will run on any recent web browsers without requiring any additional plug-in. The job of the client is simply to fetch the user input and submit to the server (which performs all the computation).

The implementation of the permutation p-value test and the selection of significantly altered proteins along with the file outputs are done on the server in the Statistical Module. The statistical module was implemented using the R programming language. R is a free and open-source programming language that is designed specifically for reading large tables of data, developing statistical computations, and generating graphical representation of the data. Using R has simplified the development process and reduced the likelihood of errors in the code. However, R is not designed for the web, and is unable to directly present the results it generates on the web. To facilitate communication between R and the web client, a Server Module was

built. The server module was implemented with a web-specific language called PHP. PHP is one of the most popular server-side scripting languages designed for web development. In addition to facilitating communication between the statistical module and the web client, the server module sets up the execution environment for the statistical modules, provides the required input, launches the execution of R scripts in a separate process, and relays the execution progress and final result to the client. The relationships between the client and server, as well as the relationships between the various languages used are shown in Figure 1A.

On the client side, there are the Data Module and the Presentation Module. The data module fetches the required data from the server module. The data sent between the data module and server module are encoded in JavaScript Object Notation (JSON). The presentation module formats and presents/displays the data. It also handles user interaction.

Computation

The main QPPC computation can logically be divided into two stages: 1) compute p-values and 2) generate a list of significantly altered proteins. The required input and the resulting output for each stage is shown in Figure 1B.

Stage 1 takes a file input containing peptide data and parameters from the user. The input file format was modeled on the peptide summary export (.ssv) file generated directly by Spectrum Mill database searching software (Agilent Technologies). To ensure wide flexibility, comma-separated values (.csv) or tab-separated values (.txt) files are also accepted, the only requirement for the input file is that the following column headers with their respective data values exist in

these files: “accession_number”, “entry_name”, and “ratio”. Alternatively the “ratio” column header can directly indicate the type of ratio, as shown in the Spectrum Mill peptide summary export files. A drop-down menu allows the selection between “ratio”, “L/H”, “L/M”, “M/L”, “M/H”, “H/L”, or “H/M”.

There are three parameters required for Stage 1 which specify:

1. The number of permutations to perform.
2. The ratio (column) to use.
3. Whether to perform outlier removal before computing P-values. If outlier removal is required, the peptide ratio threshold needs to be specified. Peptide ratios greater than this number (and its inverse) will be removed as outliers.

Prior to permutation testing, QPPC performs data pre-processing to remove

- a) ratios that are negative, not numbers or 0s;
- b) peptides that are the only single observation of its respective protein.

If Outlier Removal is selected,

- c) peptides that have a ratio outside the threshold set by Outlier Remover.

A summary of all the peptide ratios removed, their assigned proteins, and the reason for their removal is available as a downloadable excel file at the end of Stage 1 computation.

The permutation p-value algorithm implemented is as described in Nguyen et al.¹. The output for stage 1 is a .csv file containing the following statistical values for each peptide found in the input

file: the mean ratio, the standard deviation, the log of mean, the log of standard deviation, the number of observations, and the p-value. In addition to the permutation p-value, we also implemented family-wise error rate (FWER) and false discovery rate (FDR) adjusted p-value computations. The FWER and FDR adjusted p-values were computed via Bonferroni¹⁴ and Benjamini-Yekutieli¹⁵ corrections, respectively.

Stage 2 takes both the p-values file produced from Stage 1 and the user-defined parameters for determining significantly altered proteins. As the p-values file is already on the server, the user only needs to enter the parameters. The following describes the parameters to stage 2 and the computations performed with these parameters:

1. The normalization type to be applied before the analysis. User can select either None, Mean, or Median. When Mean or Median is selected, normalization will be applied to Mean or Median over all quantified protein average log-ratios. Normalization is often useful to counteract known biases such as unequal loading of samples or incomplete incorporation in the labeling stage.
2. The p-value criterion. Proteins with p-values less than the specified cut-off value are deemed significant. This is the first of two filtering criteria.
3. The second criterion. User can select between protein ratio fold-change and standard deviations. When the fold-change criterion is used, a protein is deemed significant if its average ratio is greater than the cut-off or less than the inverse of the cut-off (e.g. a cut-off of 2 fold-change implies proteins with average ratios above $\log(2)$ or below $-\log(2)$ are deemed significant). When the standard deviation cut-off is used, a protein is deemed significant if its average log-ratio is a cut-off number of standard deviations away from

the average log-ratio of all proteins quantified (e.g. if the cut-off is 2 standard deviations, and the standard deviation is 2 and mean of the average log-ratio are 2 and 0 respectively, then a protein is deemed significant if it has an average log-ratio below -4 or above 4).

4. The cut-off value for the second criterion. If fold-change is the chosen criterion, then the cut-off value should be greater than 1. If standard deviation is the chosen criterion, then the cut-off value should be greater than 0.

The output of Stage 2 contains two spreadsheets (.csv files): the first indicates which proteins are significant under either or both criteria and the second provides summary statistics for all quantified proteins. Accompanying these spreadsheets are (1) a volcano plot which is useful for visualizing the location and spread of proteins that were deemed significant under the criteria, and (2) a histogram of protein log-ratios which is helpful for distributional assessments.

The results produced by Stage 1 and 2 of QPPC were verified with simulation and experimental results. See Results and Discussion section for more detail.

Usability

QPPC is built as a single-page web application to provide a smoother and more responsive user experience. Responsive in this context refers to the application's ability to provide feedback to the users (this should not be confused with the term "responsive web design" which refers to a website that is designed to work on different screen sizes). For tasks that may take some time to complete, such as uploading the data file or computing p-values, QPPC is designed to provide feedback on the progress of these tasks and has an offline processing mode.

With file upload, QPPC is designed so that file upload will commence right after the file is selected or dragged-and-dropped onto the upload area. Furthermore, the file upload is performed in the background; this allows upload progress be shown, and at the same time the user can select parameters. For large files, or users with slow Internet upload speed, the user can chose to automatically start p-values computation right after the file upload is complete to avoid having to wait for the file upload. The maximum input file size is currently set at 20MB. File upload is implemented with the help of an open source JavaScript library called Fine-Uploader (<https://github.com/Widen/fine-uploader>).

The process of obtaining the progress of p-values computation is implemented by having the data module periodically poll the server module to obtain and display the progress. The server module in turn communicates with the statistical module by having the statistical module write its progress to a predefined file, and the server module read from it.

In the situation where the dataset is large, and the number of permutations required is high, the p-values might take over ten minutes to compute. This might be too long for users to wait. Hence, QPPC is designed with an offline processing mode. In this mode, a user can submit a p-values computation job with an email address. The user may close the browser once the job is submitted. When the result is ready, an email containing a link to the result will be sent to the specified address. In addition, this email also contains a second link to allow user to continue to determine significantly altered proteins (Stage 2).

This offline processing ability is implemented by using unique session identification (SID) numbers. A SID is generated on the server and passed to the client once the input data file has been uploaded to server. Conceptually, a SID is associated with an input data file. The client is required to include the SID as a URL parameter in all subsequent requests to the server to indicate on which dataset the computation should be performed, and which results to retrieve. Similarly, links containing SID are emailed to the user for retrieving results and to continue with Stage 2 processing.

The results generated by QPPC are all downloadable as CSV or PDF files, so users are not dependent on QPPC to view their results. In addition, QPPC provides an online interactive display of the volcano plot and histogram. The advantage of the online interactive views is that the users can hover their mouse pointer over a point on the volcano plot or a block in the histogram to visualise the details of a protein for that point or block. Furthermore, users can zoom in (or out) of the volcano plot to distinguish between tightly packed points on the original scale. The interactive view was implemented with Google Charts (<https://developers.google.com/chart/>).

Results generated by QPPC are stored in the server for two weeks. During this period, users can use the results link in the email to download their results. Results older than two weeks are automatically deleted.

Experimental Dataset

To determine the effect of the number of permutations on estimated p-values, we utilized a published SILAC dataset of peptide ratios, in which the effect of caveolin-1 loss on the murine embryonic fibroblast detergent-resistant membrane proteome was examined¹⁶. The exported peptide summary document from Spectrum Mill has ratios for 19595 peptides, out of which 2853 are negatives, 164 are non-numbers, and 115 are ratios of zero. Two input files, before and after removing known contaminants (i.e. keratins and serum albumin) from the original datasets, were submitted to Stage 1 of QPPC, either at 1000 or 10000 permutations, with an outlier removal threshold of 100. The agreement between estimated p-values after 1000 and 10000 permutations was then analyzed by observing their variability using Bland-Altman plots.

Safety considerations

There are no specific safety considerations in using QPPC.

RESULTS AND DISCUSSION

We developed the open-access web-based tool QPPC to help proteomics scientists undertaking permutation p-value calculations for quantitative proteomics experiments. Furthermore, tools for visualization and selection of significantly altered proteins were also included. The development of QPPC has been an iterative process. Once a version of QPPC is developed it is tested by the biologists. Feedback from the biologists on the Accessibility, Computation and Usability of the tool was used to derive the requirements and the design for the next version.

Figure 2A shows the current QPPC homepage where users upload a data file and select parameters for Stage 1 computation. A single data file is used for a permutation p-value calculation, which should include all biological or technical replicates of the experiment. The input data format is a list of peptide ratios with the peptides having been assigned to protein identities, for example, the peptide summary export from Spectrum Mill (Agilent) search. As QPPC makes no assessment on the peptide-protein assignment, the user should be confident of the data quality and protein assignment.

Output from a Stage 1 computation can be directly taken to Stage 2 which provides user flexibility to determine significantly altered proteins with the help of multiple criteria (Figure 2B). The user can optionally normalize the protein ratios to the median or mean ratio for the dataset at this stage. Stage 2 outputs include the interactive histogram and volcano plot, as well as 4 downloadable files: a csv file containing the data, a second csv file containing summary of the mean, median ratios and standard deviation for the experimental dataset, and pdf files of the histogram and volcano plot (Figure 3). Given the short computation time required, users can

easily repeat Stage 2 using a different input parameter after inspecting the outputs. Experimental considerations of the user parameters are examined in this paper.

Permutation level

As the permutation test is a type of randomized test, it requires aggregations over repeated computations to determine the p-values. Increasing the number of repetitions will increase the accuracy of the p-values, but will also increase the computation time. In order to determine the optimal permutation level to recommend as default, we compared the effect of different permutation levels on the p-value error and the computation time, using a simulated dataset of 10000 peptides and 500 proteins (Figure 4). Ten simulations were performed for each of the permutation levels: 100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000 and 100000.

Figure 4A shows the margin of error for a normal 99.999% CI for true p-values ranging from 0.05 to 0.001, against the permutations level. Figure 4B shows the time required in seconds for various permutation levels. From Figure 4B, we see that a quadratic relationship fits the points well. This implies that as the permutation level increases, a square factor of effort increase is needed. This is nonlinear and thus the effort grows faster than the size of the problem.

Additionally, from Figure 4A we see that the margin of error for a 99.999% CI for various true p-values is quite acceptable at the 1000 permutations level. For example if the true p-value was 0.05, the 99.999% CI would be (0.02, 0.08) thus making any true p-value of 0.05 significant at the 0.1 level 99.999% of the time. Increasing the permutation level from 1000 to 10000 according to our timing simulations would increase the effort from 13.66 seconds to 147.35

seconds, whilst only giving us a reduction in the margin of error for the 99.999% interval of a p-value=0.05 example from 0.03 to 0.01.

We further evaluated the effect of permutation levels on calculated p-values using a real biological data set from a lipid raft SILAC proteomics study¹⁶. The results were analyzed by Bland-Altman plots where the mean p-value for each protein was plotted against the difference of p-values obtained after 1000 and 10000 permutations (Figure 5A). The Bland-Altman plot is an effective way to compare observations which have small differences in measurement and/or datasets without proportional differences between the methods¹⁷. The horizontal lines represent 95% limit of agreement at ± 1.96 standard deviation away from mean of the p-value difference between 1000 and 10000 permutations. To aid visualization of significantly altered proteins, a vertical line was drawn at p-value=0.05, and the significantly altered proteins to the left were colored in magenta (Figure 5A). The result indicates that, while there is high variability in the calculated p-values (ranging from 0.2 to 0.8), the significant p-values (magenta dots) remained within ± 0.02 away from one another, hence the list of significant proteins remains the same for 1000 or 10000 permutation levels. Given this result, we have used 1000 permutations as the default value for QPPC, however, since computation of 10000 permutations of this dataset took less than 10 minutes and produced more accurate estimates of p-values, users may choose to increase the number of permutations to 10000 for small-medium datasets.

Outlier peptide ratios and contaminant proteins

QPPC performs data quality checks prior to computation. At this step, peptide ratios with non-numerical, negative and zero values are removed as they are technical codes from the database

searching software not valid for statistical computation. In addition, any single peptide-ratios that uniquely identified a protein are also removed since single point measurements are unreliable for statistical analysis, particularly as several biological or technical replicates are expected to be combined and analyzed in one data file. A protein identified/quantified based on one peptide ratio out of all replicates will be low-confidence.

Users may choose to remove outliers based on biological or technical knowledge. For example, keratins can accumulate during sample processing and cell lysates can be contaminated with albumin from serum. Since contaminants introduced during sample handling are almost always 'light' labeled in SILAC experiments, inclusion of such skewed contaminant ratios in the permutation analysis could have an impact on the calculated p-value. Arguably the best way to treat these contaminants is to manually remove them, based on biological knowledge, prior to submitting the dataset for analysis. However, this may not be practical because not all contaminants are known or can be inferred from biological knowledge, and some quantitative proteomics datasets may contain tens of thousands of peptides when several experiments are combined. Therefore, we have also included an optional threshold-based outlier remover which removes all ratios greater than the input variable or less than its inverse. This optional peptide ratio outlier removal was designed to remove any outrageous ratios which are virtually impossible based on the properties of the system. The current web default value of 100 was empirically chosen based on our experimental data in which ratios were hardly ever more than 10. The 10-fold difference minimizes any accidental removal of true data.

Effect of permutation level and contaminant proteins on the variability of p-values

Low level protein contamination is expected during proteomic sample preparation. To evaluate the effect of this on the variability of permutation p-values, we made use of the same SILAC dataset of lipid raft proteome which required significant sample handling during preparation of detergent-resistant membranes but utilized a liquid handler for in-gel digestion to minimize contamination¹⁶. In this dataset, 186 peptide ratios assigned to 14 known protein contaminants were identified and removed before submitting to QPPC. Out of the 186 peptide ratios, 130 are negatives, 18 are non-numbers and 3 are technical outliers (ratio over 100), resulting in all ratios associated with 7 contaminant keratins being removed during the QPPC pre-processing step. So the current analysis compared the effect of 35 peptide ratios mapping to 7 contaminant proteins on the permutation p-values and their biological significance. The level of contamination accounted for 0.21% of the total analyzed ratios. Table 1 shows QPPC Stage 1 output for the seven contaminant proteins.

The dataset with/without manual contaminant removal was submitted to QPPC at 1000 permutations, opting for the technical outlier removal threshold of 100. The results were again analysed by Bland-Altman plots (Figure 5B), which showed no effect of manual contaminant peptide ratio removal. It has to be noted, however, that this example contains a relatively low number of known contaminants, so their effect on p-value was minor. It is possible that a large number of contaminants and outliers in a dataset could lead to erroneous permutation p-values, hence manual inspection of datasets for potential large-scale contamination is recommended if there is any suspicion of such systemic errors.

CONCLUSIONS

QPPC provides a simple, user-friendly web-interface for permutation statistical analysis of quantitative proteomics data. By using a generic data input format consisting of peptide ratios with assigned protein accession numbers, QPPC can analyze quantitative proteomics data obtained via any instrument or database searching software. To further facilitate quantitative proteomics data analysis an additional parameter of fold-change or deviation of the ratio from group median/mean can be used, together with the permutation p-value, to compute a list of significantly altered proteins. An interactive, downloadable histogram and volcano plot allow visualization of the distribution of protein ratios according to these parameters. QPPC is free of charge and available at <http://qppc.di.uq.edu.au/>.

ACKNOWLEDGEMENTS

The authors thank Dr Leonard Foster and Dr Fiona McMillan for critical reading of the manuscript, Marcus Schull and Darren D'Souza for technical support. MMH is a Future Fellow of the Australian Research Council (FT120100251).

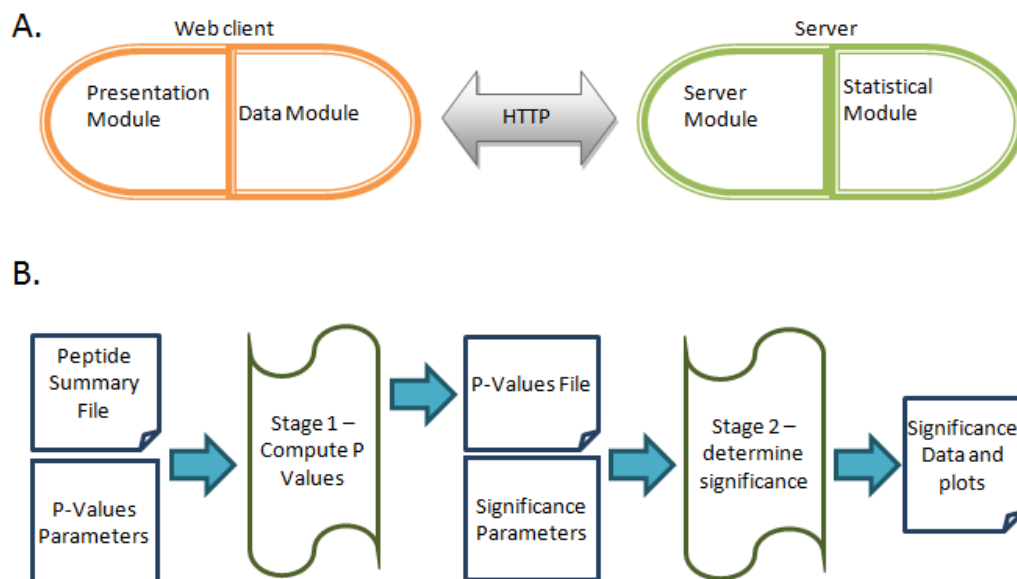

FIGURES

Figure 1. QPPC server architecture (A) and workflow (B).

A.

Quantitative Proteomics P-value Calculator



Home
Documentation ▾
Cite Us
Contact

This application assist selection of significantly altered proteins from quantitative proteomics experiments in two stages:
Stage 1 computes p-values using a distribution-free permutation method based on simulation of the log(ratio).
Stage 2 (available after completion of Stage 1) facilitates selection of significantly altered proteins based on two parameters - the p-value and fold-change or standard deviation from the mean.

Stage 1: Compute P-Values

Data to be analysed:

Click or Drop a file here

Use Sample Data File (for demonstration)

Ratio to use:

L/H ▾

Number of permutations:

1000

Remove Outlier

Outlier removal threshold:

100

Email result (optional):


✉

Start analysis after upload completes.

[Analyse](#)

B.

Quantitative Proteomics P-value Calculator



Home
Documentation ▾
Cite Us
Contact

This application assist selection of significantly altered proteins from quantitative proteomics experiments in two stages:
Stage 1 computes p-values using a distribution-free permutation method based on simulation of the log(ratio).
Stage 2 (available after completion of Stage 1) facilitates selection of significantly altered proteins based on two parameters - the p-value and fold-change or standard deviation from the mean.

Stage 1: Compute P-Values

Result - P-Values

Stage 2: Selection of Significantly Altered Proteins and Plots

Determine significantly altered proteins and produce volcano plot and histogram.

Normalization type:

None ▾

P-value cutoff level:

0.05

Type of criterion:

Fold Change ▾

Cutoff for the criterion:

2

Use adjusted p-value:


Yes ▾

[Analyse](#)

Figure 2. Web interface for Stage 1 (A) and Stage 2 (B) computation.

A.

Quantitative Proteomics P-value Calculator



Home
Documentation
Cite Us
Contact

This application assist selection of significantly altered proteins from quantitative proteomics experiments in two stages:
Stage 1 computes p-values using a distribution-free permutation method based on simulation of the log(ratio).
Stage 2 (available after completion of Stage 1) facilitates selection of significantly altered proteins based on two parameters - the p-value and fold-change or standard deviation from the mean.

Stage 1: Compute P-Values

Result - P-Values

ID	DETAIL	NORMALIZED_MEAN	STANDARD_DEVIATION	LOG2_MEAN	LOG2_SD	NUMBER_OF_OBSERVATIONS	PERM_P_VALUE	D_SCORE	SIGNIFICANT_UNDER_P_VALUE	SIGNIFICANT_UNDER_CRITERION	SIGNIFICANT_UNDER_BOTH
P1217	Quarantine protein Q110231071 subunit beta-1	0.1420902	0.22320872	0.121706749	0.217003857	126	0.88340386	-0.166284238	FALSE	FALSE	FALSE
P11055	S-mucliosin	0.2295536	0.220458857	0.211354282	0.215682776	447	1	-0.422487823	TRUE	TRUE	TRUE
Q71106	Tubulin alpha-1A chain	0.7124927	0.499758914	0.80935887	0.282701261	160	1.00604	0.889473854	TRUE	FALSE	FALSE
P60708	AdGn_viraplexin-1	0.2163831	0.27317896	0.208043766	0.21102682	267	0.482883713	0.010407806	FALSE	FALSE	FALSE
P11057	NADH-dependent proteinase 2	0.8832412	0.208483864	-0.143064893	0.215719488	58	0.871412819	-0.770883817	FALSE	FALSE	FALSE

Stage 2: Selection of Significantly Altered Proteins and Plots

Result - Significantly Altered Proteins and Plots

QPPC Stage 2 Result is ready
View:

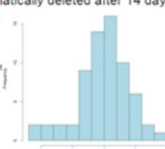
- Interactive Plots

Download:

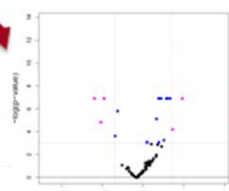
- QPPC SIG Output csv
- QPPC SIG Additional csv
- QPPC SIG Volcano plot pdf
- QPPC SIG Histogram pdf

Note: This result will be automatically deleted after 14 days.

Stage 2 Results



Histogram



Volcano Plot

Disclaimer:
This tool is provided free of charge and for research use only. The user to verify the accuracy, completeness, timeliness, quality or suitability for a particular use of the tool provided on this site. The University of Queensland (UQ) make no claims.

B.

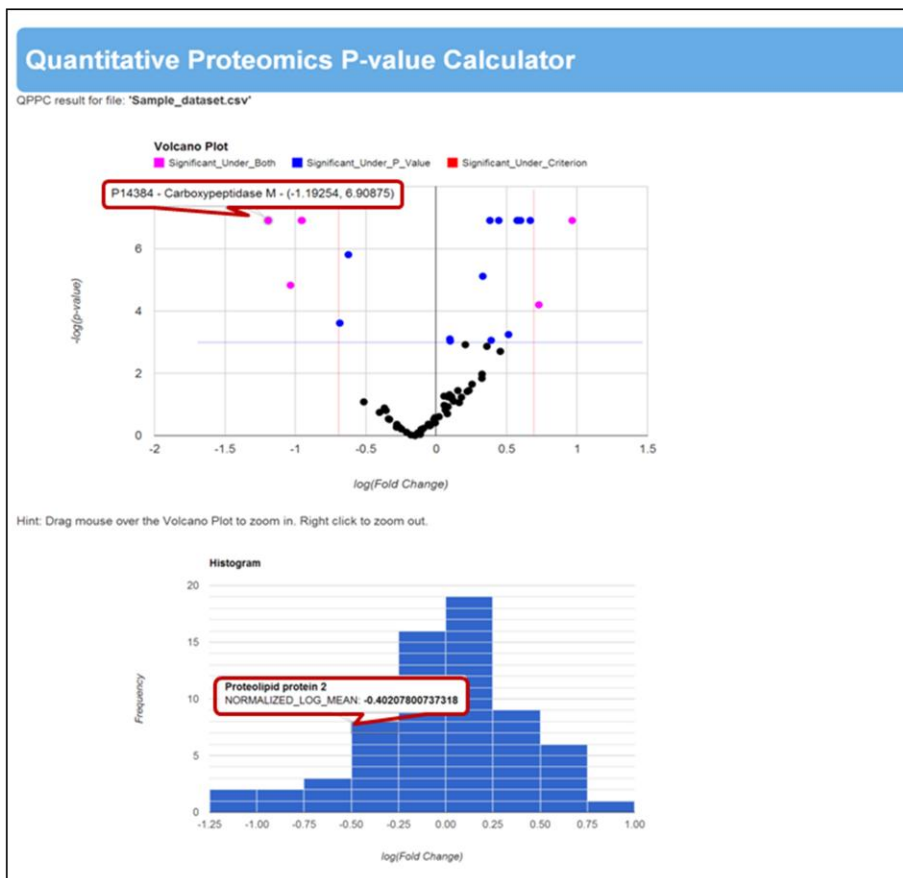


Figure 3. QPPC Stage 2 outputs.

(A) Stage 2 produces 4 downloadable files: QPPC_SIG_Output.csv reports the significance of each protein based on the two parameters: QPPC_SIG_Additional.csv reports the overall mean, median ratio and standard deviation for the dataset; QPPC_SIG_Volcano_plot.pdf and QPPC_SIG_Histogram.pdf are two graphics files for the analyzed data. (B) Interactive volcano plot and histogram allows real time evaluation of the analyzed data, and can be access from the Stage 2 output through View – Interactive Plots.

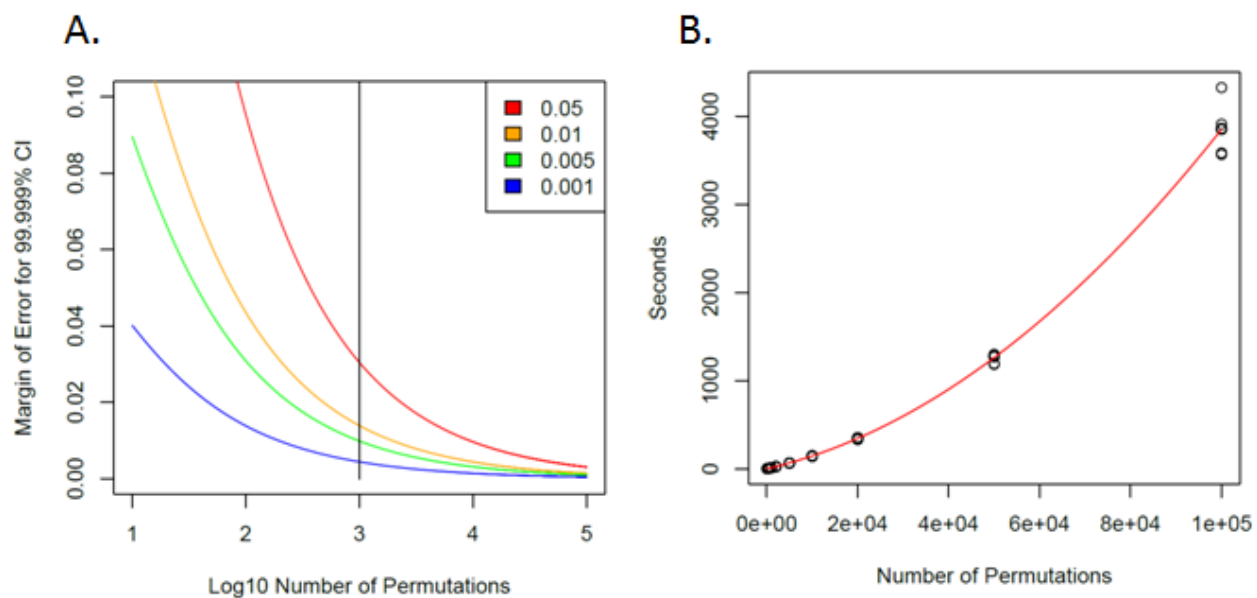


Figure 4. Effect of permutation level on margin of error and computation time.

Ten simulations for performed for each of the permutation levels 100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000 and 100000, using a 10000 peptides and 500 proteins simulated dataset. (A) plots the margin of error for 99.999% CI against the log(number of permutations). (B) plots the number of seconds used for each of the permutation runs.

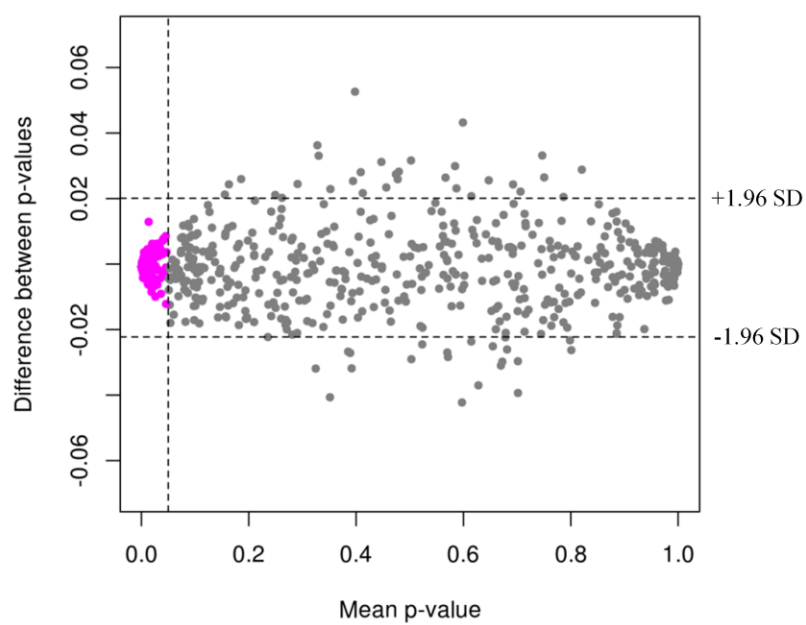
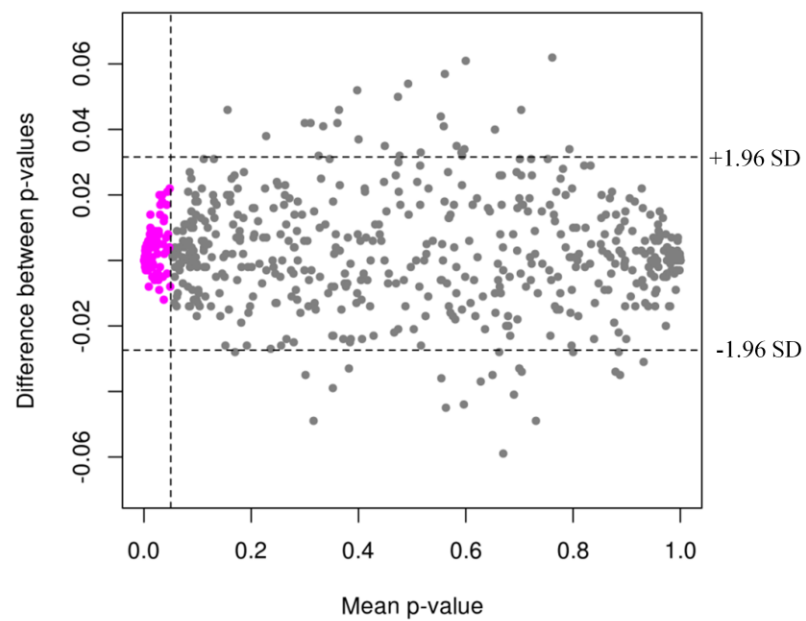
A. Effect of number of permutations on p-values**B. Effect of contaminant removal on p-values**

Figure 5. Effect of number of permutations and contaminant removal on the variability of p-values.

Bland-Altman plots measuring agreement of estimated p-values between (A) 1000 and 10000 permutations, and (B) with and without manual contaminants removal at 1000 permutations. For both experiments, the difference in the calculated p-values of the significantly altered proteins ($p < 0.05$, coloured in magenta) was less than 2 standard deviation (SD) and did not switch to $p > 0.05$.

TABLE

Protein Name	UniProt Accession	Number of ratios	Mean Ratio	St Dev	Perm p-value
Keratin, type I cytoskeletal 10	P02535	7	2.470	0.739	0.029597
Keratin, type II cytoskeletal 1	P04104	3	38.457	30.962	1.00E-04
Keratin, type II cytoskeletal 1b	Q6IFZ6	6	10.326	2.969	1.00E-04
Keratin, type II cytoskeletal 2 epidermal	Q3TTY5	5	14.652	14.793	1.00E-04
Keratin, type II cytoskeletal 5	Q922U2	6	28.446	9.759	1.00E-04
Keratin, type II cytoskeletal 7	Q9DCV7	2	1.588	0.123	0.460054
Serum albumin	P07724	6	17.561	20.477	1.00E-04

Table 1. Proteins removed as known contaminants in the experimental dataset.

QPPC Stage 1 output summary for the 7 known contaminants manually removed. Peptide ratios associated with the following contaminants were removed in the pre-processing step of QPPC: keratin, type II cytoskeletal 79; keratin, type II cytoskeletal 2 oral; keratin, type I cytoskeletal 13; keratin, type I cytoskeletal 42; keratin, type I cytoskeletal 14; keratin, type I cytoskeletal 16 and Keratin, type II cytoskeletal 8.

St Dev, standard deviation; Perm p-value, permutation p-value.

REFERENCES

1. Nguyen, H.; Wood, I.; M.M., H., A robust permutation test for quantitative SILAC proteomics experiments. *Journal of Integrated OMICS* **2012**, *2* (2), DOI: 10.5584/jiomics.v2i2.109.
2. Podwojski, K.; Stephan, C.; Eisenacher, M., Important issues in planning a proteomics experiment: statistical considerations of quantitative proteomic data. *Methods Mol Biol* **2012**, *893*, 3-21.
3. (a) Bantscheff, M.; Lemeer, S.; Savitski, M. M.; Kuster, B., Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem* **2012**, *404* (4), 939-65; (b) Nikolov, M.; Schmidt, C.; Urlaub, H., Quantitative mass spectrometry-based proteomics: an overview. *Methods Mol Biol* **2012**, *893*, 85-100.
4. Cho, H.; Smalley, D. M.; Theodorescu, D.; Ley, K.; Lee, J. K., Statistical identification of differentially labeled peptides from liquid chromatography tandem mass spectrometry. *Proteomics* **2007**, *7* (20), 3681-92.
5. Jorge, I.; Navarro, P.; Martinez-Acedo, P.; Nunez, E.; Serrano, H.; Alfranca, A.; Redondo, J. M.; Vazquez, J., Statistical model to analyze quantitative proteomics data obtained by 18O/16O labeling and linear ion trap mass spectrometry: application to the study of vascular endothelial growth factor-induced angiogenesis in endothelial cells. *Molecular & cellular proteomics : MCP* **2009**, *8* (5), 1130-49.
6. Ting, L.; Cowley, M. J.; Hoon, S. L.; Guilhaus, M.; Raftery, M. J.; Cavicchioli, R., Normalization and statistical analysis of quantitative proteomics data generated by metabolic labeling. *Molecular & cellular proteomics : MCP* **2009**, *8* (10), 2227-42.
7. Breitwieser, F. P.; Muller, A.; Dayon, L.; Kocher, T.; Hainard, A.; Pichler, P.; Schmidt-Erfurth, U.; Superti-Furga, G.; Sanchez, J. C.; Mechtler, K.; Bennett, K. L.; Colinge, J., General statistical modeling of data from protein relative expression isobaric tags. *J Proteome Res* **2011**, *10* (6), 2758-66.
8. Malinowska, A.; Kistowski, M.; Bakun, M.; Rubel, T.; Tkaczyk, M.; Mierzejewska, J.; Dadlez, M., Diffprot - software for non-parametric statistical analysis of differential proteomics data. *J Proteomics* **2012**, *75* (13), 4062-73.
9. (a) Kayala, M. A.; Baldi, P., Cyber-T web server: differential analysis of high-throughput data. *Nucleic Acids Res* **2012**, *40* (Web Server issue), W553-9; (b) Baldi, P.; Long, A. D., A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* **2001**, *17* (6), 509-19.
10. Schwammle, V.; Leon, I. R.; Jensen, O. N., Assessment and improvement of statistical tools for comparative proteomics analysis of sparse data sets with few experimental replicates. *J Proteome Res* **2013**, *12* (9), 3874-83.
11. Cox, J.; Matic, I.; Hilger, M.; Nagaraj, N.; Selbach, M.; Olsen, J. V.; Mann, M., A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. *Nat Protoc* **2009**, *4* (5), 698-705.
12. Gerster, S.; Kwon, T.; Ludwig, C.; Matondo, M.; Vogel, C.; Marcotte, E. M.; Aebersold, R.; Buhlmann, P., Statistical approach to protein quantification. *Molecular & cellular proteomics : MCP* **2014**, *13* (2), 666-77.
13. Inder, K. L.; Zheng, Y. Z.; Davis, M. J.; Moon, H.; Loo, D.; Nguyen, H.; Clements, J. A.; Parton, R. G.; Foster, L. J.; Hill, M. M., Expression of PTRF in PC-3 Cells modulates cholesterol

dynamics and the actin cytoskeleton impacting secretion pathways. *Molecular & cellular proteomics : MCP* **2012**, *11* (2), M111 012245.

14. Lehmann, E. L.; Romano, J. P., *Testing Statistical Hypotheses* Springer, New York: 2005.

15. Benjamini, Y.; Yekutieli, D., The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **2001**, *29*.

16. Inder, K. L.; Loo, D.; Zheng, Y. Z.; Parton, R. G.; Foster, L. J.; Hill, M. M., Normalization of protein at different stages in SILAC subcellular proteomics affects functional analysis. *Journal of Integrated OMICS* **2012**, *2* (2), 114-122.

17. Dewitte, K.; Fierens, C.; Stöckl, D.; Thienpont, L. M., Application of the Bland–Altman plot for interpretation of method-comparison studies: a critical investigation of its practice. *Clinical chemistry* **2002**, *48* (5), 799-801.

