Table of Contents

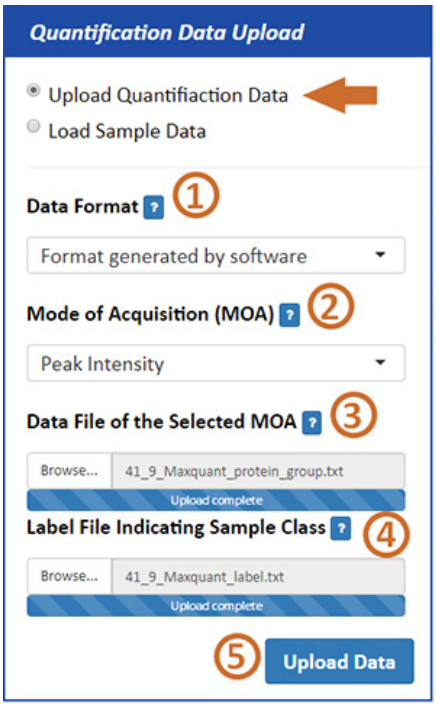
---

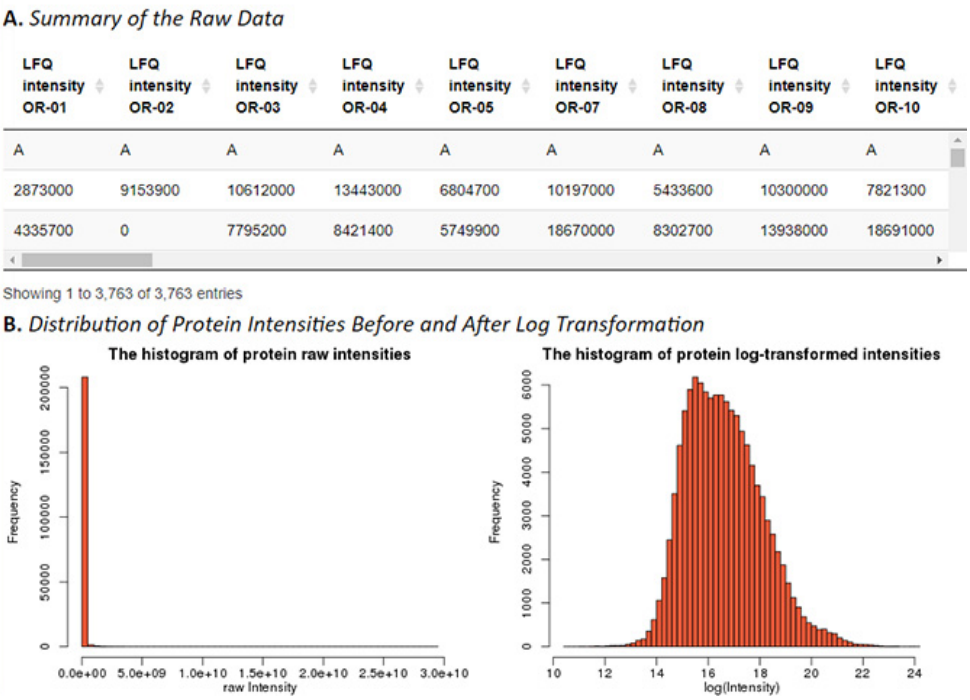
1. Step-by-step Instruction on the Usage of ANPELA

Analysis and subsequent performance assessment are started by clicking on the “Analysis” panel on the homepage of ANPELA. The collection of web services and the whole process provided by ANPELA includes: (Step 1) uploading the quantification data, (Step 2) method's assumption assessment and data transformation & pretreatment, (Step 3) data filtering & missing value imputation, and (Step 4) performance assessment of the proteome quantification.

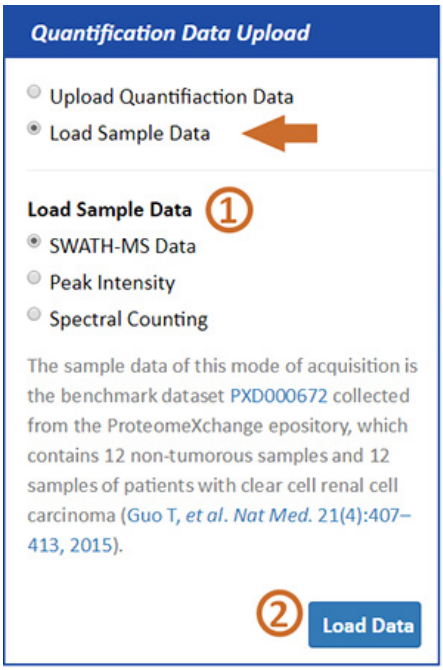
Step 1. Uploading Quantification Data

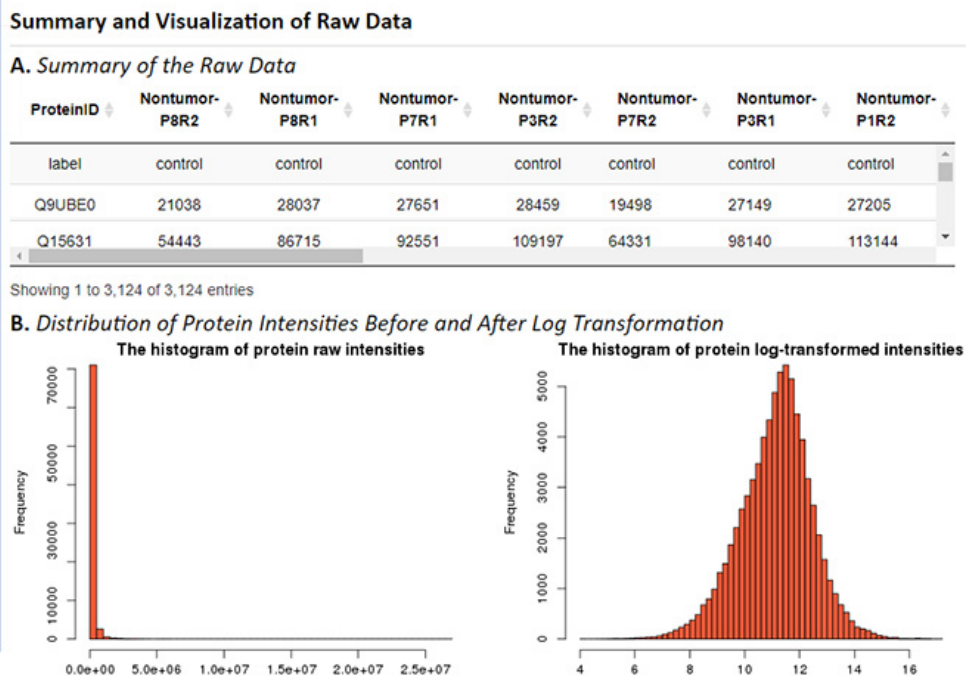
By click “Upload Quantification Data”, users are allowed to upload their data in various formats generated by popular software tools for label-free quantification. All software tools aim at processing the raw proteomics data acquired by 3 quantification measurements (SWATH-MS, peak intensity and spectral counting). Users are asked to upload the specific file containing the data generated by those tools, together with a label file indicating the classes of each sample (detail information of the file format can be found in the **Section 2** of this Manual). Moreover, in case that users want to process their data before ANPELA analysis, they are allowed to upload their processed data in a unified format defined by ANPELA which could be readily found **HERE** (Right Click to Save). By clicking the “Upload Data” button, the quantification data provided by the users can be uploaded for further analysis.





Three sets of sample data are also provided in this step facilitating a direct access and evaluation of ANPELA. These sample data are all benchmark datasets collected from the PRoteomics IDEntifications (PRIDE) database developed by the European Bioinformatics Institute. Particularly, the sample data for SWATH-MS is the dataset PXD000672 containing 12 non-tumorous samples and 12 samples of patients with clear cell renal cell carcinoma (Guo T, *et al*. *Nat Med*. 21(4):407-413, 2015); the sample data for protein intensity is the dataset PXD005144 with 66 samples of pancreatic cancer patients and 36 samples of chronic pancreatitis patients (Saraswat M, *et al*. *Cancer Med*. 6(7):1738-1751, 2017); and the sample data for spectral counting is the dataset PXD001819 providing yeast cell lysat samples of different concentrations (0.5 *vs* 50 fmol/microgram) acquired by MS2 spectral counting (Ramus C, *et al*. *J Proteomics*. 132:51-62, 2016). By clicking the "Load Data" button, the sample dataset selected by the users can be uploaded for further analysis.

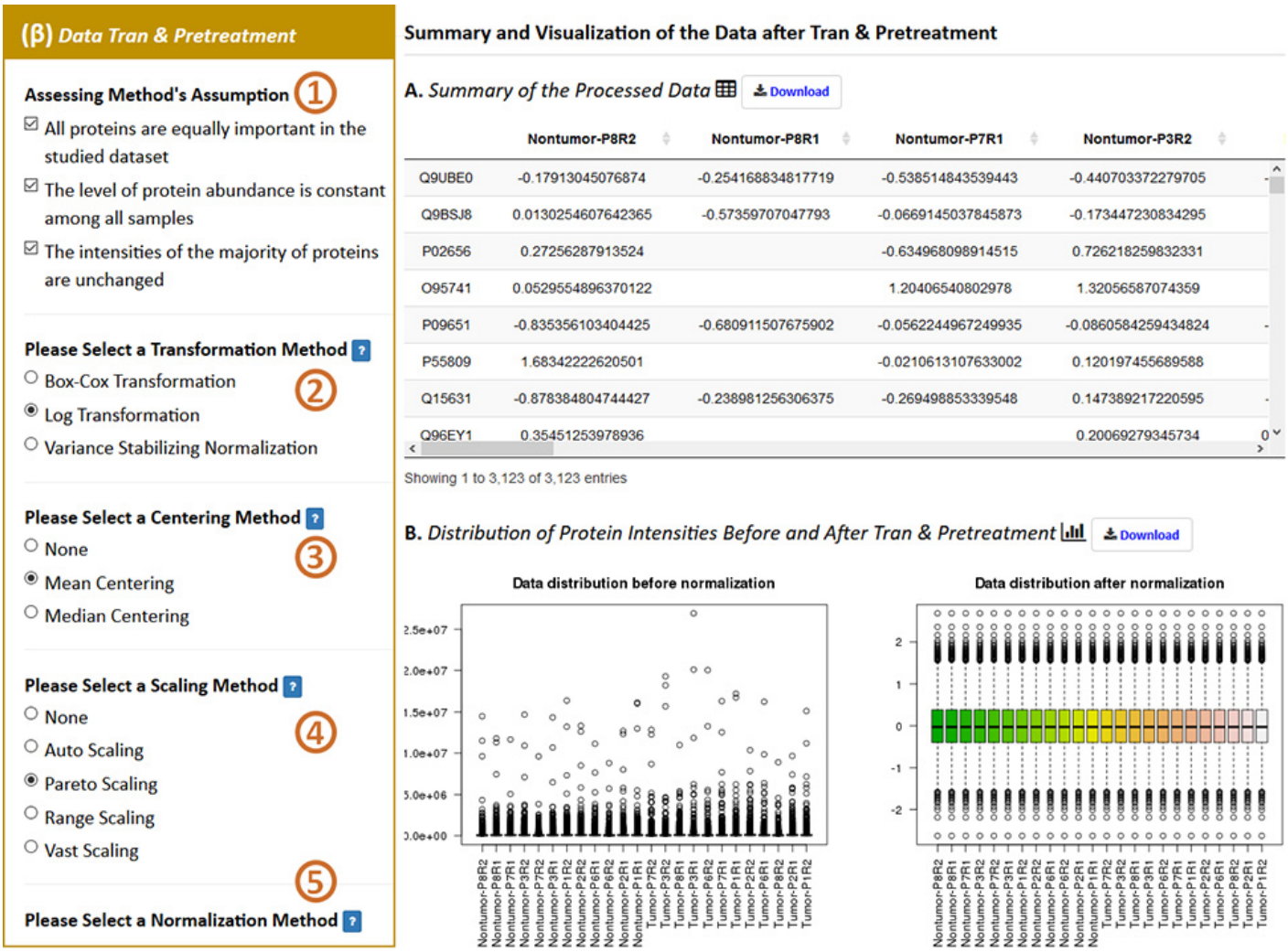




Step 2. Method's Assmuption Assessment and Data Transformation & Pretreatment

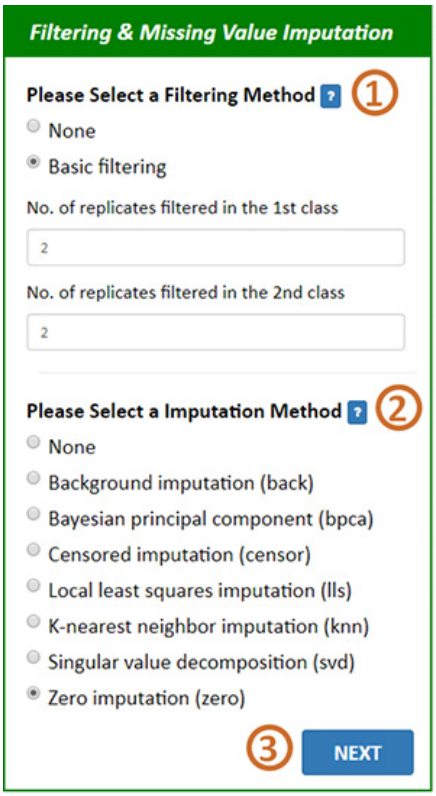
The manipulation methods were reported to be based on their own statistical assumption about the data, which might make them inappropriate for manipulating some proteomic data. Taking pretreatment methods as examples, there were generally three types of assumptions: (**Assumption A**) all proteins were assumed to be equally important; (**Assumption B**) the level of protein abundance was assumed to be constant among all samples; (**Assumption C**) the intensities of the vast majority of the proteins were assumed to be unchanged under the studied conditions. Due to these distinct assumptions, some methods may be fundamentally inappropriate for certain dataset and cannot be assessed for the studied datasets. Therefore, before any performance assessment, users should first analyze the nature of their datasets, and then assess and indicate whether the method's assumption held for these data.

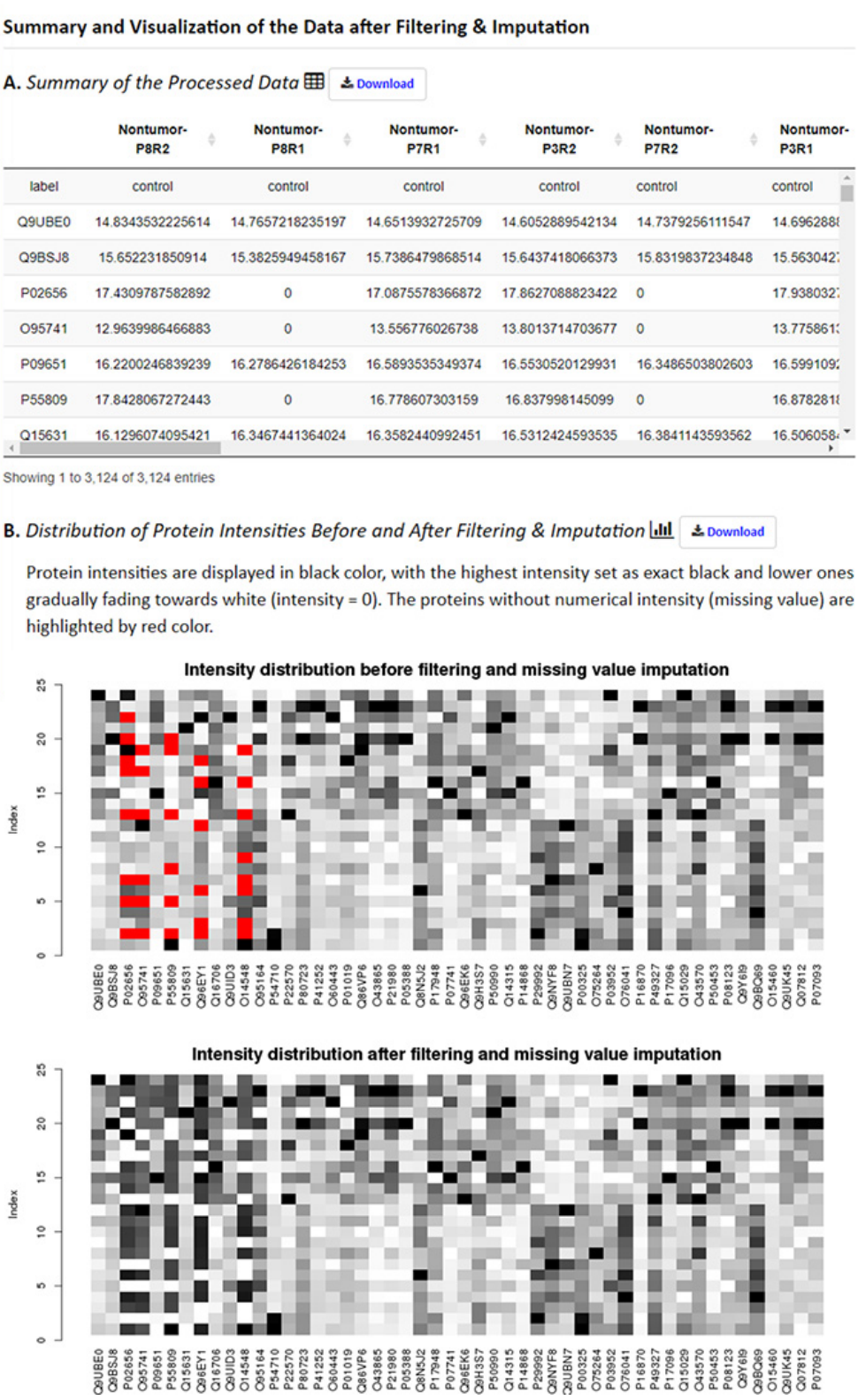
Users are provided with the option to conduct pretreatment on their uploaded data. In total, 3 types of transformation methods frequently applied to manipulate the label-free proteomics data are included. Furthermore, the current version of ANPELA offers 18 pretreatment methods popular for centering, scaling and normalizing the proteomics data. A detail explanation on each method is provided in the **Section 3** of this Manual. By clicking the "PROCESS" button, a summary of the processed data and a plot of the intensity distribution before and after data manipulation are automatically generated. All resulting data and figures can be downloaded by clicking the "Download" button. Moreover, the sample outputs of "Summary of the Processed Data" and "Distribution of Protein Intensities" that performs interactively in the same way as real output are provided.



Step 3. Data Filtering & Missing Value Imputation

Data filtering and missing value imputation are subsequently provided in this step. The filtering method used here is the basic filtering, and 7 imputation methods frequently applied to treat missing value are covered, which include *Background Imputation, Bayesian Principal Component Imputation, Censored Imputation, K-nearest Neighbor Imputation, Local Least Squares Imputation, Singular Value Decomposition and Zero Imputation*. A detail explanation on each imputation method is provided in the **Section 3** of this Manual. By clicking the “PROCESS” button, a summary of the processed data and a plot of the intensity distribution before and after data manipulation are automatically generated. All resulting data and figures can be downloaded by clicking the “Download” button. Moreover, the sample outputs of "Summary of the Processed Data" and "Distribution of Protein Intensities" that performs interactively in the same way as real output are provided.



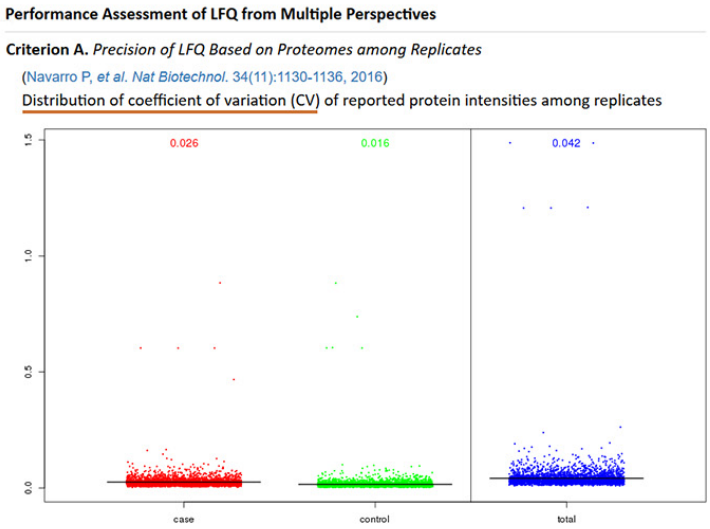


Step 4. Performance Assessment of Label-free Quantification (LFQ) from Multiple Perspectives

Five well-established criteria for a comprehensive evaluation on the performance of LFQ are provided in ANPELA, and each criterion is either quantitatively or qualitatively assessed by various metrics. These criteria include:

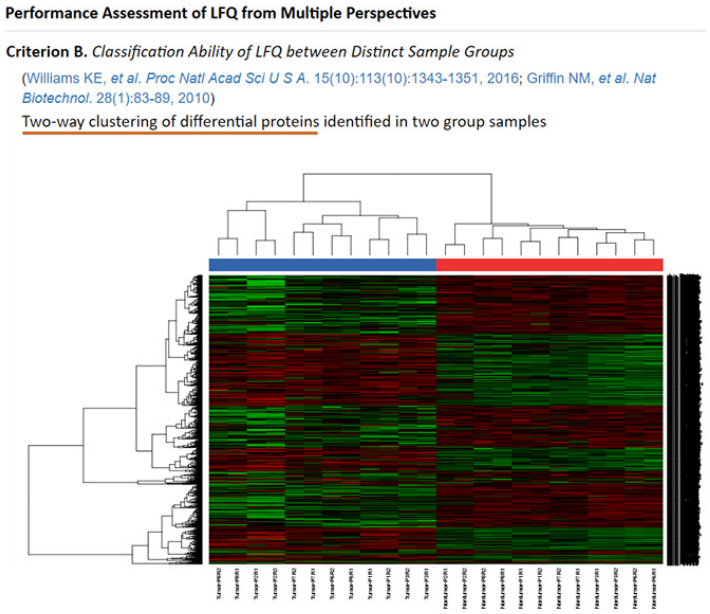
Criterion A: ***Precision of LFQ Based on Proteomes among Replicates***
(Kuharev J, *et al*. *Proteomics*. 15(18):3140-3151, 2015)

Different quantification measurements, various kinds of software for pre-processing raw proteomics data, and diverse methods for data manipulation profoundly affect the precision of LFQ, which can be assessed by the coefficient of variation (CV) of reported protein intensities among replicates (Navarro P, *et al*. *Nat Biotechnol*. 34(11):1130-1136, 2016; Kuharev J, *et al*. *Proteomics*. 15(18):3140-3151, 2015). In particular, the metric CV is designed to reflect LFQ's ability to reduce variation among replicates, and therefore to enhance the technical reproducibility (Chawade A, *et al*. *J Proteome Res*. 13(6):3114-3120, 2014). The lower value (illustrated by boxplots below) of CV denotes more thorough removal of experimentally induced noise and indicates better precision of LFQ. Moreover, the sample outputs of "Distribution of CV" that performs interactively in the same way as real output are provided.



Criterion B: ***Classification Ability of LFQ between Distinct Sample Groups***
(Griffin NM, *et al*. *Nat Biotechnol*. 28(1):83-89, 2010)

An appropriate LFQ is expected to retain or even enlarge the difference in proteomics data between two distinct sample groups (Griffin NM, *et al*. *Nat Biotechnol*. 28(1):83-89, 2010). A heatmap hierarchically clustering samples based on their protein intensities is therefore frequently used as an effective metric to assess LFQ's classification ability (Griffin NM, *et al*. *Nat Biotechnol*. 28(1):83-89, 2010). Firstly, the total number of protein intensities in each sample is reduced by feature selection. Then, proteins (rows) and samples (columns) are clustered based on their similarities in protein intensity profile. Detail process on how to assess LFQ's classification ability can be found in the prestigious publication by Griffin NM, *et al*. (Griffin NM, *et al*. *Nat Biotechnol*. 28(1):83-89, 2010). Moreover, the sample outputs of "Two-way clustering of differential proteins" that performs interactively in the same way as real output are provided.



Criterion C: ***Differential Expression Analysis Based on Reproducibility-optimization***
(Karpievitch YV, *et al*. *BMC Bioinformatics*. 13(S16):S5, 2012)

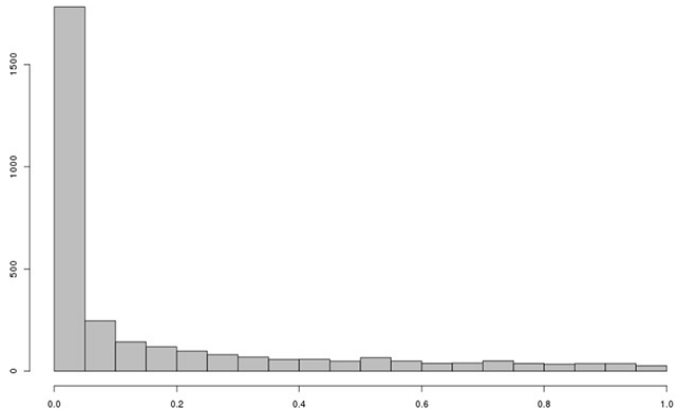
To avoid overfitting or confounding in LFQ, the distribution of *P*-values of protein intensities between distinct sample groups is examined (Risso D, *et al*. *Nat Biotechnol*. 32(9):896-902, 2014). Ideally, one expects a uniform distribution for the bulk of non-differentially expressed proteins, with a peak in the [0.00, 0.05] interval corresponding to proteins with differential intensity (Risso D, *et al*. *Nat Biotechnol*. 32(9):896-902, 2014). Moreover, the volcano plot colored proteins with differential intensity can give a glance of the total number of differentially expressed proteins (Välikangas T, *et al*. *Brief Bioinform*. doi:10.1093/bib/bbx054, 2017). In the proteomics (and other OMICs) studies that explore the mechanism underlining complex biological process, a limited number of differentially expressed proteins may resulted in false discovery (Blaise BJ. *Anal Chem*. 85(19):8943-8950, 2013). Therefore, the differential significance of protein intensities between sample groups measured by P-values is firstly calculated using the reproducibility-optimized test statistic (ROTS) package in ANPELA (Pursiheimo A, *et al*. *J Proteome Res*. 14(10):4118-4126, 2015). Secondly, the distribution of P-values and the volcano plot are provided. Skewed distribution of P-values may indicate overfitting and/or confounding (Karpievitch YV, *et al*. *BMC Bioinformatics*. 13(S16):S5, 2012). Moreover, the sample outputs of "Distriubtion of P-values" and "Volcano plot of protein markers" that perform interactively in the same way as real output are provided.

Performance Assessment of LFQ from Multiple Perspectives

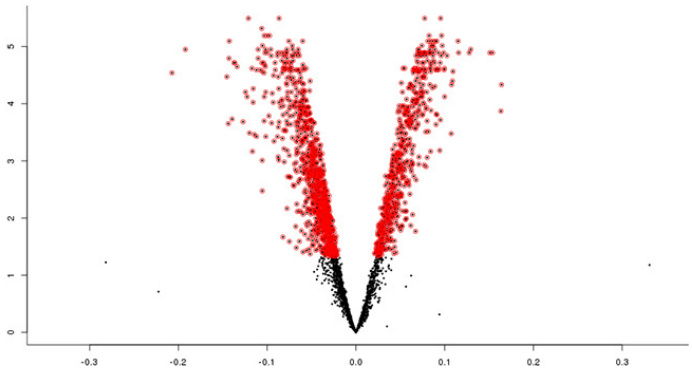
Criterion C. *Differential Expression Analysis Based on Reproducibility-optimization*

(Välikangas T, *et al. Brief Bioinform.* doi:10.1093/bib/bbx054, 2017)

A. Distribution of *P*-values of protein intensities between distinct sample groups



B. Volcano plot of the proteins identified as differentially expressed between distinct sample groups



Criterion D: ***Reproducibility of the Identified Protein Markers among Different Datasets***

(Li B, *et al*. *Nucleic Acids Res*. 45(W1):162-170, 2017)

Consistency score is a popular criterion used to represent the robustness of protein marker identification (Li B, *et al*. *Nucleic Acids Res*. 45(W1):162-170, 2017), which is calculated to quantitatively measure the overlap of identified protein markers among different partitions of a given dataset (Wang X, *et al*. *Mol Biosyst*. 11(5):1235-1240, 2015). The higher consistency score represents the more robust results in protein marker identification (Li B, *et al*. *Nucleic Acids Res*. 45(W1):162-170, 2017). Thus, the random sampling is firstly preformed within LFQ dataset to produce multiple sub-datasets. Then, each protein is ranked according to its significance measured by q-value and absolute fold changes. Thirdly, top-ranked proteins in each sub-dataset are selected as markers. Finally, a consistency score is calculated based on these markers using equation (Wang X, *et al*. *Mol Biosyst*. 11(5):1235-1240, 2015) as follow:

$$S = \sum_{i=2}^{C} \sum_{S \epsilon I_i} 2^{i-2} \cdot n_S$$

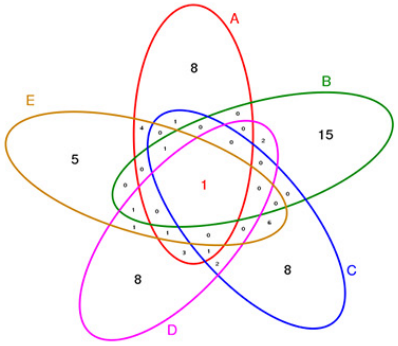
where C is the total number of sub-datasets, I_i indicates a set of significant protein makers containing the intersections of any i sub-datasets, and n_S refers to the number of markers in the intersection S. Moreover, the sample outputs of "Venn diagram illustrating marker numbers" that performs interactively in the same way as real output are provided.

Performance Assessment of LFQ from Multiple Perspectives

Criterion D. *Reproducibility of the Identified Protein Markers among Different Datasets*

(Collins BC, *et al. Nat Commun.* 8(1):291, 2017; Li B, *et al. Nucleic Acids Res.* 45(W1):162-170, 2017)

A. Venn diagram illustrating maker numbers and their overlaps among different rounds of sampling (the overlapping regions indicate the makers shared by multiple rounds)



B. The consistency score of the identified markers among different rounds of sampling is 35.3

Criterion E: ***Accuracy of LFQ Based on Spiked and Background Proteins***

(Navarro P, *et al*. *Nat Biotechnol*. 34(11):1130-1136, 2016)

Additional experimental data (e.g. spiked proteins) are frequently generated and used as references to validate or adjust the performance of LFQ (Kuharev J, *et al*. *Proteomics*. 15(18):3140-3151, 2015; Navarro P, *et al*. *Nat Biotechnol*. 34(11):1130-1136, 2016), and the expected log fold changes (logFCs) are known both for the spiked and the background proteins (the expected LogFC for background proteins equals to zero) (Välikangas T, *et al*. *Brief Bioinform*. doi:10.1093/bib/bbx054, 2017). In ANPELA, the reproducibility-optimized test statistic (ROTS) is firstly applied to identify the differentially expressed proteins. Then, the true positive rate (TPR), the true negative rate (TNR) and the precision (PRE) for the success discovery of the spiked proteins are calculated. The higher the TPR, the more accurate the LFQ achieves. Moreover, the logFCs of protein intensities (for both spiked and background proteins) between two sample groups are calculated, and the level of correspondence between the quantification and the expected logFCs is then assessed by the mean squared error (MSE). The performance of LFQ can be reflected by how well the quantification logFCs corresponded to what are expected based on the references (Välikangas T, *et al*. *Brief Bioinform*. doi:10.1093/bib/bbx054, 2017). Moreover, a boxplot illustrating the deviations of both quantification and expected logFCs of the spiked proteins is provided. The preferred median in boxplot would be zero with minimized deviations. The required format of the file providing the information of the spiked proteins can be readily downloaded **HERE** (Right Click to Save). The users will be asked to upload this file in the “Performance Assessment” step, and multiple metrics under this criterion will be calculated to the users for evaluating their selected quantification workflow. Moreover, the sample outputs of "Deviations between the quantification and the expected LogFCs of the spiked proteins", "Deviations of both spiked and background proteins between the quantification and the expected", "Metrics measuring LFQ performance" and "ROC curve of classification accuracy" that perform interactively in the same way as real output are provided.

Performance Assessment of LFQ from Multiple Perspectives

Criterion E. *Accuracy of LFQ Based on Spiked and Background Proteins*

(Dowle AA, *et al. J Proteome Res*. 15(10):3550-3562, 2016; Navarro P, *et al. Nat Biotechnol*. 34(11):1130-1136, 2016)

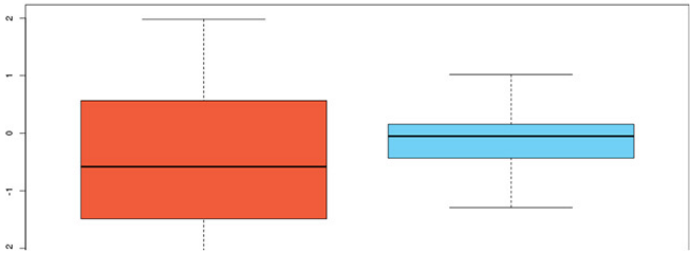
A. Deviations between the quantification and the expected logFCs of the spiked proteins

Proteins	ExpectedFC logFCs	Quantification logFCs	Deviation
P02768ups	4.60517018598809	6.58239527768116	1.97722509169307
P06396ups	4.60517018598809	6.19690364707524	1.59173346108715
P02787ups	4.60517018598809	6.48818055382322	1.88301036783513
P01008ups	4.60517018598809	5.57336889425456	0.968198708266465
P12081ups	4.60517018598809	6.44401630081293	1.83884611482484
P02788ups	4.60517018598809	5.82700386642013	1.22183368043204
P10636ups	4.60517018598809	5.72677184965377	1.12160166366568
P06732ups	4.60517018598809	6.21272443633043	1.60755425034234

Showing 1 to 48 of 48 entries

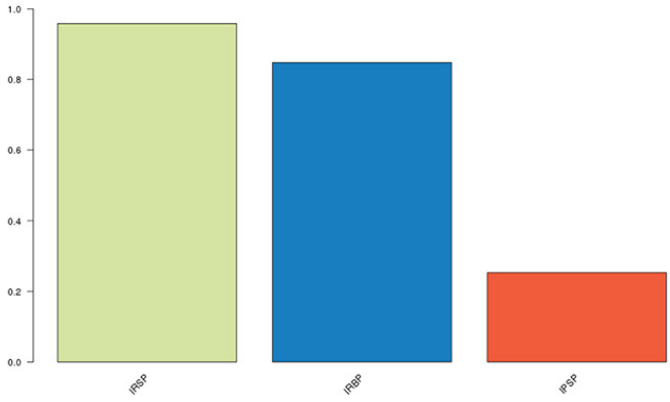
B. Deviations of both spiked (orange) and background (blue) proteins between the quantification and the expected

LogFCs measured by boxplots, the preferred median in boxplot would be ZERO with the minimized deviations.



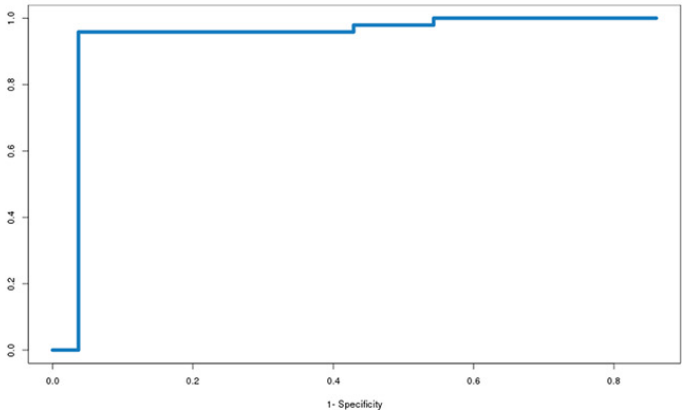
C. Metrics measuring LFQ performance on detecting the spiked and the background proteins

IRSP (identification rate of spiked proteins) = No. of true spiked proteins identified by ROTS / Total No. of true spiked proteins; IRBP (identification rate of background proteins) = No. of background proteins identified by ROTS / Total No. of background proteins; IPSP (identification precision of spiked proteins) = No. of true spiked proteins identified by ROTS / Total No. of spiked proteins found by ROTS. The preferred values of these three metrics would be ONE with the highest performances.



D. Receiver operator characteristic (ROC) curve demonstrating the diagnostic classification accuracy based on the spiked proteins

ROC curve shows the relationship between sensitivity and 1-specificity with the area under the curve denoting the overall diagnostic accuracy of the proteomics data across all thresholds.



2. Various Kinds of Quantification Software for Pre-processing Raw Proteomics Data

ANPELA accepts a variety of data generated by 18 kinds of popular quantification software, all of which aim at pre-processing the raw proteomics data acquired by 3 quantification measurements:

2.1 A List of Software for Pre-processing the Data Acquired Based on SWATH-MS

(software sorted alphabetically)

☐ **DIA-UMPIRE** (*http://diaumpire.sourceforge.net*)

A comprehensive computational workflow and open-source software for processing the data independent acquisition (DIA) mass spectrometry-based proteomics data (Tsou CC, *et al*. *Nat Methods*. 12(3):258-264, 2015). It enables untargeted protein quantification based on the SWATH-MS data obtained by the Orbitrap family of mass spectrometers (Tsou CC, *et al*. *Proteomics*. 16(15-16):2257-2271, 2016), and also enables targeted extraction of quantitative information based on peptides initially identified in only a subset of the samples, resulting in more consistent quantification across multiple samples (Tsou CC, *et al*. *Nat Methods*. 12(3):258-264, 2015). It has been widely used to identify similar number of peptide ions with better identification reproducibility between replicates and samples, than conventional data-dependent acquisition (Bruderer R, *et al*. *Mol Cell Proteomics*. 14(5):1400-1410, 2015). Moreover, it has also been frequently used to process untargeted data for identifying host cell proteins (Kreimer S, *et al*. *Anal Chem*. 89(10):5294-5302, 2017) and to export the peptide identification results of pseudo-MS2 spectra (Wu L, *et al*. *Proteomics*. 16(15-16):2272-2283, 2016). **The resulting file of DIA-UMPIRE accepted by ANPELA** is the "DIAumpire_ProteinSummary_XXXX" file, and the format of which could be readily found **HERE** (Right Click to Save).

☐ **OpenSWATH** (*http://www.openswath.org*)

An open-source software that allows targeted analysis of DIA data based on SWATH-MS in an automated, high-throughput fashion (Röst HL, *et al*. *Nat Biotechnol*. 32(3):219-223, 2014). It is a cross-platform software, written in C++, that relies only on open data formats, allowing it to analyze DIA data from multiple instrument vendors and is integrated and distributed together with OpenMS (Röst HL, *et al*. *Nat Methods*. 13(9):777-783, 2016). It is widely applied to analyze the proteome of streptococcus pyogenes (Röst HL, *et al*. *Nat Biotechnol*. 32(3):219-223, 2014), to estimate q-values of peptide and protein level (Rosenberger G, *et al*. *Nat Methods* 14(9):921-927, 2017). Its generic utility for all types of modification and its scalability could enable confident quantification of the post-translational modifications in DIA-based large-scale studies (Rosenberger G, *et al*. *Nat Biotechnol* 35(8):781-788, 2017). **The resulting file of OpenSWATH accepted by ANPELA** is the OpenSWATH file in *csv* format, and the format of which could be readily found **HERE** (Right Click to Save).

☐ **PeakView** (*https://sciex.com/products/software/peakview-software*)

A commercial software which covers all major components of in-silico processes in a SWATH workflow, from extended assay library building to final statistical analysis and reporting (Li S, *et al*. *J Proteome Res*. 16(2):738-747, 2017; Wu JX, *et al*. *Mol Cell Proteomics*. 15(7):2501-2514, 2016). PeakView uses a set of processing settings to

filter the ion library and determine which peptides or transitions should be used for proteome quantification (Anjo SI, *et al*. *Proteomics*. 17(3-4):1600278, 2017), which is demonstrated to be a powerful strategy particularly for biomarker discovery and clinical field (Anjo SI, *et al*. *Proteomics*. 17(3-4):1600278, 2017). It was used for the N-linked glycoproteins enrichment prior to tryptic digestion, library creation, and analysis (Liu Y, *et al*. *Proteomics*. 13(8):1247-1256, 2013), evaluating the amount of sample needed for PCT-SWATH analysis (Shao S, *et al*. *Proteomics*. 15(21):3711-3721, 2015) and selecting the best method for extracting green algae (Gao Y, *et al*. *Electrophoresis*. 37(10):1270-1276, 2016). **The resulting file of PeakView accepted by ANPELA** is the "ProtSummary_XXXX" file, and the format of which could be readily found **HERE** (Right Click to Save).

☐ **Skyline** (*http://skyline.maccosslab.org*)

A freely-available and open source Windows client application for building selected reaction monitoring, multiple reaction monitoring, parallel reaction monitoring (targeted MS/MS), DIA/SWATH and targeted DDA with MS1 quantitative methods (Broudy D, *et al*. *Bioinformatics*. 30(17):2521-2523, 2014). Skyline was explicitly designed to accelerate targeted proteomics experimentation and foster broad sharing of both methods and results across instrument platforms (MacLean B, *et al*. *Bioinformatics*. 26(7):966-968, 2010). So far, it has been applied to the peptide and transition selection for targeted experiments (Schilling B, *et al*. *Mol Cell Proteomics*. 11(5):202-214, 2012), the retention time determination for scheduled MS experiments (Escher C, *et al*. *Proteomics*. 12(8):1111-1121, 2012) and the isolation window determination for DIA experiments (Zhang Y, *et al*. *J Proteome Res*. 14(10):4359-4371, 2015). **The resulting file of Skyline accepted by ANPELA** is the "Skyline_XXXX_XXXX" file in *tsv* format, and the format of which could be readily found **HERE** (Right Click to Save).

☐ **Spectronaut** (*http://www.spectronaut.org*)

A computational tool for targeted analysis of DIA measurement based on SWATH-MS independent of mass spectrometer vendor (Bruderer R, *et al*. *Proteomics*. 16(15-16):2246-2256, 2016; Bruderer R, *et al*. *Mol Cell Proteomics*. 14(5):1400-1410, 2015). It demonstrates a powerful ability to peak picking and automatic interference correction by utilizing the spectral libraries generated from the raw data acquired on various instrument platforms, and is specifically designed to support spectral-library-free workflow and targeted analysis of OMICs data by hyper reaction monitoring (Navarro P, *et al*. *Nat Biotechnol*. 34(11):1130-1136, 2016; Li S, *et al*. *J Proteome Res*. 16(2):738-747, 2017). It has been widely applied to DIA-based quantitative proteome profiling (Bruderer R, *et al*. *Mol Cell Proteomics*. 14(5):1400-1410, 2015), improved proteomic quantification by sequential window acquisition (Li S, *et al*. *J Proteome Res*. 16(2):738-747, 2017) and high-precision indexed retention time prediction in targeted DIA analysis (Bruderer R, *et al*. *Proteomics*. 16(15-16):2246-2256, 2016). **The resulting file of Spectronaut accepted by ANPELA** is the Spectronaut file in *tsv* format, and the format of which could be readily found **HERE** (Right Click to Save).

2.2 A List of Software for Pre-processing the Data Acquired Based on Precursor Ion Signal Intensity (Peak Intensity)

(software sorted alphabetically)

☐ **MaxQuant** (*http://www.maxquant.org*)

An integrated suite of algorithms specifically developed for processing the high-resolution, quantitative mass-spectrometry data, which is one of the most frequently used platforms for analyzing the MS-based proteome information (Cox J, *et al*. *Nat Biotechnol*. 26(12):1367-1372, 2008; Tyanova S, *et al*. *Nat Protoc*. 11(12):2301-2319, 2016). It is widely used to analyze the tandem spectra generated by collision-induced dissociation (CID), the higher-energy collisional dissociation (HECD) and the electron transfer dissociation (ETD) (Tyanova S, *et al*. *Proteomics*. 15(8):1453-1456, 2015). MaxQuant is used for analyzing data derived from all major relative quantification techniques, including label-free quantification (Tyanova S, *et al*. *Nat Protoc*. 11(12):2301-2319, 2016), MS1-level labeling readouts (Millikin RJ, *et al*. *J Proteome Res*, 2017) and isobaric MS2-level labeling readouts (Weisser H, *et al*. *J Proteome Res*. 12(4):1628-1644, 2013). **The resulting file of MaxQuant accepted by ANPELA** is the "proteinGroups.txt" under a folder named "txt", and the format of which could be readily found **HERE** (Right Click to Save).

☐ **MFPaQ** (*http://mfpaq.sourceforge.net*)

A web-based application that runs on a server on which Mascot Server 2.1 and Perl 5.8 must be installed. To perform quantification, the external module—Extract Daemon is developed to extract intensity values from raw proteomics data (Bouyssié D, *et al*. *Mol Cell Proteomics*. 6(9):1621-1637, 2007). A distinguished key feature of MFPaQ lines in its quantification module, which provides information on protein relative expression following the isotopic labeling and identification with the Mascot. (Gautier V, *et al*. *Mol Cell Proteomics*. 11(8):527-539, 2012). Moreover, it has been applied to the quantitative study of membrane proteins from primary human endothelial cells (Bouyssié D, *et al*. *Mol Cell Proteomics*. 6(9):1621-1637, 2007), and SILAC-based proteomic profiling of the human MDA-MB-231 metastatic breast cancer cell line (Hoedt E, *et al*. *PLoS One*. 9(8):e104563, 2014). **The resulting file of MFPaQ accepted by ANPELA** is the MFPaQ file, and the format of which could be readily found **HERE** (Right Click to Save).

☐ **OpenMS** (*http://www.openms.de*)

A robust, open-source, cross-platform software specifically designed for the flexible and reproducible analysis of high-throughput MS data (Sturm M, *et al*. *BMC Bioinformatics*. 9:163, 2008; Weisser H, *et al*. *J Proteome Res*. 12(4):1628-1644, 2013). It uses modern software engineering concepts with an emphasis on modularity, reusability and extensive testing using continuous integration, and implements common mass spectrometric data processing tasks through a well-defined application programming interface and through the standardized open data format. (Röst HL, *et al*. *Nat Methods*. 13(9):741-748, 2016). OpenMS is widely applied to the quantitative and variant enabled mapping of peptides to genomes (Schlaffner CN, *et al*. *Cell Syst*. 5(2):152-156, 2017), the analysis of cerebrospinal fluid proteome in alzheimer's (Khoonsari PE, *et al*. *PLoS One*. 11(3):e0150672 2016) and quantitative analysis of label-free LC-MS data for comparative candidate biomarker studies (Hoekman B, *et al*. *Mol Cell Proteomics*. 11(6):M111, 2012). **The resulting file of OpenMS accepted by ANPELA** is the OpenMS file, and the format of which could be readily found **HERE** (Right Click to Save).

☐ **PEAKS** (*http://www.bioinfor.com*)

A software platform with complete solution for discovery proteomics, including the protein identification and quantification, analysis of posttranslational modification and sequence variants, and peptide/protein de novo sequencing (Ma B, *et al*. *Rapid Commun Mass Spectrom*. 17(20):2337-2342, 2003). It relies on a sophisticated dynamic programming algorithm to efficiently compute the best peptide sequences whose fragment ions can best interpret the peaks in the MS/MS spectrum. It is thus a useful tool for the analysis of protein identification and quantification of known and unknown genomes (Ma B, *et al*. *Rapid Commun Mass Spectrom*. 17(20):2337-2342, 2003). PEAKS has matured into a comprehensive proteomics platform supporting the analysis of label-free data and label-based data, such as TMT(MS2,MS3)/iTRAQ, SILAC, ICAT and so on. It achieves significantly improved accuracy and sensitivity over other commonly applied software packages (Zhang J, *et al*. *Mol Cell Proteomics*. 11(4):M111, 2012). **The resulting file of PEAKS accepted by ANPELA** is the "proteins.csv" file under a folder named "PEAKS XXXX", and the format of which could be readily found **HERE** (Right Click to Save).

☐ **Progenesis** (*http://www.nonlinear.com/progenesis*)

A new generation of bioinformatics vehicle targeting small molecule discovery analysis for metabolomics and proteomics, which quantifies proteins based on peptide ion signal peak intensity (Zhang J, *et al*. *Anal Bioanal Chem*. 408(14):3881-3890, 2016). Progenesis allows full operator control over every processing step including alignment of peptide ion signal landscapes and indeed individual peptide ion signal peaks (Al Shweiki MR, *et al*. *J Proteome Res*. 16(4):1410-1424, 2017). It can be used for protein label-free quantification and peak picking with the automatic sensitivity method, that uses a noise estimation algorithm to determine the noise level of the data (Almeida AM, *et al*. *J Proteomics*. 152:206-215, 2017). **The resulting file of Progenesis accepted by ANPELA** is the Progenesis file in *csv* format, and the format of which could be readily found **HERE** (Right Click to Save).

☐ **Proteios SE** (*http://www.proteios.org*)

ProSE integrates protein identification search engine access into several proteomic workflows, both gel-based and liquid chromatography-based, and allows seamless combination of search results, protein inference, protein annotation and quantitation tools (Gärdén P, *et al*. *Bioinformatics*. 21(9):2085-2087, 2005). It is targeted for large projects with shared data, integrates sample tracking and aims at becoming a standard analysis platform for proteomics, whose major feature is the automated linking of data from different parts of proteomic workflows (Häkkinen J, *et al*. *J Proteome Res*. 8(6):3037-3043, 2009). It has built-in support to several protein identification engines such as Mascot, X!Tandem, and combines search results from multiple search engines (Végvári A, *et al*. *Mol Cell J Proteomics*. 75(1):202-210, 2011), and automatically generates the protein identification reports containing information required for publication of proteomics results (Levander F, *et al*. *Proteomics*. 7(5):668-674, 2007). **The resulting file of Proteios SE accepted by ANPELA** is the Proteios SE file, and the format of which could be readily found **HERE** (Right Click to Save).

☐ **Scaffold** (*http://www.proteomesoftware.com/products/scaffold*)

A commercial bioinformatic tool, which attempts to increase the confidence in protein identification reports through the use of several statistical methods (Searle BC. *Proteomics*. 10(6):1265-9, 2016). It supports a wide variety of search engines and uses a pipeline of several peptide and protein validation methods after an initial database-search analysis (Codrea MC, *et al*. *Adv Exp Med Biol*. 919:203-215, 2016). Scaffold has been widely applied to the identification of proteome for a new target to inhibit yellow fever virus replication (Vidotto A, *et al*. *J Proteome Res*. 16(4):1542-1555, 2017), analysis of the follicle fluid proteome for preconception folic acid use (Twigt JM, *et al*. *Eur J Clin Invest*. 45(8):833-41, 2015) and identifiable analysis of effects of cadmium exposure on the gill proteome of Cottus gobio (Dorts J, *et al*. *Aquat Toxicol*. 154:87-96, 2014). **The resulting file of Scaffold accepted by ANPELA** is the "scaffoldXXXX" file, and the format of which could be readily found **HERE** (Right Click to Save).

☐ **Thermo Proteome Discoverer** (*http://thermo-msf-parser.googlecode.com*)

A software for workflow-driven data analysis in proteomics integrating all different steps in a quantitative proteomics experiment (MS/MS spectrum extraction, peptide identification and quantification) into the user-configurable, automated workflows (Colaert N, *et al*. *J Proteome Res*. 10(8):3840-3843, 2011; Veit J, *et al*. *J Proteome Res*. 15(9):3441-3448, 2016). It has a convenient graphical user interface in which users can load raw data directly from the instrument and explore and analyze it because it supports multiple sequence database search engines (e.g., *Sequest HT*, *Mascot*), the spectral library searching, the peptide spectrum-match validation (e.g., *Percolator*), as well as various quantification techniques, like isobaric mass tagging (e.g., *iTRAQ*, *TMT*) or SILAC (Veit J, *et al*. *J Proteome Res*. 15(9):3441-3448, 2016). Proteome Discoverer is applied to the iTRAQ (isobaric tag for relative and absolute quantitation)-based quantitative analysis of protein mixtures (Casado-Vela J, *et al*. *Proteomics*. 10(2):343-347, 2010). **The resulting file of Thermo Proteome Discoverer accepted by ANPELA** is the Proteome Discoverer file, and the format of which could be readily found **HERE** (Right Click to Save).

2.3 A List of Software for Pre-processing the Data Acquired Based on Spectral Counting

(software sorted alphabetically)

☐ **Abacus** (*http://abacustpp.sourceforge.net*)

A computational tool for extracting and preprocessing spectral count data for label-free quantitative proteomic analysis (Fermin D, *et al*. *Proteomics*. 11(7):1340-1345, 2011). It aims at streamlining analysis of spectral count data by providing an automated, easy to use solution which extracts the information from proteomic datasets for subsequent statistical analysis (Fermin D, *et al*. *Proteomics*. 11(7):1340-1345, 2011). However, the approach has the disadvantage of losing information or attempting to apportion large numbers of spectra on the basis of relatively small sets of differentiating spectra (Chen YY, *et. al*. *Anal Bioanal Chem*. 404(4):1115-1125, 2012). It is compatible with the widely used trans-proteomic pipeline suite of tools and comes with a graphical user interface making it easy to interact with the program (Fermin D, *et al*. *Proteomics*. 11(7):1340-1345, 2011). **The resulting file of Abacus accepted by ANPELA** is the "abacus_data output.csv" file under a folder named "abacus_data", and the format of which could be readily found **HERE** (Right Click to Save).

☐ **Census** (*http://fields.scripps.edu/census*)

A quantitative software tool which can analyze high-throughput mass spectrometry data from shotgun proteomics experiments in an efficient way and various stable isotope labeling experiments (e.g., 15N, 18O, SILAC, iTRAQ and TMT) in addition to the labeling-free experiments (Park SK, *et al*. *Curr Protoc Bioinformatics*. Chapter 13:Unit 13.12.1-11, 2010). What makes Census differentiated most from other quantitative tools is its flexibility to handle most types of quantitative proteomics labeling strategies, as well as label-free experiment with multiple statistical algorithms to improve quality of results (Park SK, *et al*. *Nat Methods*. 5(4):319-322, 2008). Census can be used for large-scale differential proteome analysis in plasmodium falciparum under drug treatment (Prieto JH, *et al*. *PLoS One*. 3(12):e4098, 2008), and proteomic analysis of protein turnover by metabolic whole rodent pulse-chase isotopic labeling (Savas JN, *et al*. *Methods Mol Biol*. 1410:293-304, 2016). **The resulting file of Census accepted by ANPELA** is the Census file, and the format of which could be readily found **HERE** (Right Click to Save).

☐ **DTASelect** (*http://fields.scripps.edu/DTASelect*)

A Java tool that is used to organize, filter, and interpret results generated by SEQUEST (one of the most widely used protein database searching programs for tandem mass spectrometry) (Cociorva D, *et al*. *Curr Protoc Bioinformatics*. Chapter 13:Unit 13.4, 2007). It assembles protein-level information from peptide data and focuses on peptides of interest by sweeping away the less likely identification (Cociorva D, *et al*. *Curr Protoc Bioinformatics*. Chapter 13:Unit 13.4, 2007). It makes more complex experiments feasible by streamlining data analysis for proteomics (Tabb DL, *et al*. *J Proteome Res*. 1(1):21-6, 2002). It can be used for a proteogenomic study with a controlled protein false discovery rate (Park GW, *et al*. *J Proteome Res*. 15(11):4082-4090, 2016), and data analysis of palmitoylated protein identifications (Wan J, *et al*. *Nat Protoc*. 2(7):1573-1584, 2007). **The resulting file of DTASelect accepted by ANPELA** is the DTASelect file, and the format of which could be readily found **HERE** (Right Click to Save).

☐ **IRMa-hEIDI** (*http://biodev.extra.cea.fr/docs/irma*)

The IRMa toolbox provides an interactive application to assist in the validation of Mascot search results, and allows automatic filtering of Mascot identification results as well as manual confirmation or rejection of individual PSM (a match between a fragmentation mass spectrum and a peptide) (Dupierris V, *et al*. *Bioinformatics*. 25(15):1980-1981, 2009). Its main originality is to filter matches rather than identified proteins and its features are easy navigation within identification result and batch mode to automatically validate multiple identification results (Dupierris V, *et al*. *Bioinformatics*. 25(15):1980-1981, 2009). The IRMa-hEIDI is used to filter the spectral count workflows results of several label-free bioinformatics tools, coupling Mascot peptide identification with IRMa validation and hEIDI grouping and comparison ended up with the best compromise between sensitivity and false discovery proportion (Ramus C, *et al*. *J Proteomics*. 132:51-62, 2016; Ramus C, *et al*. *Data Brief*. 6:286-294, 2015). **The resulting file of IRMa-hEIDI accepted by ANPELA** is the IRMa-hEIDI file, and the format of which could be readily found **HERE** (Right Click to Save).

☐ **MaxQuant** (*http://www.maxquant.org*)

An integrated suite of algorithms specifically developed for processing high-resolution, quantitative MS data, which has been keeping up with recent advances in high-resolution instrumentation and with the development of fragmentation techniques (Cox J, *et al*. *Nat Biotechnol*. 26(12):1367-1372, 2008; Tyanova S, *et al*. *Nat Protoc*. 11(12):2301-2319, 2016). It has matured into a comprehensive proteomics platform supporting the analysis of MS data generated by the MS systems from most vendors, and integrated a multitude of algorithms (Tyanova S, *et al*. *Proteomics*. 15(8):1453-1456, 2015). It substantially improves mass precision as well as mass accuracy (Neuhauser N, *et al*. *Mol Cell Proteomics*. 11(11):1500-1509, 2012). It can be used for identification of ubiquitylation and SUMOylation sites after the ubiquitylated peptides and SCX-fractionated SUMO peptides which were analyzed separately by LC–MS/MS (McManus FP, *et al*. *Nat Protoc*. 12(11):2342-2358, 2017), comprehensive profiling of pancreatic tissue proteome (Liu CW, *et al*. *Methods Mol Biol*. doi: 10.1007/7651_2017_77, 2017) and label-free quantitative analysis on the cisplatin resistance in ovarian cancer cells (Wang F, *et al*. *Cell Mol Biol*. 63(5):25-28, 2017). **The resulting file of MaxQuant accepted by ANPELA** is the "proteinGroups.txt" under a folder named "txt", and the format of which could be readily found **HERE** (Right Click to Save).

☐ **MFPaQ** (*http://mfpaq.sourceforge.net*)

A software tool that facilitates organization, mining, and validation of Mascot results and offers different functionalities to work on validated protein lists, as well as data quantification using isotopic labeling methods or label free approaches (Bouyssié D, *et al*. *Mol Cell Proteomics*. 6(9):1621-1637, 2007). MFPaQ extracts quantitative data from raw files obtained by nano-LC-MS/MS, calculates peptide ratios, and generates a non-redundant list of proteins identified in a multisearch experiment with their calculated averaged and normalized ratio (Bouyssié D, *et al*. *Mol Cell Proteomics*. 6(9):1621-1637, 2007). It has been applied to the large scale analysis of the human inflammatory endothelial cells (Gautier V, *et al*. *Mol Cell Proteomics*. 11(8):527-539, 2012), and the label-free quantification of cerebrospinal fluid by combining peptide ligand library treatment (Mouton-Barbosa E, *et al*. *Mol Cell Proteomics*. 9(5):1006-1021, 2010). **The resulting file of MFPaQ accepted by ANPELA** is the MFPaQ file, and the format of which could be readily found **HERE** (Right Click to Save).

☐ **ProteinProphet** (*http://proteinprophet.sourceforge.net*)

A statistical model which is designed for computing probabilities that proteins are present in a sample on the basis of peptides assigned to tandem mass (MS/MS) spectra acquired from a proteolytic digest of the sample (Nesvizhskii AI, *et al*. *Anal Chem*. 75(17):4646-4658, 2003). It allows the filtering of large-scale proteomics data with predictable sensitivity and false positive identification error rates (Nesvizhskii AI, *et al*. *Anal Chem*. 75(17):4646-4658, 2003). It was used to discriminate true assignments of MS/MS spectra to peptide sequences from false assignments, to assign a probability value for each identified peptide, and to compute sensitivity and error rate for the assignment of spectra to sequences in each experiment (Keller A, *et al*. *Anal Chem*. 74(20):5383-5392, 2002). It was also used to infer the protein identifications and to compute probabilities that a protein had been correctly identified, based on the available peptide sequence evidence (Nesvizhskii AI, *et al*. *Anal Chem*. 75(17):4646-4658, 2003; Yan W, *et al*. *Mol Cell Proteomics*. 3(10):1039-1041, 2004). **The resulting file of ProteinProphet accepted by ANPELA** is the ProteinProphet file, and the format of which could be readily found **HERE** (Right Click to Save).

☐ **Scaffold** (*http://www.proteomesoftware.com/products/scaffold*)

A feature-rich software suite to assist in analysis, visualization, quantification, annotation and validation of complex LC-MS/MS experiments (Codrea MC, *et al*. *Adv Exp Med Biol*. 919:203-215, 2016). It exports the sample reports to Microsoft Excel, and relevant information and annotations for each protein were searched from databases including Swiss-Prot, Human Protein Reference Database, Entrez Gene, and the Plasma Proteome Database (Cho CK, *et al*. *Mol Cell Proteomics*. 6(8):1406-1415, 2007). Scaffold was applied for the tryptic peptide product ion MS2 spectral processing, false discovery rate assessment, and protein identification (Garbis SD, *et al*. *Anal Chem*. 83(3):708-718, 2011). It has been widely used to the label free quantitation and labeled quantitation in breast cancer patients with thick white or thick yellow tongue fur (Cao MQ, *et al*. *Zhong Xi Yi Jie He Xue Bao*. 9(3):275-280, 2011), and the analysis of the follicle fluid proteome in preconception folic acid use (Twigt JM, *et al*. *Eur J Clin Invest*. 45(8):833-841, 2015). **The resulting file of Scaffold accepted by ANPELA** is the "scaffoldXXXX" file, and the format of which could be readily found **HERE** (Right Click to Save).

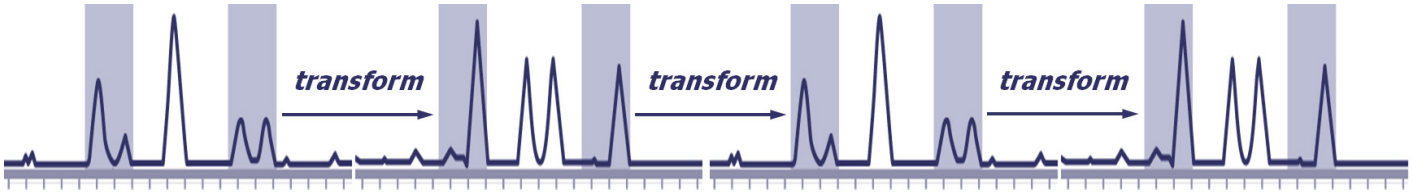
---

3. A Variety of Methods for Data Manipulation at Different Manipulation Stages

Users are provided with the option to conduct transformation, pretreatment and imputation on their uploaded data. In total, 3 transformation, 18 pretreatment and 7 imputation methods frequently applied to manipulate the label-free proteomics data are provided in the current version of ANPELA.

3.1 Methods for Data Transformation

(methods sorted alphabetically)



☐ **Box-cox Transformation**

Box & Cox proposed a parametric power transformation technique in order to reduce anomalies such as non-additivity, non-normality and heteroscedasticity (Sakia RM. *The Statistician*. 41:169-178, 1992). This transformation has been extensively studied, and an attempt is made to review the corresponding studies relating to this transformation (Sakia RM. *The Statistician*. 41:169-178, 1992).

☐ **Log Transformation**

Log transformation was carried out almost routinely for obtaining a more symmetric distribution prior to statistical analysis (De Livera AM, *et al*. *Anal Chem*. 84(24):10768-10776, 2012). It works for data where you can see that the residuals get bigger for bigger values of the dependent variable (De Livera AM, *et al*. *Anal*

Chem. 84(24):10768-10776, 2012). Such trends in the residuals occur often, because the error or change in the value of an outcome variable is often a percent of the value rather than an absolute value (De Livera AM, *et al*. *Anal Chem*. 84(24):10768-10776, 2012).

☐ Variance Stabilization Normalization

Variance Stabilization Normalization (VSN) is one of the non-linear methods aiming to keep the variance constant over the entire data range (Huber W, *et al*. *Bioinformatics*. 18 S1:96-104, 2002; Kohl SM, *et al*. *Metabolomics*. 8(S1):146-160, 2012). VSN approaches the logarithm for large values to remove heteroscedasticity using the inverse hyperbolic sine (Kohl SM, *et al*. *Metabolomics*. 8(Suppl 1):146-160, 2012). For small intensities, it performs linear transformation behavior to make the variance unchanged (Kohl SM, *et al*. *Metabolomics*. 8(Suppl 1):146-160, 2012). VSN was originally developed as normalization method for label-free relative quantification of endogenous peptides (Kultima K, *et al*. *Mol Cell Proteomics*. 8(10):2285-2295, 2009).

3.2 Methods for Data Pretreatment

Pretreatment Methods include 2 centering methods, 4 scaling methods and 12 normalization methods.

3.2.1 Methods for Data Centering

(methods sorted alphabetically)

☐ Mean Centering

Used for facilitating the improvement of the sensitivity of significance test in spectral counting-based comparative discovery proteomics (Gregori. J, *et al*. *J Proteomics*. 75(13):3938-51, 2012).

☐ Median Centering

Facilitating the normalization procedures in LC-MS proteomics experiments through dataset dependent ranking of normalization scaling factors (Webb-Robertson. B. J, *et al*. *Proteomics*. 2011, 11(24): 4736-41).

3.2.2 Methods for Data Scaling

(methods sorted alphabetically)

☐ Auto Scaling

Auto Scaling (Unit Variance Scaling, UV) is one of the simplest methods adjusting proteomics variances, which scales protein intensities based on the standard deviation of proteomics data (Kohl SM, *et al*. *Metabolomics*. 8(Suppl 1):146-160, 2012). This method scales all protein intensities to unit variance, and all intensities are equally important and comparably scaled (Gromski PS, *et al*. *Metabolomics*. 11:684-695, 2015). The data is analyzed on the basis of correlations and standard deviation of all intensities, but the disadvantage of auto scaling is that analytical errors may be amplified due to dilution effects (Kohl SM, *et al*. *Metabolomics*. 8(Suppl 1):146-160, 2012). Auto scaling has been used to identify proteomic biomarkers for psoriasis and psoriasis arthritis (Reindl J, *et al*. *J Proteomics*. 140:55-61, 2016) and normalize LC-MS proteomics data based on scan-level information (Nezami Ranjbar MR, *et al*. *Proteome Sci*. 11(Suppl 1):S13, 2013).

☐ Pareto Scaling

Pareto Scaling uses the square root of the standard deviation of the data as scaling factor (Kohl SM, *et al*. *Metabolomics*. 8(Suppl 1):146-160, 2012). This method is able to reduce the weight of large fold changes in protein intensities, which is more significantly than auto scaling (Kohl SM, *et al*. *Metabolomics*. 8(Suppl 1):146-160, 2012). But the dominant weight of extremely large fold changes may still be unchanged (Kohl SM, *et al*. *Metabolomics*. 8(Suppl 1):146-160, 2012). Therefore, the disadvantage of pareto scaling is the sensitivity to large fold changes (Van den Berg RA, *et al*. *BMC Genomics*. 7:142, 2006). Pareto scaling was applied to normalize LC-MS proteomics data using scan-level information in the Gaussian process regression model (Nezami Ranjbar MR, *et al*. *Proteome Sci*. 11(Suppl 1):S13, 2013).

☐ Range Scaling

Range scaling uses this biological range as the scaling factor (Smilde AK, *et al*. *Anal Chem* 2005, 77:6729-6736). A disadvantage of range scaling with regard to the other scaling methods tested is that only two values are used to estimate the biological range, while for the standard deviation all measurements are taken into account. This makes range scaling more sensitive to outliers. To increase the robustness of range scaling, the range could also be determined by using robust range estimators. Manipulating the non-targeted ultra-high performance liquid chromatography tandem mass spectrometry (UHPLC-MS) proteomic/metabolomic data (Guida R. Di, *et a*l. *Metabolomics*. 2016, 1293).

☐ Vast Scaling

Vast scaling is an acronym of variable stability scaling and it is an extension of autoscaling (Keun HC, *et al*. *Anal Chim Acta*. 2003, 490:265-276). It focuses on stable variables, the variables that do not show strong variation, using the standard deviation and the so-called coefficient of variation (cv) as scaling factors. Vast scaling can be used unsupervised as well as supervised. When vast scaling is applied as a supervised method, group information about the samples is used to determine group specific cvs for scaling. Assessing the impact of delayed storage on the measured proteome and metabolome of human cerebrospinal fluid (Rosenling T, *et al*. *Clin Chem*. 2011, 57(12): 1703-11).

3.2.3 Methods for Data Normalization

(methods sorted alphabetically)

☐ Cyclic Loess

Cyclic Loess (Cyclic Locally Weighted Regression) originates from the combination of MA-plot and logged Bland-Altman plot by assuming the existence of non-linear bias (Kohl SM, *et al*. *Metabolomics*. 8(Suppl 1):146-160, 2012), and can estimate a regression surface using multivariate smoothing procedure (Webb-Robertson BJ, *et al*. *Metabolomics*. 10(5):897-908, 2014). However, cyclic loess is one of the most time-consuming one among the normalization methods, and the amount of time grows exponentially as the number of sample increases (Ballman KV, *et al*. *Bioinformatics*. 20(16):2778-86, 2004). Cyclic loess has been applied in proteomics profiling in the context of common experimental designs (Keeping AJ, *et al*. *J Proteome Res*. 10(3):1353-60, 2011).

☐ EigenMS

EigenMS removes bias of unknown complexity from the Liquid Chromatography coupled with Mass Spectrometry (LC/MS)-based proteomics data, allowing for increased sensitivity in differential analysis. EigenMS normalization aims at preserving the original differences while removing the bias from the data (Välikangas T, *et al*. *Brief Bioinform*. pii: bbw095, 2016). It works by 3 steps (Karpievitch YV, *et al*. *PLoS One*. 9(12):e116221, 2014): (1) EigenMS preserves the true differences in the proteomics data by estimating treatment effects with an ANOVA model; (2) singular value decomposition of the residuals matrix is used to determine bias trends in the data; (3) the number of bias trends is estimated via a permutation test and the effects of the bias trends are eliminated. EigenMS has been applied in the profiling of MS-based quantitative label-free proteomics and LC-based proteomics (Karpievitch YV, *et al*. *BMC Bioinformatics*. 13(S16):S5, 2012; Karpievitch YV, *et al*. *Ann Appl Stat*. 4(4):1797-1823,2010).

☐ Linear Baseline

Linear Baseline (Linear Baseline Scaling) maps each spectrum to the baseline based on the assumption of a constant linear relationship between each feature of a given spectrum and the baseline (Kohl SM, *et al*. *Metabolomics*. 8(Suppl 1):146-160, 2012). The baseline is the median of each feature across all spectra and the scaling factor is computed as the ratio of the mean intensity of the baseline to the mean intensity of each spectrum (Kohl SM, *et al*. *Metabolomics*. 8(Suppl 1):146-160, 2012). The intensities of all spectra are multiplied by their particular scaling factors (Kohl SM, *et al*. *Metabolomics*. 8(Suppl 1):146-160, 2012). However, this assumption of a linear correlation among sample spectra may be oversimplified (Kohl SM, *et al*. *Metabolomics*. 8(Suppl 1):146-160, 2012).

☐ Locally Weighted Scatterplot Smoothing

Locally Weighted Scatterplot Smoothing (Lowess) is a method used to normalize a two-color array gene expression dataset to compensate for non-linear dye-bias. In this approach, the log-ratio for each sample is adjusted by the lowess fitted value (Yang YH, *et al*. *Proc Spie*. 6(10):1-21, 2003). Lowess normalization assumes that the dye bias appears to be dependent on spot intensity (Yang YH, *et al*. *Proc Spie*. 6(10):1-21, 2003). Lowess normalization can be applied to complete or incomplete datasets and may be applied to a two-color array expression dataset (Yang YH, *et al*. *Proc Spie*. 6(10):1-21, 2003). This method has been used in MS-based proteomics (Karpievitch YV, *et al*. *BMC Bioinformatics*. 13(S16):S5, 2012)

☐ Mean

Mean Normalization normalizes the data by mean value of all signals to eliminate background effect (Andjelkovic V, *et al*. *Plant Cell Rep*. 25(1):71-9, 2006). Intensity of each protein in a given sample is used by the mean of intensity of all variables in the sample (De Livera AM, *et al. Anal Chem*. 84(24):10768-10776, 2012). In order to make the samples comparable, the means of the intensities for each experimental run are forced to be equal to one another using this method (Ejigu BA, *et al*. *OMICS*. 17(9):473-485, 2013). For example, each sample is scaled such that the mean of all abundances in a sample equals one (De Livera AM, *et al*. *Anal Chem*. 84(24):10768-10776, 2012). This method has been used in the profiling of urine peptidome (Padoan A, *et al*. *Proteomics*. 15(9):1476-1485, 2015).

☐ Median

Median normalization is based on the assumption that the samples of a dataset are separated by a constant. It scales the samples so that they have the same median. For example, the median of the protein intensities in the sample equals one (Bolstad BM, et al. Bioinformatics. 19(2):185-93, 2003). The median normalization, the commonly used method without the need for internal standards, is more practical than the sum normalization especially in situations where several saturated abundances may be associated with some of the factors of interest (Bolstad BM, et al. Bioinformatics. 19(2):185-93, 2003). It has previously been used in MS-based label-free proteomics analysis for removing systematic biases associated with mass spectrometry (Callister SJ, et al. J Proteome Res. 5(2):277-86, 2006).

☐ Median Absolute Deviation

The Median Absolute Deviation (MAD) is a robust measure of the spread of the data, and is used as an estimate of the sample standard deviation if scaled by a factor of 1.483, and it is a simple way to quantify variation (Matzke MM, *et al*. *Bioinformatics*. 27(20):2866-2872, 2011). This method has been used to improve the quality control process of peptide-centric LC-MS proteomics data (Matzke MM, *et al*. *Bioinformatics*. 27(20):2866-2872, 2011).

☐ Probabilistic Quotient Normalization

PQN (Probabilistic Quotient Normalization) transforms the proteomics spectra according to an overall estimation on the most probable dilution (Dieterle F, *et al. Anal Chem*. 78(13):4281-4290, 2006). This algorithm has been reported to be significantly robust and accurate comparing to the integral and the vector length normalizations (Dieterle F, *et al. Anal Chem*. 78(13):4281-4290, 2006). There are three steps in the procedure of PQN24: (1) perform an integral normalization of each spectrum, then select a reference spectrum such as the median spectrum; (2) calculate the quotient between a given test spectrum and reference spectrum, then estimate the median of all quotients for each variable; (3) all variables of the test spectrum are divided by the median quotient. PQN has been applied in MALDI-TOF mass spectrometry knowledge discovery (López-Fernández H, *et al*. *BMC Bioinformatics*. 16:318, 2015).

☐ Quantile

Quantile (Quantile Normalization) aims at achieving the same distribution of protein intensities across all samples, and the quantile-quantile plot in this method is used to visualize the distribution similarity (Kohl SM, *et al*. *Metabolomics*. 8(Suppl 1):146-160, 2012). Quantile normalization is motivated by the idea that the distribution of two data vectors is the same if the quantile-quantile plot is a straight diagonal line (Bolstad BM, *et al. Bioinformatics*. 19(2):185-93, 2003). While a common and non-data driven distribution is generated using quantile normalization, an agreed standard could not be reached (Bolstad BM, *et al. Bioinformatics*. 19(2):185-93, 2003). Quantile normalization has been adopted for removing systematic biases associated with mass spectrometry and label-free proteomics (Callister SJ, *et al*. *J Proteome Res*. 5(2):277-86, 2006).

☐ Robust Linear Regression

Robust Linear Regression is used for transference: when you want to rescale one reference interval to another scale. The robust linear regression is more robust against outliers in the data than linear regression using least squares estimation (Välikangas T, *et al*. *Brief Bioinform*. pii: bbw095, 2016). This method has been used to minimize plate effects of suspension bead array data (Hong MG, *et al*. *J Proteome Res*. 15(10):3473-3480, 2016).

☐ Tol Ion Current

Tol Ion Current sum all the separate ion currents carried by the ions of different m/z contributing to a complete mass spectrum or in a specified m/z range of a mass spectrum. And the sum of all peak areas of peptides unique to a particular organism was here called pTIC (proteome total ion current) (Gaspari M, *et al*. *Anal Chem*. 88(23):11568-11574, 2016). This method has been used in MALDI-TOF and SELDI-TOF mass spectra proteomic profiling (Borgaonkar SP, *et al*. *OMICS*. 14(1):115-26, 2010).

☐ Trimmed Mean of M Values

Trimmed Mean of M Values (TMM) normalization is a simple and effective method for estimating relative RNA production levels from RNA-seq data (Lin Y, *et al*. *BMC Genomics*. 17:28, 2016). It estimates scale factors between samples that can be incorporated into currently used statistical methods for differential expression analysis (Lin Y, *et al*. *BMC Genomics*. 17:28, 2016). The Trimmed Mean of M-values normalization methods were sensitive to the removal of low-expressed genes from the data set in RNA-Seq data (Lin Y, *et al*. *BMC Genomics*. 17:28, 2016).

3.3 Methods for Missing Value Imputation

(methods sorted alphabetically)

☐ **Background Imputation (BACK)**

All missing values were replaced with the lowest detected intensity value of the data set. This imputation simulates the situation where protein values are missing because of having small concentrations in the sample and thus cannot be detected during the MS run (Chai LE, *et al*. *Malays J Med Sci*. 21(2):20-2, 2014). The lowest intensity value detected is therefore imputed for the missing protein values as a representative of the background (Chai LE, *et al*. *Malays J Med Sci*. 21(2):20-2, 2014).

☐ **Bayesian Principal Component Imputation (BPCA)**

Bayesian Principal Component Imputation (BPCA) out-performs the kNN and SVD imputation methods (Chai LE, *et al*. *Malays J Med Sci*. 21(2):20-22, 2014). One of the features of BPCA that allows it to provide a better performance than the latter two methods is its capacity to auto-select the parameters used in the estimation (Chai LE, *et al*. *Malays J Med Sci*. 21(2):20-22, 2014). This method also produces improved estimation performance when the number of the samples is huge (Chai LE, *et al*. *Malays J Med Sci*. 21(2):20-22, 2014).

☐ **Censored Imputation (CENSOR)**

If only a single NA for a protein in a sample group was found, it was considered as being 'missing completely at random', and no value was imputed for it (Välikangas T, *et al*. *Brief Bioinform*. doi: 10.1093/bib/bbx054, 2017). If a protein contained more than one missing value in a sample group (consisting of technical replicates), they were considered missing because of being below detection capacity, and the lowest intensity value in the data set was imputed for them. (Välikangas T, *et al*. *Brief Bioinform*. doi: 10.1093/bib/bbx054, 2017).

☐ **K-nearest Neighbor Imputation (KNN)**

The kNN impute method aims to identify k genes that are very similar to the genes with missing values, where the similarity is estimated by the Euclidean distance measure, and the missing values are imputed with the values of weighted average from these neighbouring genes (Chai LE, *et al*. *Malays J Med Sci*. 21(2):20-2, 2014). KNN-based methods tend to select genes with expression profiles similar to the gene of interest to impute missing values, and KNN outperforms BPCA and LLS with relatively small size datasets (Chai LE, *et al*. *Malays J Med Sci*. 21(2):20-2, 2014).

☐ **Local Least Squares Imputation (LLS)**

Local Least Squares Imputation (LLS) exploits the local similarity structures in the data, as well as the least squares optimisation process (Chai LE, *et al*. *Malays J Med Sci*. 21(2):20-2, 2014). The proposed local least squares imputation method (LLSimpute) represents a target gene that has missing values as a linear combination of similar genes (Kim H, *et al*. *Bioinformatics*. 21(2):1-12, 2004). The similar genes are chosen by k-nearest neighbors or k coherent genes that have large absolute values of Pearson correlation coefficients. Nonparametric missing values estimation method of LLSimpute are designed by introducing an automatic k-value estimator (Kim H, *et al*. *Bioinformatics*. 21(2):1-12, 2004).

☐ **Singular Value Decomposition (SVD)**

Singular value decomposition (SVD) SVD is also known as Karhunen–Loève expansion in pattern recognition and as principal-component analysis in statistics (Alter O, *et al*. *PNAS*. 97(18):10101-10106, 2000). SVD is a linear transformation of the expression data from the genes × arrays space to the reduced "eigengenes" × "eigenarrays" space (Alter O, *et al*. *PNAS*. 97(18):10101-10106, 2000). In contrast to the KNN imputation which utilizes local pairwise information between genes in the gene expression matrix, SVD imputation attempts to utilize the global information in the entire matrix in predicting the missing values (Gan X, *et al*. *Nucleic Acids Res*. 34(5):1608-1619, 2006). The basic concept about this method is to find the dominant components summarizing the entire matrix and then to predict the missing values in the target genes by regressing against the dominant components (Gan X, *et al*. *Nucleic Acids Res*. 34(5):1608-1619, 2006).

☐ **Zero Imputation (ZERO)**

The simplest imputation method is by replacing the missing values with zeros. This zero replacement method does not utilize any information about the data (Gan X, *et al*. *Nucleic Acids Res*. 34(5):1608-1619, 2006). In fact, the integrity and usefulness of the data can be jeopardized by zero imputation since erroneous relationships between genes can be artificially created due to the imputation (Gan X, *et al*. *Nucleic Acids Res*. 34(5):1608-1619, 2006).

---

4. Diverse MS Systems for Proteome Quantification

Those popular kinds of software listed in the Section 2 of this Manual aim at quantifying the raw proteomics data derived from a diverse set of MS systems including the AB SCIEX Q-TOF systems, the Agilent Q-TOF mass spectrometer, the Bruker hybrid Q-TOF mass spectrometer and the Thermo Fisher Scientific Orbitrap.

4.1 AB SCIEX Q-TOF Systems

☐ AB SCIEX QTRAP Systems (QTRAP 6500+ System, QTRAP 6500 System, QTRAP 5500 System, QTRAP 4500 System, QTRAP 4000 System, QTRAP 3200 System)

☐ AB TOF/TOF Systems (TOF/TOF 5800 System)

☐ AB Triple Quad Systems (Triple Quad 6500+ System, Triple Quad 6500 System, Triple Quad 5500 System, Triple Quad 4500 System, Triple Quad 3500 System, API 4000 System, API 3200 System)

☐ TripleTOF Systems (TripleTOF 6600 System, TripleTOF 5600+ System, TripleTOF 4600 System)

☐ X-Series QTOF Systems (X500B QTOF system, X500R QTOF System)

4.2 Agilent Q-TOF

☐ Agilent 6530 Accurate-Mass Quadrupole Time-of-Flight LC/MS system

4.3 Bruker Hybrid Q-TOF Mass Spectrometer

4.4 Thermo Fisher Scientific Orbitrap

---

5. References

Al Shweiki MR, et al. Assessment of Label-Free Quantification in Discovery Proteomics and Impact of Technological Factors and Natural Variability of Protein Abundance. J Proteome Res. 16(4):1410-1424, 2017

Almeida AM, et al. The longissimus thoracis muscle proteome in Alentejana bulls as affected by growth path. J Proteomics. 152:206-215, 2017

Alter O, et al. Singular value decomposition for genome-wide expression data processing and modeling. PNAS. 97(18):10101-10106, 2000

Andjelkovic V, et al. Changes in gene expression in maize kernel in response to water and salt stress. Plant Cell Rep. 25(1):71-99, 2006

Anjo SI, et al. SWATH-MS as a tool for biomarker discovery: From basic research to clinical applications. Proteomics. 17(3-4), 2017

Ballman KV, et al. Faster cyclic loess: normalizing RNA arrays via linear models. Bioinformatics. 20(16):2778-86, 2004

Blaise BJ. Data-driven sample size determination for metabolic phenotyping studies. Anal Chem. 85(19):8943-8950, 2013

Bolstad BM, et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 19(2):185-93, 2003

Borgaonkar SP, et al. Comparison of normalization methods for the identification of biomarkers using MALDI-TOF and SELDI-TOF mass spectra. OMICS. 14(1):115-26, 2010

Bouyssié D, et al. Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: application to the proteomics study of membrane proteins from primary human endothelial cells. Mol Cell Proteomics. 6(9):1621-1637, 2007

Broudy D, et al. A framework for installable external tools in Skyline. Bioinformatics. 30(17):2521-2523, 2014

Bruderer R, et al. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. Mol Cell Proteomics. 14(5):1400-1410, 2015

Bruderer R, et al. High-precision iRT prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation. Proteomics. 16(15-16):2246-2256, 2016

Callister SJ, et al. Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. J Proteome Res. 5(2):277-86, 2006

Cao MQ, et al. Identification of salivary biomarkers in breast cancer patients with thick white or thick yellow tongue fur using isobaric tags for relative and absolute quantitative proteomics. Zhong Xi Yi Jie He Xue Bao. 9(3):275-280, 2011

Casado-Vela J, et al. iTRAQ-based quantitative analysis of protein mixtures with large fold change and dynamic range. Proteomics. 10(2):343-347, 2010

Chai LE, et al. Investigating the effects of imputation methods for modelling gene networks using a dynamic bayesian network from gene expression data. Malays J Med Sci. 21(2):20-22, 2014

Chawade A, et al. Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. J Proteome Res. 13(6):3114-3120, 2014

Cheadle C, et al. Analysis of microarray data using Z score transformation. J Mol Diagn. 5(2):73-81, 2003

Chen YY, et al. Refining comparative proteomics by spectral counting to account for shared peptides and multiple search engines. Anal Bioanal Chem. 404(4):1115-1125, 2012

Cho CK, et al. Proteomics analysis of human amniotic fluid. Mol Cell Proteomics. 6(8):1406-15, 2007

Cociorva D, et al. Validation of tandem mass spectrometry database search results using DTASelect. Curr Protoc Bioinformatics. Chapter 13:Unit 13.4, 2007

Codrea MC, et al. Platforms and Pipelines for Proteomics Data Analysis and Management. Adv Exp Med Biol. 919:203-215, 2016

Colaert N, et al. Thermo-msf-parser: an open source Java library to parse and visualize Thermo Proteome Discoverer msf files. J Proteome Res. 10(8):3840-3843, 2011

Cox J, et al. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 26(12):1367-1372, 2008

De Livera AM, et al. Normalizing and integrating metabolomics data. Anal Chem. 84(24):10768-10776, 2012

Dieterle F, et al. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. Anal Chem. 78(13):4281-4290, 2006

Dike AO. The Distribution of Cube Root Transformation of the Error Component of the Multiplicative Time Series Model. Global Journal Inc. 16(5):49-60, 2016

Dorts J, et al. Effects of cadmium exposure on the gill proteome of Cottus gobio: modulatory effects of prior thermal acclimation. Aquat Toxicol. 154:87-96, 2014

Dupierris V, et al. A toolbox for validation of mass spectrometry peptides identification and generation of database: IRMa. Bioinformatics. 25(15):1980-1981, 2009

Ejigu BA, et al. Evaluation of normalization methods to pave the way towards large-scale LC-MS-based metabolomics profiling experiments. OMICS. 17(9):473-485, 2013

Escher C, et al. Using iRT, a normalized retention time for more targeted measurement of peptides. Proteomics. 12(8):1111-1121, 2012

Fermin D, et al. Abacus: A computational tool for extracting and pre-processing spectral count data for label-free quantitative proteomic analysis. Proteomics. 11(7):1340-1345, 2011

Gan X, et al. Microarray missing data imputation based on a set theoretic framework and biological knowledge. Nucleic Acids Res. 34(5):1608-1619, 2006

Gao Y, et al. Evaluation of sample extraction methods for proteomics analysis of green algae Chlorella vulgaris. Electrophoresis. 37(10):1270-1276, 2016

Garbis SD, et al. A novel multidimensional protein identification technology approach combining protein size exclusion prefractionation, peptide zwitterion-ion hydrophilic interaction chromatography, and nano-ultraperformance RP chromatography/nESI-MS2 for the in-depth analysis of the serum proteome and phosphoproteome: application to clinical sera derived from humans with benign prostate hyperplasia. Anal Chem. 83(3):708-18, 2011

Gaspari M, et al. Proteome Speciation by Mass Spectrometry: Characterization of Composite Protein Mixtures in Milk Replacers. Anal Chem. 88(23):11568-11574, 2016

Gautier V, et al. Label-free quantification and shotgun analysis of complex proteomes by one-dimensional SDS-PAGE/NanoLC-MS: evaluation for the large scale analysis of inflammatory human endothelial cells. Mol Cell Proteomics. 11(8):527-539. 2012

Griffin NM, et al. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nat Biotechnol. 28(1):83-89, 2010

Gromski PS, et al. The influence of scaling metabolomics data on model classification accuracy. Metabolomics. 11:684-695, 2015

Guo T, et al. Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. Nat Med. 21(4):407-413, 2015

Gärdén P, et al. PROTEIOS: an open source proteomics initiative. Bioinformatics. 21(9):2085-2087, 2005

Hoedt E, et al. SILAC-based proteomic profiling of the human MDA-MB-231 metastatic breast cancer cell line in response to the two antitumoral lactoferrin isoforms: the secreted lactoferrin and the intracellular delta-lactoferrin. PLoS One. 9(8):e104563, 2014

Hoekman B, et al. msCompare: a framework for quantitative analysis of label-free LC-MS data for comparative candidate biomarker studies. Mol Cell Proteomics. 11(6):M111, 2012

Hong MG, et al. Multidimensional Normalization to Minimize Plate Effects of Suspension Bead Array Data. J Proteome Res. 15(10):3473-3480, 2016

Huber W, et al. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics. 18 Suppl 1:S96-104, 2002

Häkkinen J, et al. The proteios software environment: an extensible multiuser platform for management and analysis of proteomics data. J Proteome Res. 8(6):3037-3043, 2009

Karpievitch YV, et al. Liquid Chromatography Mass Spectrometry-Based Proteomics: Biological and Technological Aspects. Ann Appl Stat. 4(4):1797-1823,2010

Karpievitch YV, et al. Metabolomics data normalization with EigenMS. PLoS One. 9(12):e116221, 2014

Karpievitch YV, et al. Normalization and missing value imputation for label-free LC-MS analysis. BMC Bioinformatics. 13(S16):S5, 2012

Keeping AJ, et al. Data variance and statistical significance in 2D-gel electrophoresis and DIGE experiments: comparison of the effects of normalization methods. J Proteome Res. 10(3):1353-60, 2011

Keller A, et al. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem. 74(20):5383-5392, 2002

Khoonsari PE, et al. Analysis of the Cerebrospinal Fluid Proteome in Alzheimer's Disease. PLoS One. 11(3):e0150672, 2016

Kim H, et al. Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics. 21(2):1-12, 2004

Kohl SM, et al. State-of-the art data normalization methods improve NMR-based metabolomic analysis. Metabolomics. 8(Suppl 1):146-160, 2012

MacLean B, et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments Bioinformatics. 26(7):966-968, 2010

Matzke MM, et al. Improved quality control processing of peptide-centric LC-MS proteomics data. Bioinformatics. 27(20):2866-2872, 2011

McManus FP, et al. Identification of cross talk between SUMOylation and ubiquitylation using a sequential peptide immunopurification approach. Nat Protoc. 12(11):2342-2358, 2017

Millikin RJ, et al. Ultrafast Peptide Label-Free Quantification with FlashLFQ. J Proteome Res, 2017

Mouton-Barbosa E, et al. In-depth exploration of cerebrospinal fluid by combining peptide ligand library treatment and label-free protein quantification. Mol Cell Proteomics. 9(5):1006-1021, 2010

Navarro P, et al. A multicenter study benchmarks software tools for label-free proteome quantification. Nat Biotechnol. 34(11):1130-1136, 2016

Nesvizhskii AI, et al. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem. 75(17):4646-58, 2003

Neuhauser N, et al. Expert system for computer-assisted annotation of MS/MS spectra. Mol. Cell. Proteomics. 11(11):1500-1509, 2012

Nezami Ranjbar MR, et al. Gaussian process regression model for normalization of LC-MS data using scan-level information. Proteome Sci. 11(Suppl 1):S13, 2013

Padoan A, et al. Reproducibility in urine peptidome profiling using MALDI-TOF. Proteomics. 15(9):1476-1485, 2015

Park GW, et al. Integrated Proteomic Pipeline Using Multiple Search Engines for a Proteogenomic Study with a Controlled Protein False Discovery Rate. J Proteome Res. 15(11):4082-4090, 2016

Park SK, et al. A quantitative analysis software tool for mass spectrometry-based proteomics. Nat Methods. 5(4):319-322, 2008

Park SK, et al. Census for proteome quantification. Curr Protoc Bioinformatics. Chapter 13:Unit 13.12.1-11, 2010

Prieto JH, et al. Large-scale differential proteome analysis in Plasmodium falciparum under drug treatment. PLoS One. 3(12):e4098, 2008

Pursiheimo A, et al. Optimization of Statistical Methods Impact on Quantitative Proteomics Data. J Proteome Res. 14(10):4118-4126, 2015

Ramus C, et al. Benchmarking quantitative label-free LC-MS data processing workflows using a complex spiked proteomic standard dataset. J Proteomics. 132:51-62, 2016

Ramus C, et al. Spiked proteomic standard dataset for testing label-free quantitative software and statistical methods. Data Brief. 6:286-294, 2015

Reindl J, et al. Proteomic biomarkers for psoriasis and psoriasis arthritis. J Proteomics. 140:55-61, 2016

Risso D, et al. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 32(9):896-902, 2014

Rosenberger G, et al. Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS. Nat Biotechnol 35(8):781-788, 2017

Rosenberger G, et al. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. Nat Methods 14(9):921-927, 2017

Röst HL, et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods. 13(9):741-748, 2016

Röst HL, et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat Biotechnol. 32(3):219-223, 2014

Röst HL, et al. TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. Nat Methods. 13(9):777-783. 2016

Sakia RM. The Box-Cox transformation technique: a review. The Statistician. 41:169-178, 1992

Saranya C, et al. A Study on Normalization Techniques for Privacy Preserving Data Mining. International Journal of Engineering and Technology. 5(3):2701-2704, 2013

Saraswat M, et al. Comparative proteomic profiling of the serum differentiates pancreatic cancer from chronic pancreatitis. Cancer Med. 6(7):1738-1751, 2017

Savas JN, et al. Proteomic Analysis of Protein Turnover by Metabolic Whole Rodent Pulse-Chase Isotopic Labeling and Shotgun Mass Spectrometry Analysis Methods Mol Biol. 1410:293-304, 2016

Schilling B, et al. Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation. Mol Cell Proteomics. 11(5):202-214, 2012

Schlaffner CN, et al. Fast, Quantitative and Variant Enabled Mapping of Peptides to Genomes. Cell Syst. 5(2):152-156, 2017

Searle BC. Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. Proteomics. 10(6):1265-9, 2016

Shao S, et al. Minimal sample requirement for highly multiplexed protein quantification in cell lines and tissues by PCT-SWATH mass spectrometry Proteomics. 15(21):3711-3721, 2015

Sturm M, et al. OpenMS - an open-source software framework for mass spectrometry. BMC Bioinformatics. 9:163, 2008

Tabb DL, et al. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J Proteome Res. 1(1):21-6, 2002

Tsou CC, et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat Methods. 12(3):258-264, 2015

Tsou CC, et al. Untargeted, spectral library-free analysis of data-independent acquisition proteomics data generated using Orbitrap mass spectrometers. Proteomics. 16(15-16):2257-2271, 2016

Twigt JM, et al. Preconception folic acid use influences the follicle fluid proteome. Eur J Clin Invest. 45(8):833-41, 2015

Tyanova S, et al. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc. 11(12):2301-2319, 2016

Tyanova S, et al. Visualization of LC-MS/MS proteomics data in MaxQuant. Proteomics. 15(8):1453-1456, 2015

Van den Berg RA, et al. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics. 7:142, 2006

Veit J, et al. LFQProfiler and RNP(xl): Open-Source Tools for Label-Free Quantification and Protein-RNA Cross-Linking Integrated into Proteome Discoverer. J Proteome Res. 15(9):3441-3448, 2016

Vidotto A, et al. Systems Biology Reveals NS4B-Cyclophilin A Interaction: A New Target to Inhibit YFV Replication. J Proteome Res. 16(4):1542-1555, 2017

Välikangas T, et al. A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. Brief Bioinform. doi:10.1093/bib/bbx054, 2017

Välikangas T, et al. A systematic evaluation of normalization methods in quantitative label-free proteomics. Brief Bioinform. pii: bbw095, 2016

Végvári A, et al. Bioinformatic strategies for unambiguous identification of prostate specific antigen in clinical samples Mol Cell J Proteomics. 75(1):202-210, 2011

Wan J, et al. Palmitoylated proteins: purification and identification. Nat Protoc. 2(7):1573-1584, 2007

Wang F, et al. Label free quantitative proteomics analysis on the cisplatin resistance in ovarian cancer cells. Cell Mol Biol (Noisy-le-grand). 63(5):25-28, 2017

Wang X, et al. Optimal consistency in microRNA expression analysis using reference-gene-based normalization. Mol Biosyst. 11(5):1235-1240, 2015

Webb-Robertson BJ, et al. A Statistical Analysis of the Effects of Urease Pre-treatment on the Measurement of the Urinary Metabolome by Gas Chromatography-Mass Spectrometry. Metabolomics. 10(5):897-908, 2014

Webb-Robertson BJ, et al. A Statistical Selection Strategy for Normalization Procedures in LC-MS Proteomics Experiments through Dataset Dependent Ranking of Normalization Scaling Factors. Proteomics. 11(24):4736-4741, 2011

Weisser H, et al. An automated pipeline for high-throughput label-free quantitative proteomics. J Proteome Res. 12(4):1628-1644, 2013

Wu JX, et al. SWATH Mass Spectrometry Performance Using Extended Peptide MS/MS Assay Libraries. Mol Cell Proteomics. 15(7):2501-2514, 2016

Wu L, et al. A hybrid retention time alignment algorithm for SWATH-MS data. Proteomics. 16(15-16):2272-2283, 2016

Yan W, et al. A dataset of human liver proteins identified by protein profiling via isotope-coded affinity tag (ICAT) and tandem mass spectrometry. Mol Cell Proteomics. 3(10):1039-1041, 2004

Yang YH, et al. Normalization for cDNA Microarray Data. Proc Spie. 6(10):1-21, 2003

Zhang J, et al. An intelligentized strategy for endogenous small molecules characterization and quality evaluation of earthworm from two geographic origins by ultra-high performance HILIC/QTOF MS(E) and Progenesis QI. Anal Bioanal Chem. 408(14):3881-3890, 2016

Zhang J, et al. PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. Mol Cell Proteomics. 11(4):M111, 2012

Zhang Y, et al. The Use of Variable Q1 Isolation Windows Improves Selectivity in LC-SWATH-MS Acquisition. J Proteome Res. 14(10):4359-4371, 2015

Zhang Z. Recombinant human activated protein C for the treatment of severe sepsis and septic shock: a study protocol for incorporating observational evidence using a Bayesian approach. BMJ Open. 4(7):e005622, 2014

College of Pharmaceutical Sciences, Zhejiang University
Hangzhou, P.R. China, 310058.
Contact number: +86-(0)571-8820-8444

Last Updated by: 6/8/2024