*Bioinformatics*, 37(2), 2021, 273–275
doi: 10.1093/bioinformatics/btaa1088
Advance Access Publication Date: 8 January 2021
Applications Note

Gene expression

# ProteomeExpert: a Docker image-based web server for exploring, modeling, visualizing and mining quantitative proteomic datasets

**Tiansheng Zhu [1,2,3,4], Hao Chen[1,2,5], Xishan Yan[6], Zhicheng Wu[1,2,3,4], Xiaoxu Zhou[1,2,3], Qi Xiao[1,2,3], Weigang Ge[1,2,5], Qiushi Zhang[1,2,5], Chao Xu[6], Luang Xu[1,2,3], Guan Ruan[1,2,5], Zhangzhi Xue[1,2,3], Chunhui Yuan[1,2,3,*], Guo-Bo Chen[7,8,*] and Tiannan Guo [1,2,3,*]**

[1]Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou, Zhejiang, China, [2]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang, China, [3]Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou, Zhejiang, China, [4]Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, China, [5]Westlake Omics (Hangzhou) Biotechnology Co., Ltd., Hangzhou, Zhejiang Province, China, [6]College of Mathematics and Informatics, Digital Fujian Institute of Big Data Security Technology, Fujian Normal University, China,[7]Clinical Research Institute, Zhejiang Provincial People's Hospital, People's Hospital of Hangzhou Medical College, Hangzhou, Zhejiang, China and [8]Key Laboratory of Endocrine Gland Diseases of Zhejiang Province, Hangzhou, Zhejiang, China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on September 14, 2020; revised on November 13, 2020; editorial decision on December 20, 2020; accepted on December 23, 2020

## Abstract

**Summary:** The rapid progresses of high-throughput sequencing technology-based omics and mass spectrometry-based proteomics, such as data-independent acquisition and its penetration to clinical studies have generated increasing number of proteomic datasets containing hundreds to thousands of samples. To analyze these quantitative proteomic datasets and other omics (e.g. transcriptomics and metabolomics) datasets more efficiently and conveniently, we present a web server-based software tool ProteomeExpert implemented in Docker, which offers various analysis tools for experimental design, data mining, interpretation and visualization of quantitative proteomic datasets. ProteomeExpert can be deployed on an operating system with Docker installed or with R language environment.

**Availability and implementation:** The Docker image of ProteomeExpert is freely available from https://hub.docker.com/r/lifeinfo/proteomeexpert. The source code of ProteomeExpert is also openly accessible at http://www.github.com/guomics-lab/ProteomeExpert/. In addition, a demo server is provided at https://proteomic.shinyapps.io/peserver/.

**Contact:** yuanchunhui@westlake.edu.cn; chen.guobo@foxmail.com; guotiannan@westlake.edu.cn

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Recent advances in liquid chromatographic mass spectrometry-based proteomics technology permit acquisition of hundreds to thousands of proteomics datasets in a relatively short time, especially using data-independent acquisition (DIA) mass spectrometry (MS) strategy (Gillet *et al.*, 2012; Guo *et al.*, 2015; Yue *et al.*, 2020; Zhang *et al.*, 2020). The substantial increase of proteomics data during the last few years necessitates effective algorithms and software tools for automatic interpretation of the resultant quantitative datasets to obtain valuable biological insights or assist clinical diagnosis. Currently, existing tools require programming skills, such as SWATH2stats (Blattmann *et al.*, 2016) and MSstats (Choi *et al.*, 2014). Another tool, mapDIA (Teo *et al.*, 2015), is a statistical analysis package for differentially expressed protein using DIA fragment-level intensities. PANDA-view (Chang *et al.*, 2018) and Perseus (Tyanova *et al.*, 2016) both depend on other packages and do not include functionalities in feature selection, peptide–protein inference and experimental design. Therefore, proteomics analysis tools supporting various functions across tasks still need replenishment.

Here, we provide an easy-to-use web server-based comprehensive data analysis platform for quantitative proteomics data and other omics (e.g. transcriptomics and metabolomics, hereafter) data that covers experimental design, data preprocessing, data quality control, protein inference, statistics, feature selection, unsupervised learning, supervised learning and visualization. In order to facilitate installing and deployment, we release this platform as a Docker image except for GitHub, which is easy to install and share on operating systems, such as Linux, Mac, and Windows.

## 2 Materials and methods

### 2.1 Architecture and implementation

ProteomeExpert is built as a Docker image. Figure 1 shows the overview modules of the platform. It includes an interactive web interface based on the Shiny package of R language, which integrates a collection of modules. The analysis can be customized after tuning the parameters in each interface; the resource-demanding computation is done in a remote server that can be further upgraded accordingly. This comprehensive architecture is designed to alleviate the computational burden in handling the increasing volume of proteomic data. Even with little programming skills, users can conduct most analyses that meet their requirements. The detailed description including power design, protein inference, feature selection and help information is in the Supplementary File.

### 2.2 Data input and output

'Data upload' is the core data input interface to upload the data file through a web interface. The data console allows uploading of the user-specific protein matrix and experiment meta-data (including experiment run sample and individual/patient information) as the input data for generating summary statistics, machine learning and data preprocessing. The modules for batch design, protein inference, annotation and supervised learning for the testing set have their own data upload functions. ProteomeExpert also provides interactive parameter settings and options to download processed data and plots.

### 2.3 Experimental design

The experimental design includes two sub-modules: power analysis and batch design. Power analysis enables the estimation of sample size for detecting the statistically significantly differentiated proteins given the balanced type I and type II error rates. Batch design is suitable for analyzing emerging large cohorts containing hundreds or thousands of samples, a scale that should be processed in multiple batches. However, unwanted variations are often creeped in due to technical variability among batches, so that a balanced batch design is implemented in this module that maximizes the statistical power but controls the technical variation. Batch design and batch correction methods are necessary for downstream data analysis, especially for large cohorts.

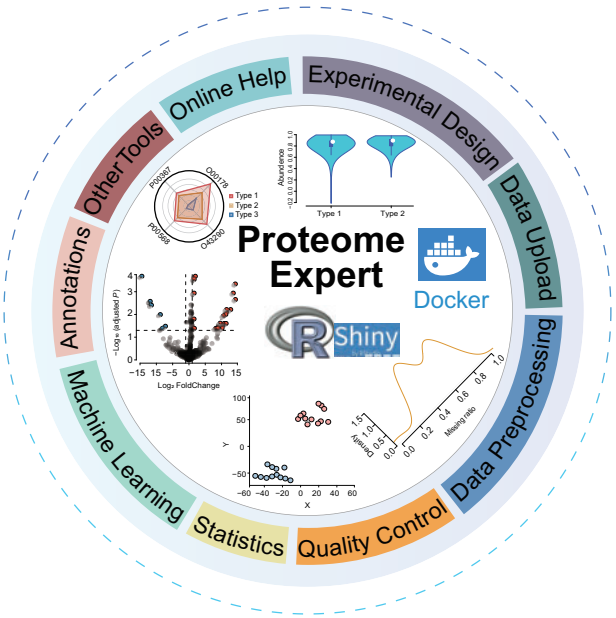


**Fig. 1.** The ProteomeExpert data analysis platform

### 2.4 Data preprocessing

The data preprocessing module includes methods for log transform, missing value substitution, normalization, batch effect correction, and replicates treatment. In particular, batch effect correction is important in data preprocessing. Batch effects may obscure biological signals, we recommend performing batch diagnostics, normalization, and adjustment right after peptide identification.

### 2.5 Machine learning

In the feature selection module, users not only use filter methods including near zero variance and high correlation but also exercise additional feature selection methods: LASSO (Tibshirani, 1996), genetic algorithm and random forest. As in clinical applications classifying disease into subtypes is of great interest in the fields of diagnosis and prognosis, users can perform various machine-learning analyses: unsupervised analysis PCA, t-SNE, and UMAP for dimensionality reduction, and supervised algorithm decision tree, random forest, and XGBoost for clustering.

## 3 Conclusions

In summary, we have developed ProteomeExpert to meet the requirement for processing large-scale quantitative proteomics datasets. Most, if not all, quantitative proteomic datasets can be fed into ProteomeExpert, including but not limited to DIA-MS with or without ion mobility, label-free or stable isotope labeling-based data-dependent acquisition MS, parallel reaction monitoring MS and multiple reaction monitoring MS. Transcriptomic and metabolomic datasets can also be processed by this tool. ProteomeExpert is compatible with other omics tools by uploading their results in tab-delimited or comma-separated text file format or excel file. Moreover, ProteomeExpert includes comprehensive methods for data preprocessing, visualization, statistics, and machine learning. It can be hosted within R shiny environment under Windows, Linux and Mac system or deployed in Docker available as a web server.

## Funding

This work was supported by grants from the National Key R&D Program of China [No. 2020YFE0202200]; Zhejiang Provincial Natural Science Foundation for Distinguished Young Scholars [LR19C050001]; Hangzhou Agriculture and Society Advancement Program [20190101A04]; National Natural Science Foundation of China [81972492]; and National Science Fund for Young Scholars [21904107].

*Conflict of Interest*: Tiannan Guo is shareholder of Westlake Omics Inc. Hao Chen, Weigang Ge and Qiushi Zhang are employees of Westlake Omics Inc. The remaining authors declare no competing interests.

## References

Blattmann,P. *et al.* (2016) SWATH2stats: an r/bioconductor package to process and convert quantitative SWATH-MS proteomics data for downstream analysis tools. *PLoS One*, **11**, e0153160.

Chang,C. *et al.* (2018) PANDA-view: an easy-to-use tool for statistical analysis and visualization of quantitative proteomics data. *Bioinformatics*, **34**, 3594–3596.

Choi,M. *et al.* (2014) MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics*, **30**, 2524–2526.

Gillet,L.C. *et al.* (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics*, **11**, O111016717.

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

Teo,G. *et al.* (2015) mapDIA: preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry. *J. Proteomics*, **129**, 108–120.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Methodol.*, **58**, 267–288.

Tyanova,S. *et al.* (2016) The perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods*, **13**, 731–740.

Yue,L. *et al.* (2020) Generating proteomic big data for precision medicine. *Proteomics*, **20**, e1900358.

Zhang,F. *et al.* (2020) Data-independent acquisition mass spectrometry-based proteomics and software tools: a glimpse in 2020. *Proteomics*, **20**, e1900276.