proteiNorm – A user-friendly tool for normalization and analysis of TMT and label-free protein quantification

Stefan Graw 1,2, Jillian Tang 1, Stephanie D Byrum1,2,\*

1 Department of Biochemistry and Molecular Biology, University of Arkansas for Medical Sciences, Little Rock, AR

2 Arkansas Children's Research Institute, Little Rock, AR

\* Corresponding author

ABSTRACT (Word Style “BD\_Abstract”).

The technological advances in mass spectrometry allow us to collect more comprehensive data with higher quality and increasing speed. With the rapidly increasing amount of data generated, the need for streamlining analyses becomes more apparent. Proteomic data is known to be often affected by systemic bias from unknown sources and failing to adequately normalize the data can lead to erroneous conclusions. To allow researchers to easily evaluate and compare different normalization via a user-friendly interface, we have developed “proteiNorm”.

The current implementation of proteiNorm accommodates some preliminary filter on peptide and sample level, followed by an evaluation of several popular normalization methods and identification of missing value pattern. The user then selects an adequate normalization method and one of several imputation methods used for the subsequent comparison of different differential abundance/expression methods and estimation of statistical power. The application of proteiNorm and interpretation of results is demonstrated on a Tandem Mass Tag mass spectrometry example, where the proteome of three different (breast cancer?) cell lines was profiled.

With proteiNorm, we provide a user-friendly tool to identify an adequate normalization method and to select an appropriate method for a differential abundance/expression analysis.

KEYWORDS: Proteomics, Normalization, Differential analysis

1. **Introduction**

With the latest technological advances in the development of mass spectrometer, the detection accuracy of the instruments and the speed with which data is being produced have drastically increased. The amount of data generated by these high-throughput devices increases rapidly and the need for tools that allow to streamline analyses becomes more apparent.

In a single run the peptide levels of thousands of proteins are being measured for up to several patients. However, data generated from mass spectrometry is often affected by systemic bias, variation caused by non-biological sources, which can lead to erroneous conclusions [1]. These sources of variation can include sample preparation and handling, device calibration, changes in temperature and other unknown sources [2]. Unlike the effect of batching, these sources of variation are either not measured or simply cannot be measured, and therefore cannot be controlled for in the statistical modelling. Normalization is a technique that aims to account for these systematic biases and make samples more comparable, while preserving the signal. Many normalization methods have been proposed with most of them being adopted from DNA microarray technology [3]. Välikangas et al. systemically evaluated normalization methods used in quantitative label-free proteomics and while there are normalization methods like VSN, linear regression normalization and local regression normalization, that consistently rank amount the top performing methods, it is curial to select a suitable normalization method depending on the data set [2, 4]. The need for tools that identify proper methods for normalization was recognized and addressed by Webb-Robertson et al. and Chawade et al. with their tools SPAN and Normalyzer, respectivly. SPAN combines eight peptides selection methods to select peptides subsequently used during the normalization and five methods for normalization [4]. Normalyzer includes popular normalization methods such as linear regression, local regression, total intensity, average intensity, median intensity, VSN, and quantile normalization. The performance of each normalization methods is individually evaluated by comparing their pooled coefficient of variance (PCV), pooled median absolute deviation (PMAD), and pooled estimate of variance (PEV) [1]. However, there remains an outstanding need for a user-friendly and publicly available tool that provides a systematically evaluation of normalization methods but also facilities imputation of missing data and compares different differential abundance methods depending on a chosen normalization method.

Here, we present proteiNorm, a user-friendly tool for a systematic evaluation of normalization methods, imputation of missing values and comparisons of different differential abundance methods. Our tool provides a point-and-click R-Shiny interface for researchers that are less comfortable or unfamiliar with the R environment. After loading the raw peptide data, the user can choose to remove peptide outliers, create and load a filtered protein file or to load a raw protein file. The meta data (i.e., custom sample names, group labels and batch labels) can be specified in the user-interface after loading peptide and protein data. Next, unwanted samples such as blanks, samples with poor quality or pooled samples can be remove from subsequent steps. The remaining data is then subjected to different normalization methods, which are compared among each other. Once a suitable normalization method is identified, missing values can be imputed, the data can be saved, and evaluation of different differential abundance methods can be performed for a selected normalization method. In addition, the statistical power of a given test of differential abundance can be assessed.

1. **Methods**
   1. **Implementation**

proteiNorm is implemented in R statistical language (<http://r-project.org>, version 3.6) using the package “shiny” [5] to provide a user-friendly user interface. proteiNorm is publicly available and can be downloaded from GitHub (<https://github.com/ByrumLab/proteiNorm>).

* 1. **Data Requirements**

As input, proteiNorm expects tab-separated raw peptide and/or protein data (not on logarithmic scale) as produced by software such as XXX *(where is the data coming from)*, where the column names of the measured intensities start with “Reporter intensity corrected” followed by an integer and an optional label (e.g. “Reporter intensity corrected 5 TMT2”,) and each row represents a peptide or protein. Data from both mass spectrometry quantitation methods (Tandem Mass Tag (TMT) and label-free mass spectrometry) are supported.

Due to the detection limit of mass spectrometry instruments many peptide/protein measurements have an intensity measurement of zero. These values will be considered as missing values (denoted as NA) and can be eventually imputed with precaution.

Filter function

* 1. **Normalization Methods**

The current implementation of proteiNorm includes several popular normalization methods, including log2 normalization (Log2), median normalization (Median), mean normalization (Mean), variance stabilizing normalization (VSN) [6], quantile normalization (Quantile) [7], cyclic loess normalization (Cyclic Loess) [8], global robust linear regression normalization (RLR) [1], and global intensity normalization (Global Intensity) [1]. The individual performance of each methods can be evaluated by comparing of the following metrices: total intensity, Pooled intragroup Coefficient of Variation (PCV), Pooled intragroup Median Absolute Deviation (PMAD), Pooled intragroup estimate of variance (PEV), intragroup correlation, sample correlation heatmap (Pearson), and log2-ratio distributions.

* 1. **Differential Analysis**

For the comparison of different differential abundance and expression methods proteiNorm employs the R package DAtest [9]. As the DAtest package requires complete data (no missing values), the user can user can select an appropriate method for normalization and imputation for the differential analysis. This choice should be guided by the comparison of the normalization methods and a heatmap visualizing patterns of missing values (missing at random (MAR) or missing not at random (MNAR)). The current implementation of proteiNorm provides the following imputation methods: k-nearest neighbors (KNN) [10], Quantile Regression Imputation of Left-Censored data (QRILC) [11], deterministic minimal value imputation (MinDet) [11], stochastic minimal value imputation (MinProb) [11], minimal value imputation (Min), and zero imputation (Zero).

In addition, proteiNorm provides the opportunity to adjust the DAtest-specific parameters (i.e., number of times to run the tests, effect size for the spike-ins (on natural intensity scale), and number of cores to use for parallel computing) and to exclude certain statistical tests. This latter option can be useful, as the exclusion of test with high computation demand (such as the permutation-based test) will drastically reduce the run-time.

DAtest automatically selects appropriate statistical tests from the remaining pool of tests and evaluate their performance. Therefore, the data is shuffled and spiked-in for randomly chosen features. DAtest then assesses if the spiked-in features are correctly identified by each test and whether the false discovery rate is controlled. The results are summarized in a table and figures comparing and ranking the performance of each individual test. Additionally, proteiNorm provides the opportunity to investigate the performance of each individual test by estimating statistical power (probability of correctly detecting a true signals) over a range of effect sizes (around the previously specified effect size for the spike-ins), utilizing the “powerDA” function from the “DAtest” package.

1. **Results**
2. **Discussion**
3. **Conclusion**
4. **Author Information**

Corresponding Author

\*(Word Style “FA\_Corresponding\_Author\_Footnote”). \* (Word Style “FA\_Corresponding\_Author\_Footnote”). Give contact information for the author(s) to whom correspondence should be addressed.

Present Addresses

†If an author’s address is different than the one given in the affiliation line, this information may be included here.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. ‡These authors contributed equally. (match statement to author names with a symbol)

Funding Sources

Any funds used to support the research of the manuscript should be placed here (per journal style).

Notes

Any additional relevant notes should be placed here.

1. **Acknowledgement**

(Word Style “TD\_Acknowledgments”). Generally the last paragraph of the paper is the place to acknowledge people, organizations, and financing (you may state grant numbers and sponsors here). Follow the journal’s guidelines on what to include in the Acknowledgments section.

1. **Abbreviations**
2. **References**

1. Chawade, A., E. Alexandersson, and F. Levander, *Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets.* J Proteome Res, 2014. **13**(6): p. 3114-20.

2. Valikangas, T., T. Suomi, and L.L. Elo, *A systematic evaluation of normalization methods in quantitative label-free proteomics.* Brief Bioinform, 2018. **19**(1): p. 1-11.

3. Karpievitch, Y.V., A.R. Dabney, and R.D. Smith, *Normalization and missing value imputation for label-free LC-MS analysis.* BMC Bioinformatics, 2012. **13 Suppl 16**: p. S5.

4. Webb-Robertson, B.J., et al., *A statistical selection strategy for normalization procedures in LC-MS proteomics experiments through dataset-dependent ranking of normalization scaling factors.* Proteomics, 2011. **11**(24): p. 4736-41.

5. RStudio, I., *Easy web applications in R.* 2013

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

7. Bolstad, B., *preprocessCore: A collection of pre-processing functions.* 2019.

8. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.

9. Russel, J., et al., *DAtest: a framework for choosing differential abundance or expression method.* bioRxiv, 2018: p. 241802.

10. Trevor Hastie, R.T., Balasubramanian Narasimhan and Gilbert Chu *impute: impute: Imputation for microarray data.* 2019.

11. Lazar, C., *imputeLCMD: A collection of methods for left-censored missing data imputation.* 2015.