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

Stefan Graw 1,2, Jillian Tang 1, Maroof Zafar 1, Alicia Byrd 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

KEYWORDS: Proteomics, Normalization, Differential analysis

**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 methods via a user-friendly interface, we have developed “proteiNorm”.

The current implementation of proteiNorm accommodates preliminary filter on peptide and sample level, followed by an evaluation of several popular normalization methods and visualization of missing value. 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 its results is demonstrated on a Tandem Mass Tag mass spectrometry example data set, where the proteome of three different breast cancer cell lines was profiled with and without Hydroxyurea treatment.

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.

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 and more apparent.

In a single mass spectrometry run peptide levels of thousands of proteins are can be measured for several patients. However, mass spectrometry data is often affected by systemic bias, variations 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 during the statistical analysis. 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]. Valikangas et al. systemically evaluated normalization methods used in quantitative label-free proteomics. While there are certain normalization methods that consistently rank among the top performing methods (i.e., variance stabilizing normalization (VSN), linear regression normalization and local regression normalization), it is curial to select a suitable normalization method for a specific 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, respectively. 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 and also facilities imputation of missing data and compares different differential abundance methods.

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 user-friendly point-and-click R-Shiny interface for researchers that are less comfortable or unfamiliar with the R programming environment.

After loading the raw peptide data, the user can choose to filter the peptide data using the “Top3” method [5] and create and load a filtered protein file or to load the corresponding 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 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 (optional) and the normalized protein data can be saved. In addition, different differential abundance methods can be evaluated for a selected normalization method and the statistical power of a given test of differential abundance can be assessed.

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

Our tool “proteiNorm” is implemented in R statistical language (<http://r-project.org>, version 3.6) using the package “shiny” [6] to provide a user-friendly user interface. It is publicly available and can be downloaded from GitHub (<https://github.com/ByrumLab/proteiNorm>). Figure 1 shows its workflow and illustrates an overview involving the following steps: Uploading peptide-level data, filtering peptides to create new filtered protein-level data (optional), upload protein-level data (original or filtered), provide meta data (e.g. batch and treatment groups), normalize data with different normalization methods and evaluate their performance, select appropriate normalization methods, and select appropriate imputation method (optional), export normalized (and imputed) protein-level data.

If the user uncertain about the choice of differential analysis method, proteiNorm provides the option so compare different differential abundance and expression methods provided by the R package “DAtest” [7].

Additional package dependencies include: **PACKAGES**

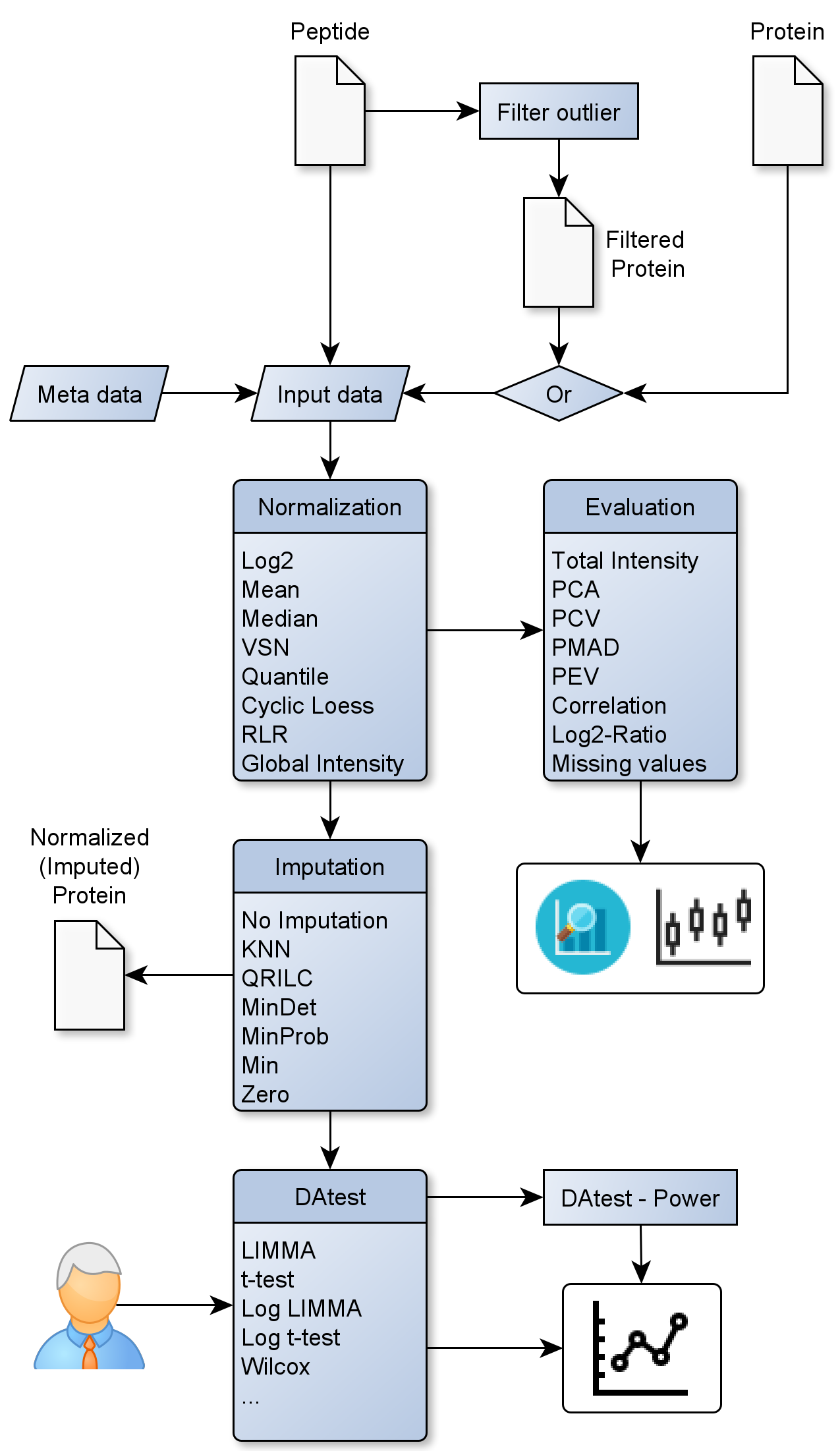


Figure Overview of proteiNorm’s workflow, inputs and outputs

* 1. **Data Requirements**

As input, proteiNorm expects tab-separated peptide (optional) and protein data (not on logarithmic scale) as produced by software such as MaxQuant [8], 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 measurements of peptides or proteins result in intensity levels of zero. These values will be considered as missing values (denoted as NA) and can be imputed with precaution. A modified heatmap of missing values (Figure 2I) from the DEP Bioconductor package [9] can be helpful to determine if data is missing at random (MAR) or missing not at random (MNAR). The MSnbase vignette describes different imputation methods used for different types of missing data [10].

* 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) [11], quantile normalization (Quantile) [12], cyclic loess normalization (Cyclic Loess) [13], 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 metrics: Total intensity (Figure 2A), principal component analysis (PCA; Figure 2B), intragroup correlation (Figure 2C), correlation heatmap (Pearson; Figure 2D), pooled intragroup coefficient of variation (PCV; Figure 2E), pooled intragroup estimate of variance (PEV; Figure 2F), pooled intragroup median absolute deviation (PMAD; Figure 2G), and log2-ratio distributions (Figure 2H).

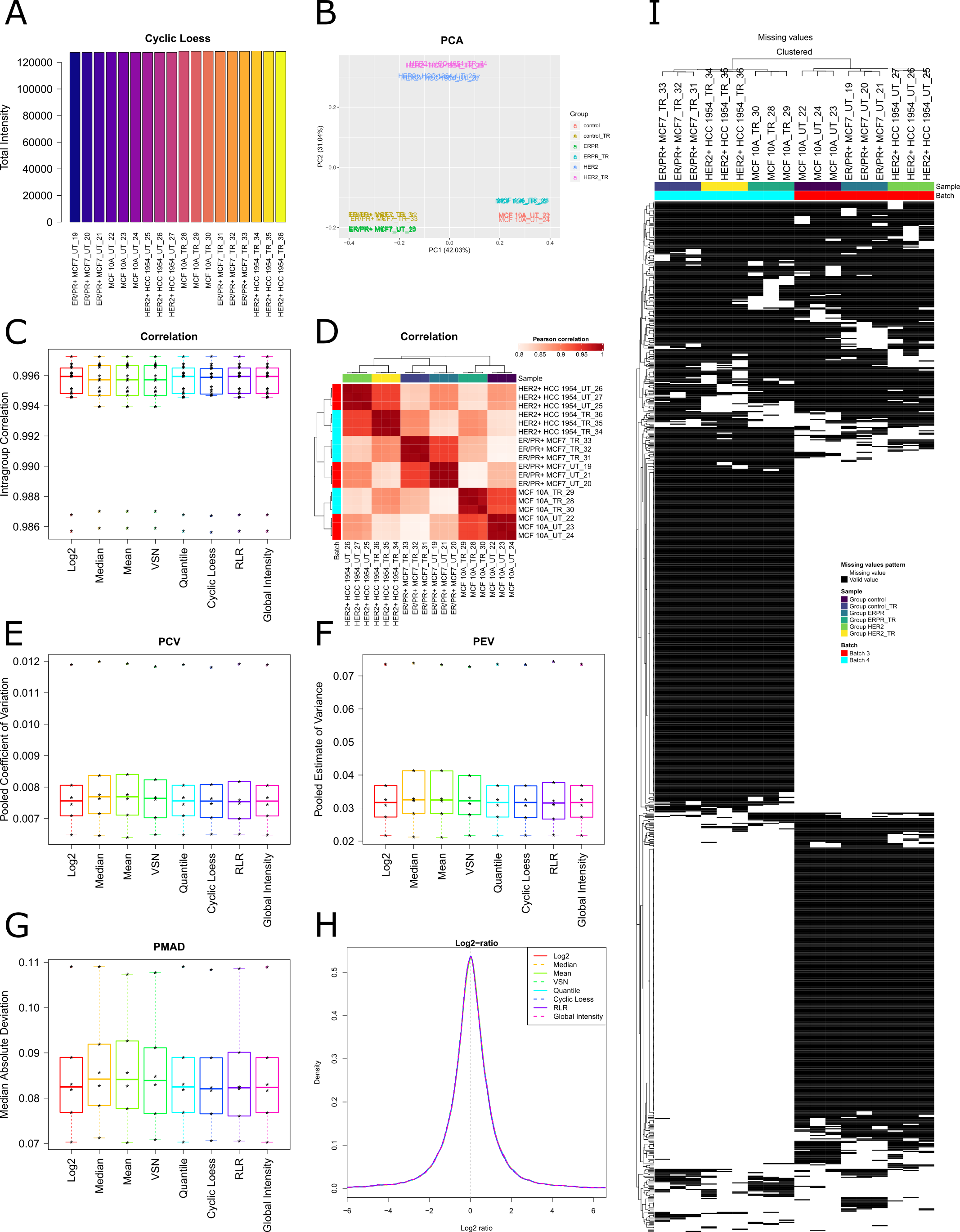


Figure Evaluation of normalization and missing values. (A) Sum of normalized intensities using cyclic loess normalization by sample. (B) Principle component analysis plot based on data normalized by cyclic loess normalization. (C) Pair-wise sample correlations within a group for different normalization methods. (C) Correlation heatmap (all pair-wise samples). (E) Pooled intragroup coefficient of variance comparing different normalization methods. (F) Pooled intragroup estimate of variance comparing different normalization methods. (G) Pooled intragroup median absolute deviation comparing different normalization methods. (H) Distribution of log2-ratios for different normalization methods. (I) Sample-clustered heatmap of missing data.

* 1. **Differential Analysis**

For the comparison of different differential abundance and expression methods proteiNorm employs the R package DAtest [7]. As the DAtest package requires complete data (no missing values), the 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 the heatmap visualizing patterns of missing values (Figure 2). The current implementation of proteiNorm provides the following imputation methods: k-nearest neighbors (KNN) [14], Quantile Regression Imputation of Left-Censored (QRILC) [15], deterministic minimal value imputation (MinDet) [15], stochastic minimal value imputation (MinProb) [15], 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. The latter option can be beneficial, 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 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. **Data**

An example data set consists of three different breast cancer cell lines (MCF10A, MCF7, HCC1954) with and without Hydroxyurea treatment with three replicates for each cell-line-treatment combination. The 18 samples were arrange on two Tandem Mass Tag (TMT) mass spectrometry batches such that untreated and treated cell lines assembled one batch each, and were profiled by **MS MACHINE NAME**.

1. **Results**

To demonstrate the application of proteiNorm and the interpretation of results, we analyzed proteomic data from the 18 samples (additional information in section 2.5). First, the raw peptide file produced by MaxQuant is uploaded and filtered using the “top3” method [5] and then exported as filtered proteinGroups file. Next, the newly created and filtered proteinGroups file is uploaded (alternatively, the unfiltered raw proteinGroups file generated by MaxQuant can be uploaded) and the corresponding meta-data is specified including treatment group, batch number and an optional custom name (as shown in Table 1). Next, pooled samples were unselected and therefore excluded from the subsequent evaluation. No additional samples were excluded, because no sample was identified as an outlier sample or a sample with poor quality based on the intensity distributions and the PCA plot. It can be observed that a batch effect is more prevalent in the raw peptide data compared to the filtered proteinGroups data. Afterwards, the performance of each normalization method is evaluated. Therefore the following figures are inspected (Figure 2): (**A**) Total intensity, (**B**) PCA, (**C**) Intragroup Correlation, (**D**) Correlation heatmap, (**E**) PCV, (**F**) PEV, (**G**) PMAD, and (**H**) LogRatio density. Subpanel A, B and D visualize the total intensity, PCA and correlation heatmap of the cyclic loess normalization, respectively (subpanels A, B and D of the remaining normalization methods not shown, as they produced similar results).

The total intensity of each sample after applying cyclic loess normalization is nearly indistinguishable (Figure 2A) and the remaining normalization methods produced very similar results (not shown). Near identical sums of intensities are to be expected, unless a treatment is expected to have a major impact on a subset of the most abundant proteins. Subpanel B of Figure 2 shows the PCA figure of the cyclic loess normalization (remaining normalization methods not shown, because of similar results) and demonstrates a stronger separation of cell lines after normalization, while the cell-line-treatment combination cluster tighter than before normalizing. Figure 2C compares the pairwise correlation within cell-line-treatment combinations of the normalization methods. High intragroup correlation values indicate high correlation of replicates within a cell-line-treatment combination and are desirable. All pairwise correlation are presented in a clustered heatmap (Figure 2D). The highest correlations are observed among samples from the same cell-line with either no treatment or treatment. The next strongest correlations are demonstrated between treated and untreated samples from the same cell-line. The second weakest correlations are produced by samples from different cell-lines with the same treatment, which are also samples within the same batch. Samples originating from different cell-lines with different treatment (in different batches) exhibited the weakest correlation. Figure 2’s subpanel E, F and G compare different measurements of intragroup variations (PCV, PEV and PMAD, respectively) for different normalization methods. Small values are desirable here as they indicate limited variation of sample replicates. It can be observed that all evaluated normalization methods produced similar results, with cyclic loess normalization ranking among the bester performing methods. In subpanel H of Figure 2 the density distribution of log2-ratios are compared for each normalization methods. For our example data set no differences are apparent among the different normalization methods. Unless an unbalanced regulation (primarily up- or down-regulation) of the data is to be expected, the densities should be centered around zero. Deviations of the density center from zero can potentially indicate biases in the normalized data.

Lastly, in Figure 2 subpanel I, a clustered heatmap of missing values for each sample is shown. Only proteins with at least one measured value and one missing value are shown (proteins with complete measurement or only missing values are not shown). The purpose of this heatmap is to identify patterns of missing values (i.e., MAR or MNAR) and guide the choice of imputation method, if desired. Here, it can be observed that samples cluster primarily by batch and secondarily by cell-line-treatment combination. It is apparent that measurements are not missing at random, but rather the missingness of values is influenced by batch and cell-line.

Based on the evaluation and comparison of normalization methods and the heatmap of missing values, the user can choose an adequate normalization method and has the choice to select an imputation method. Because some statistical methods do not require complete data (e.g. limma [16]), imputation is not always necessary. Following the selection of normalization and imputation method (“no imputation” is an imputation choice), the normalized (and imputed) protein data can be exported.

If the user is undecided about the statistical method for the differential abundance/expression analysis, different differential abundance/expression methods can be compared using the incorporated functions of DAtest (note: DAtest requires complete data and therefore an imputation methods need to be selected unless the data has no missing measurements). DAtest evaluate all appropriate statistical methods for the provided type of data and then repeatedly evaluate the performance of each method by creating an artificial spike-in signal for a number for random proteins and assessing the ability of a giving method to correctly identify proteins with artificial spike-in signal. The different differential abundance/expression methods are then rank-ordered and their power can be evaluated for a range of effect sizes.

1. **Discussion**

In our example data set we evaluated different popular normalization methods and while most methods performed very similar, we decided to use the cyclic loess normalization for this data set, based on its marginally better performance. In addition, cyclic loess normalization is one of the normalization methods that consistently ranks among the top performing methods [2]. While it is important to identify an adequate normalization method for the data set at hand, Valikangas et al. [2] have demonstrated that most normalization methods result in better performance than the simple log2 transformation.

The currently implementation of proteiNorm is limited to the evaluation of normalization methods utilizing intra-group variance and correlation. A more suitable form of evaluations for normalization methods is the assessment of true spike-in signal and/or proteomics standards. However, this requires additional steps during the sample preparation and might not always be possible. In future versions of proteiNorm we intend to include the option of evaluating normalization methods by assessing spike-in signal and/or proteomic standards. Further extensions include a broader spectrum of input file, and additional normalization and imputation methods.

1. **Conclusion**

When analyzing proteomic data measured by mass spectrometer, normalization of the data is an integral data processing step to adjust for unwanted systematic biases. Failing to acknowledge such issues can lead to erroneous results and inference, which limits reproducibility, scientific progress and increases the cost of science.

With proteiNorm we present a user-friendly tool with the goal to assist researchers in the evaluation of popular normalization methods and to identify an adequate normalization method suited for a given data set. Researchers also have the option to impute missing values and/or compare different differential abundance/expression methods and their power for a range of effect size.

1. **Author Information**

Corresponding Author

Stephanie Byrum, [sbyrum@uams.edu](mailto:sbyrum@uams.edu), (501) 686-5783, University of Arkansas for Medical Sciences, 4301 W Markham St. slot 516, Little Rock, AR 72205

Author Contributions

SG tool development manuscript writing and editing, JT developed tool and manuscript editing

MZ and AB sample preparation, SB tool development manuscript writing and editing.

Funding Sources

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

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**

VSN variance stabilizing normalization, PCV pooled coefficient of variance, PMAD pooled median absolute deviation, PEV pooled estimate of variance, TMT tandem mass tag, MAR missing at random, MNAR missing not at random, RLR robust linear regression, PCA principal component analysis, KNN k-nearest neighbors, QRILC Quantile Regression Imputation of Left-Censored

1. **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. Silva, J.C., et al., *Absolute quantification of proteins by LCMSE - A virtue of parallel MS acquisition.* Molecular & Cellular Proteomics, 2006. **5**(1): p. 144-156.

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

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

8. Tyanova, S., T. Temu, and J. Cox, *The MaxQuant computational platform for mass spectrometry-based shotgun proteomics.* Nature Protocols, 2016. **11**(12): p. 2301-2319.

9. Zhang, X.F., et al., *Proteome-wide identification of ubiquitin interactions using UbIA-MS.* Nature Protocols, 2018. **13**(3): p. 530-550.

10. Gatto, L. and K.S. Lilley, *MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation.* Bioinformatics, 2012. **28**(2): p. 288-289.

11. 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.

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

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

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

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

16. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Research, 2015. **43**(7).

Table Meta data for the example data set. Protein sample names are automatically generated from the proteinGroups file. Custom sample names is optional and replaces protein sample names when provided (needs to be unique name). Group specifies individual treatment groups. Batch is optional and indicates the batch for each sample.

|  |  |  |  |
| --- | --- | --- | --- |
| **Protein.Sample.Names** | **Custom.Sample.Names** | **Group** | **Batch** |
| Reporter.intensity.corrected.1.TMT3 | ER/PR+ MCF7\_UT\_19 | ERPR | 3 |
| Reporter.intensity.corrected.2.TMT3 | ER/PR+ MCF7\_UT\_20 | ERPR | 3 |
| Reporter.intensity.corrected.3.TMT3 | ER/PR+ MCF7\_UT\_21 | ERPR | 3 |
| Reporter.intensity.corrected.4.TMT3 | MCF 10A\_UT\_22 | control | 3 |
| Reporter.intensity.corrected.5.TMT3 | MCF 10A\_UT\_23 | control | 3 |
| Reporter.intensity.corrected.6.TMT3 | MCF 10A\_UT\_24 | control | 3 |
| Reporter.intensity.corrected.7.TMT3 | HER2+ HCC 1954\_UT\_25 | HER2 | 3 |
| Reporter.intensity.corrected.8.TMT3 | HER2+ HCC 1954\_UT\_26 | HER2 | 3 |
| Reporter.intensity.corrected.9.TMT3 | HER2+ HCC 1954\_UT\_27 | HER2 | 3 |
| Reporter.intensity.corrected.10.TMT3 | pool\_3 | pool | 3 |
| Reporter.intensity.corrected.1.TMT4 | MCF 10A\_TR\_28 | ERPR\_TR | 4 |
| Reporter.intensity.corrected.2.TMT4 | MCF 10A\_TR\_29 | ERPR\_TR | 4 |
| Reporter.intensity.corrected.3.TMT4 | MCF 10A\_TR\_30 | ERPR\_TR | 4 |
| Reporter.intensity.corrected.4.TMT4 | ER/PR+ MCF7\_TR\_31 | control\_TR | 4 |
| Reporter.intensity.corrected.5.TMT4 | ER/PR+ MCF7\_TR\_32 | control\_TR | 4 |
| Reporter.intensity.corrected.6.TMT4 | ER/PR+ MCF7\_TR\_33 | control\_TR | 4 |
| Reporter.intensity.corrected.7.TMT4 | HER2+ HCC 1954\_TR\_34 | HER2\_TR | 4 |
| Reporter.intensity.corrected.8.TMT4 | HER2+ HCC 1954\_TR\_35 | HER2\_TR | 4 |
| Reporter.intensity.corrected.9.TMT4 | HER2+ HCC 1954\_TR\_36 | HER2\_TR | 4 |
| Reporter.intensity.corrected.10.TMT4 | pool\_4 | pool | 4 |