This is an open access article published under a Creative Commons Attribution (CC-BY) License, which permits unrestricted use, distribution and reproduction in any medium, provided the author and source are cited.

ACS AUTHORCHOICE

Journal of proteome research

pubs.acs.org/jpr

Article

Isobaric Matching between Runs and Novel PSM-Level Normalization in MaxQuant Strongly Improve Reporter Ion-Based Quantification

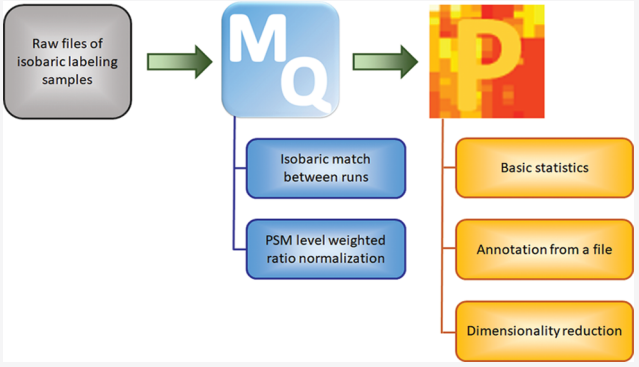
Sung-Huan Yu, Pelagia Kyriakidou, and Jürgen Cox*

Cite This: *J. Proteome Res.* 2020, 19, 3945−3954

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Isobaric labeling has the promise of combining high sample multiplexing with precise quantification. However, normalization issues and the missing value problem of complete *n*-plexes hamper quantification across more than one *n*-plex. Here, we introduce two novel algorithms implemented in MaxQuant that substantially improve the data analysis with multiple *n*-plexes. First, isobaric matching between runs makes use of the three-dimensional MS1 features to transfer identifications from identified to unidentified MS/MS spectra between liquid chromatography−mass spectrometry runs in order to utilize reporter ion intensities in unidentified spectra for quantification. On typical datasets, we observe a significant gain in MS/MS spectra that can be used for quantification. Second, we introduce a novel PSM-level normalization, applicable to data with and without the common reference channel. It is a weighted median-based method, in which the weights reflect the number of ions that were used for fragmentation. On a typical dataset, we observe complete removal of batch effects and dominance of the biological sample grouping after normalization. Furthermore, we provide many novel processing and normalization options in Perseus, the companion software for the downstream analysis of quantitative proteomics results. All novel tools and algorithms are available with the regular MaxQuant and Perseus releases, which are downloadable at http://maxquant.org.



KEYWORDS: *isobaric labeling, tandem mass tag, multiplexed quantification, normalization, match between runs, batch effects, missing values, MaxQuant, Perseus*

INTRODUCTION

Mass spectrometry (MS) has revolutionized the way researchers can monitor protein abundance changes on a proteome-wide scale. Several techniques have been established for quantifying relative amounts of proteins or peptides between related samples as, for instance, stable isotope labeling on the level of first-stage MS (MS1) spectra,[1−4] label-free quantification[5,6] (LFQ), and isobaric labeling.[7−9] The latter has the advantage of allowing for relatively high sample multiplexing and is often done in the form of tandem mass tags[7] (TMTs) or isobaric tags for relative and absolute quantitation.[7,10] Isobaric labeling can substantially improve on a genuine issue for shotgun proteomics, which is the missing value problem. In unlabeled samples, values for quantification can be missing because of several reasons, as for instance low abundance of the protein or lack of identification of peptides.[11,12] It has been observed that in label-free samples, the fraction of proteins containing one or more missing values can dominate the proteome data.[13] Isobaric labeling improves the situation to some extent because within an *n*-plex, the chances of getting missing values are strongly reduced. However, the absence of complete *n*-plexes over experimental designs distributing samples over several *n*-plexes occurs with the same likelihood as single values are missing in label-free data because comparisons across multiple *n*-plex sets are subject to the same stochastic sampling and possibilities of missing values as label-free proteomics.

MaxQuant is one of the most widely used platforms for analyzing shotgun proteomics data.[14−17] In order to recover features for quantification, beyond those that were directly identified by fragmentation spectra, several methods have been developed and integrated into the MaxQuant software in the past. Match between runs[18,19] (MBR) is one of the methods to decrease the number of missing values in label-

Received: March 30, 2020
Published: September 7, 2020

ACS Publications

© 2020 American Chemical Society

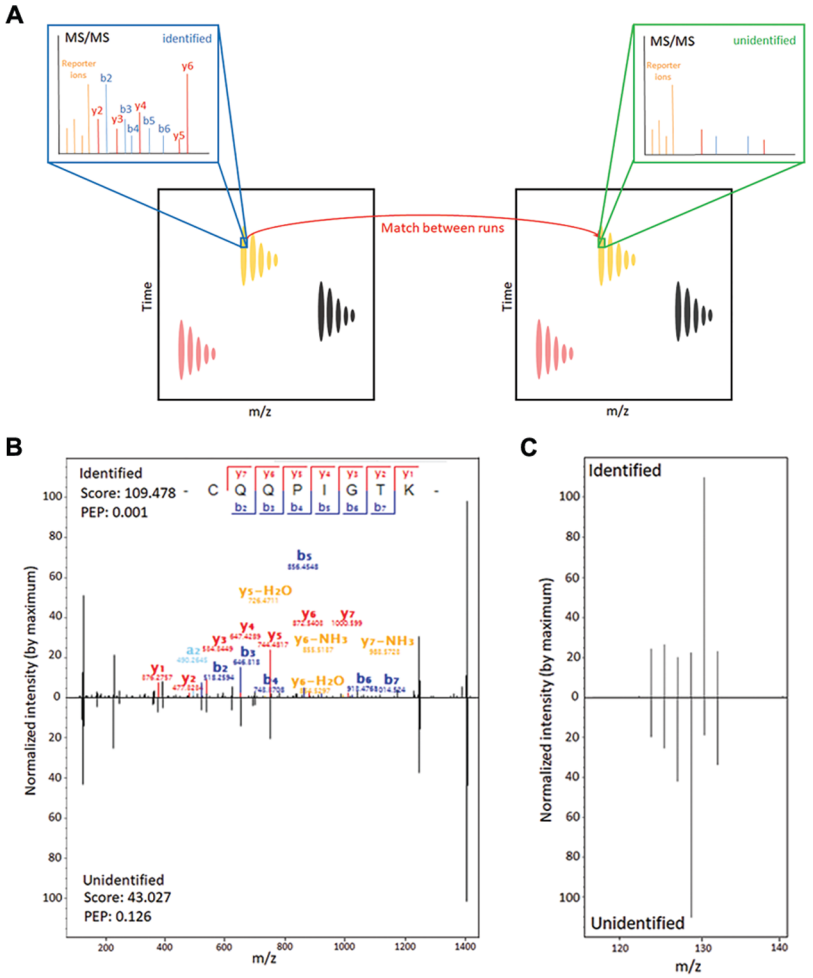


Figure 1. Overview of IMBR. (A) Schematic diagram explaining IMBR. The reporter ions in the unidentified MS/MS spectrum can be used for quantification based on the matching of MS1-level 3D isotope patterns. For instance, the yellow MS1 isotope pattern appears in both LC−MS runs. It has been identified by an MS/MS spectrum in the left run. In the right run, the same MS1 feature appears, however with an MS/MS spectrum that did not lead to the identification of the peptide because of poor coverage of the y and b ion series. This MS/MS spectrum does however contain reporter ion intensities, which are then used for quantification. (B) Example of a pair of identified and matched spectra. The upper MS/MS spectrum (raw file: 29May3013_DJB_mouse_tmt8_BR1_unfrac_165min_dda15_1.raw, scan number 8353) was identified with an Andromeda score of 109.5. The lower MS/MS spectrum (raw file: 29May3013_DJB_mouse_tmt8_BR3_unfrac_165min_dda15_1.raw, scan number 8676) has not been identified because of the absence of several peaks in the y- and b-ion series and was fond by IMBR. (C) Zoom of the spectra displayed in (B) into the mass region where the reporter ion signals are located.

free and MS1-level labeled data. It can transfer peptide identifications from a liquid chromatography (LC)−MS run, in which the peptide is identified by MS/MS, to another LC−MS run, in which the same peptide exists as an MS1 feature, but was not identified, either because no fragmentation spectrum was recorded for this MS1 feature or because the recorded fragmentation spectrum was not identified by the peptide search engine.[20] Requirements for this transfer of identifications between similar samples to happen with low rates of false positives are high mass accuracy obtained after nonlinear mass recalibration and comparable retention times after retention time alignment in MaxQuant. "Re-quantify" provides another method to increase the coverage of protein quantification for the case of MS1-level labeling. An isotope pattern that has not been paired with any labeling partners can, after it has been identified, be restored for quantification of ratios by integrating the MS signal at the expected position in m/z and retention time coordinates in the same LC−MS run.[20] Although these options can improve the protein quantification for label-free or MS1-labeling experiments, a method of recovering values for isobaric labeling still needs to be developed in MaxQuant.

Besides procedures for recovering missing values, normalization is another important step in isobaric labeling-based quantification. For removing batch effects, several useful normalization methods have recently been developed.[21−27] Almost all of the methods apply corrections at the level of protein quantification. An exception is the "compositional proteomics" strategy[28] which removes effects because of constraints imposed on the sum over channels for each PSM. We propose a new and straightforward PSM-level normalization method based on the weighted median of ratios. For MS1-level labeling, it is advantageous to define the protein ratio as the median of the peptide feature ratios, as done in MaxQuant. For MS1 signals, which are not affected by cofragmentation, the gain in robustness resulting from applying a median approach to the ratios outweighs the lack of weighing ratios by their signal strength. This is not the case for isobaric labeling signals, which often show inferior results if the protein ratio is taken as the unweighted median of PSM

https://dx.doi.org/10.1021/acs.jproteome.0c00209
J. Proteome Res. 2020, 19, 3945−3954

ratios, as shown in this manuscript. Rather simplistic methods that are not robust in the sense of the median but weighted by the signal intensity are better suited here because they reduce the influence of cofragmentation. The novel normalization method that we propose combines the strength of both approaches, the robustness of the median and the weighting of PSMs by signal strength.

In this manuscript, we present a novel isobaric MBR (IMBR) and a PSM-level normalization method that were built into MaxQuant. They can significantly increase the number of peptide features that are available for quantification and efficiently remove the batch effects. Moreover, numerous plug-ins of Perseus,[29,30] which is a powerful platform for the downstream analysis of data generated with MaxQuant or other platforms, have been developed for processing isobaric labeling datasets. All the software is available for download at http://maxquant.org.

EXPERIMENTAL SECTION

Datasets

For the evaluation of our newly developed methods, we use well-established datasets that were submitted to open access community databases. We downloaded isobaric labeling datasets of two types: those with a reference channel, in which one of the channels carries the signal of the mixture of samples suitable for forming ratios to the other channels that carry the actual samples, and those without a reference channel. As an example for the latter, we obtained data from Bailey et al.[31] It is an 8-plex dataset consisting of eight organs (kidney, lung, heart, muscle, liver, cerebrum, cerebellum, and spleen) harvested from four different mice. After tryptic digestion, the peptides of each organ were labeled with a TMT 8-plex in a randomized design and applied to two different data acquisitions—data dependent acquisition (DDA) and intelligent data acquisition (IDA). The purpose of using a dataset employing more than one acquisition methods is to demonstrate applicability of our newly developed methods to heterogeneous data. The dataset was obtained from the CHORUS database (http://chorusproject.org/; 298: Elution Order Algorithm). Moreover, we downloaded a TMT 10-plex labeling dataset containing a reference channel acquired by Lereim et al.[32] It consists of brain tissues from WT mice and *Peli1* knock-out mice. *Peli1* functions as a regulator of the immune response during experimental autoimmune encephalomyelitis (EAE). Each type of mice contains three samples based on the number of days post EAE infection: 0, 10, and 20. The samples are randomly assigned to two TMT 10-plex sets, and TMT131 is for the pooled samples consisting of a mixture of same amounts of all samples. The dataset is available in PRIDE[33] (PXD003710).

Data Processing

For both datasets, we used mouse UniProt sequences (UP000000589, reviewed at 24-07-2018, 16,992 proteins). All searches were performed with oxidation of methionine and protein N-terminal acetylation as variable modifications and cysteine carbamidomethylation as fixed modification. Trypsin was selected as protease allowing for up to two missed cleavages, and the peptide mass was limited to a maximum of 4600 Da. The initial mass tolerance was 20 ppm for precursor ions and 20 ppm for fragment ions. PSM and protein false discovery rates were both applied at 1%. In general, values of parameters in MaxQuant have not been changed from their default values unless explicitly stated.

Isobaric MBR

Matching between runs of MS1 features is performed exactly as it was done before for the quantification of unlabeled samples (Figure 1A). Prior to matching of MS1 features, their masses are recalibrated and their retention times are aligned in MaxQuant. The yellow MS1 isotope pattern in Figure 1A appears in both LC−MS runs. It has been identified by an MS/MS spectrum in the left run. In the right run, the same MS1 feature appears, however with an MS/MS spectrum that did not lead to the identification of the peptide because of poor coverage of the y and b ion series. This MS/MS spectrum does however contain reporter ion intensities, which are then used for quantification. A concrete example of an MS/MS spectrum pair of which one was identified and the other one was found by IMBR is shown in Figure 1B. While the lower MS/MS spectrum was not identified, it still contains reporter ion signals that are comparable in intensity to the ones present in the identified MS/MS spectrum (Figure 1C).

PSM-Level Weighted Ratio Normalization

In order to remove batch effects, we developed a novel method for normalizing reporter ion intensities in isobaric labeling datasets at the PSM-level and integrated it into MaxQuant software. First, for each protein, the quantifiable PSMs are retrieved. These follow, in the first instance, the same rules as for label-free or MS1-level labeling quantification. For instance, if the protein quantification should be based only on protein group-level unique peptide sequences, then these are selected. On top of this, second-stage MS (MS2)-level labeling specific filters can be applied, as for instance, based on the precursor ion fraction or base peak ratio. The resulting set of filtered PSMs is then subjected to a weighted median calculation of reporter ion intensities. How exactly the values of the weights are determined is described later in this section. Their purpose is to give higher weights to more abundant signals. For now, we consider them as constant nonnegative weights that sum up to one. The weighted median of the ratios x_1, x_2, ..., x_n for a particular isobaric labeling channel to the reference channel over n valid PSMs matching to the protein group is then taken. For the weighted median calculation, we assume that the ratios x_i are sorted in the ascending order and that the nonnegative final weights w_1, w_2, ..., w_n are normalized such that they sum up to one. The weighted median is the ratio x_k with the index k satisfying

$$\sum_{i=1}^{k-1} w_i \leq 1/2 \text{ and } \sum_{i=k+1}^{n} w_i \leq 1/2$$

The weights w_i are the product of the precursor ion intensity exactly at the retention time at which the MS/MS has been recorded times the fill time of the MS/MS spectrum. This is supposed to be proportional to the number of ions that are used for fragmentation. The weights are then exponentiated with a constant which can be set by the user (the"isobaric weight exponent" parameter in the graphical user interface). How the optimal value of the isobaric weight exponent is determined is described in the Results section. The weights w_i are the exponentially converted values, and the normalization is applied such that these sum up to one. All calculations are done for raw intensities as well as for

https://dx.doi.org/10.1021/acs.jproteome.0c00209
J. Proteome Res. 2020, 19, 3945−3954

intensities corrected for impurities, which are both reported in the output tables. An example for the calculation procedures of the normalization is shown in Supporting Information Figure S1.

Analysis Workflow in MaxQuant and Perseus

For running isobaric labeling data in MaxQuant, first, the reference channel(s) need to be assigned after loading the raw data. Multiple reference channels per *n*-plex are supported, in which case the sum of signals over the reference channels is used for normalization. If the dataset does not contain reference channels, all channels have to be assigned as reference channels and their total signal sum is taken in that case. Second, "reporter ion MS2" has to be selected as "type" under "group-specific parameters", and a suitable set of isobaric labels has to be chosen. Furthermore, if IMBR should be applied, the "MBR" option in "identification" under "global parameters" needs to be turned on. In addition, "normalization" needs to be specified as "ratio to reference channel", which is normalizing the data without weight or "weighted ratio to reference channel". If "weighted ratio to reference channel" is selected, the weight can be defined in "isobaric weight exponent" on the "misc." page. The parameters of newly developed methods are shown in Supporting Information Figure S2.

For analysis of the output tables of MaxQuant in Perseus, protein groups which are known contaminants, only identified by site or reverse, are removed. Data is then usually logarithmized for further analysis, and, if desired, mean or median subtraction per sample can be performed. Note that the channel intensities obtained from MaxQuant are not mean or median centered automatically. We removed protein groups with more than 30% missing values in total across all channels and *n*-plexes. There are many alternative ways of applying missing value filters, for instance, filtering for a minimum valid value percentage in at least one of the groups of samples. The optimal criteria for filtering may depend on the dataset and on the question that is investigated. Because this is not the subject of this manuscript, we chose the simplest possible form of filtering here by the total percentage of valid values. Moreover, the annotations of cell types were generated automatically in Perseus. The imputation of remaining missing values was performed based on sampling from a normal distribution with a value of 0.3 for the width parameter and 1.8 for the down-shift parameter, which are the defaults in Perseus,[30] with the purpose of simulating low abundant expression values. Further explanation of the imputation method can be found in Figure S3 of the original paper[30] (https://media.nature.com/original/nature-assets/nmeth/journal/v13/n9/extref/nmeth.3901-S1.pdf). We compare the results of the imputation with noise levels from empty channels in Supporting Information Figure S3, indicating that the distribution of signals from empty channels shows a strong overlap with the distribution of missing values.

Uniform manifold approximation and projection (UMAP)[34] is a newly developed nonlinear dimensionality reduction method. UMAP is similar to t-distributed stochastic neighbor embedding (t-SNE) but has some advantages over it. It is faster and preserves the global structure better than t-SNE. Moreover, UMAP is created based on manifold approximation techniques, whereas t-SNE is mainly a visualization heuristic. Therefore, the distance between clusters is more meaningful. Moreover, a study using these two methods for bioinformatic analyses in the single cell dataset was published in 2018.[35] It shows that UMAP is better performing than t-SNE. In Perseus, UMAP, t-SNE, and principal component analysis are all provided for the downstream analysis of MaxQuant results.

Additionally, the columns of reference channels were excluded from the analysis. The workflow is presented in Supporting Information Figure S4. The details of the newly developed plugins for isobaric labeling data analysis and dimensionality reduction methods are shown in Supporting Information Figures S5 and S6.

RESULTS AND DISCUSSION

Reduction of Missing Values

We developed IMBR as an extension of the already existing MBR algorithm in MaxQuant in order to reduce the missing *n*-plexes problem in isobaric labeling data. While within *n*-plex sets, the likelihood of a value missing is low, complete *n*-plex sets are missing with a similar probability as values are missing in label-free datasets (Supporting Information Figure S7). For benchmarking IMBR we use two publicly available TMT datasets, one with and one without a reference channel. To study the influence of IMBR on missing values, we performed MaxQuant analyses once without and once with IMBR with all other parameters at default values as described in the Experimental Section on both datasets. An evidence entry corresponds to a 3D MS1 feature that has at least one MS/MS spectrum attached that is used for quantification. On both datasets, we see a consistent increase of evidence entries carrying an MS/MS spectrum that can be used for quantification by 7−9% through IMBR. This increase will presumably be higher in datasets in which many *n*-plexes are combined because also the label-free matching between runs has a tendency toward higher rates of matched features in datasets with more samples. In Figure 2, the distribution of peptides found in a certain number of isobaric labeling batches is displayed with and without IMBR for the dataset by Bayley et al. As can be seen, by applying IMBR, the amounts of quantified peptides and proteins that are consistently found in all samples are 2.5 and 1.5 fold more than without using IMBR. There is no strong bias in reporter intensities between identified and matched MS/MS spectra (see Supporting Information Figure S8). The median log2 reporter intensities for identified and matched features are 23.043 and 23.030 in Supporting Information Figure S8A, respectively, and 22.34 and 20.66 in Supporting Information Figure S8B, respectively.

Determination of the Optimal Isobaric Weight Exponent

For determining the optimal value for the parameter "isobaric weight exponent", we scan W different values of the weight parameter between 0 and 1 with increment 0.05. Consider a dataset with a biological or technical replicate grouping into G groups and quantitative data for P proteins (or, more specifically, protein groups). The samples within the replicate groups should be completely randomized over the isobaric labeling batches, as it has been done in the datasets we are analyzing. We want to monitor a measure for the variability within replicate groups in relation to the total variability. For this purpose, we define

$$F_{p,g,w} = \frac{\text{variance within group } g \text{ for protein } p}{\text{total variance for protein } p}$$

https://dx.doi.org/10.1021/acs.jproteome.0c00209
J. Proteome Res. 2020, 19, 3945−3954

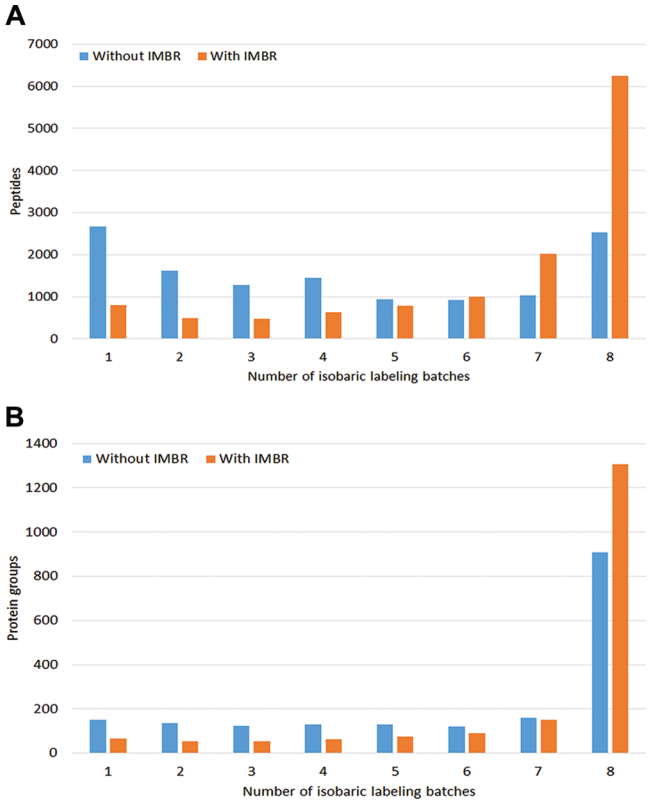


Figure 2. Improvement by IMBR. (A) For the Bailey et al. dataset, it is shown how many peptides are found in n out of eight isobaric labeling batches. Blue and orange bars represent results without and with IMBR, respectively. (B) Same as (A) for protein groups.

where $w = 1, ..., W$ is indexing the different values for the isobaric weight exponent. All variance calculations are performed on logarithmized and imputed values as described above. We take the median of these quantities over all proteins to obtain

$$G_{g,w} = \text{median}_{p=1,...,P} F_{p,g,w}$$

which is a measure for the relative spread of group g when using weight parameter w. In order to get a balanced contribution from all groups, we rescale $G_{g,w}$ within each group over the weight parameter values to a range from 0 to 100

$$M_{g,w} = \frac{100}{(\bar{G}_g - \underline{G}_g)} \cdot (G_{g,w} - \underline{G}_g)$$

where

$$\bar{G}_g = \max_{w=1,...,W} G_{g,w}$$

$$\underline{G}_g = \min_{w=1,...,W} G_{g,w}$$

The quantity

$$M_w = \sum_{g=1}^{G} M_{g,w}$$

is then optimized over the different values of the isobaric weight exponent. Figure 3 reveals that M_w is minimal if the weight exponent assumes the value 0.75 on the dataset by Bailey et al. Based on these findings, the default value for the isobaric weight exponent is set to 0.75 in the software. Because the optimal value might slightly change between datasets, the user may reoptimize the value for their data and change the value accordingly in MaxQuant. The source code and documentation of a script (MedianVar.R) for the variance calculations can be found at https://github.com/cox-labs/tools.

Effects of Weighted Median-Based Normalization

In order to test the effect of the novel PSM-level normalization, we ran MaxQuant with and without applying the normalization on the dataset of Bailey et al. We then performed UMAP analysis on both outputs. Without normalization, the result of UMAP analysis is dominated by the split into two clusters which separate the data by the acquisition method (Figure 4A). When applying weighted median-based normalization (Figure 4B), UMAP analysis results in strongly focused clusters by tissues across the acquisition methods and all isobaric labeling batches. Hence, the biases caused by

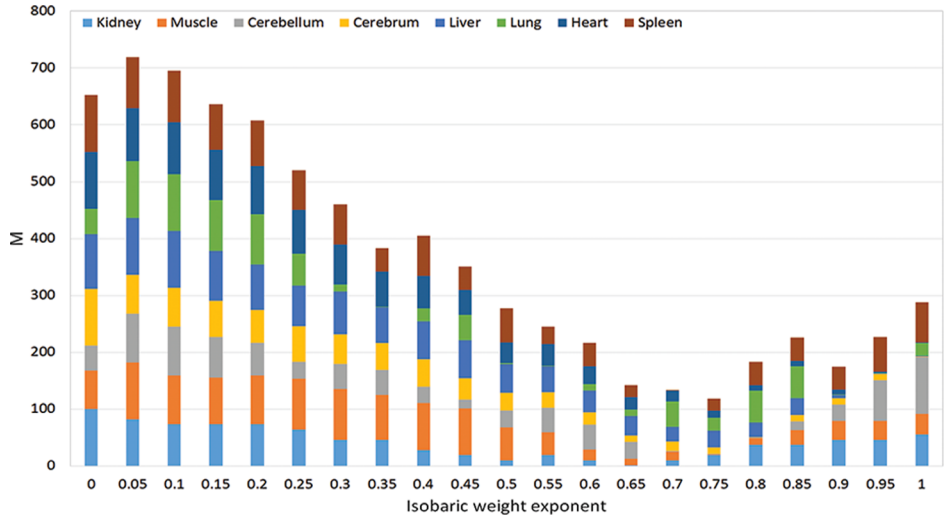


Figure 3. Optimization of the isobaric weight exponent parameter on the dataset by Bailey et al. M_w as defined in the Results section is plotted as a function of the weight exponent. The contributions of separate tissues, $M_{g,w}$, are color-coded.

https://dx.doi.org/10.1021/acs.jproteome.0c00209
J. Proteome Res. 2020, 19, 3945−3954

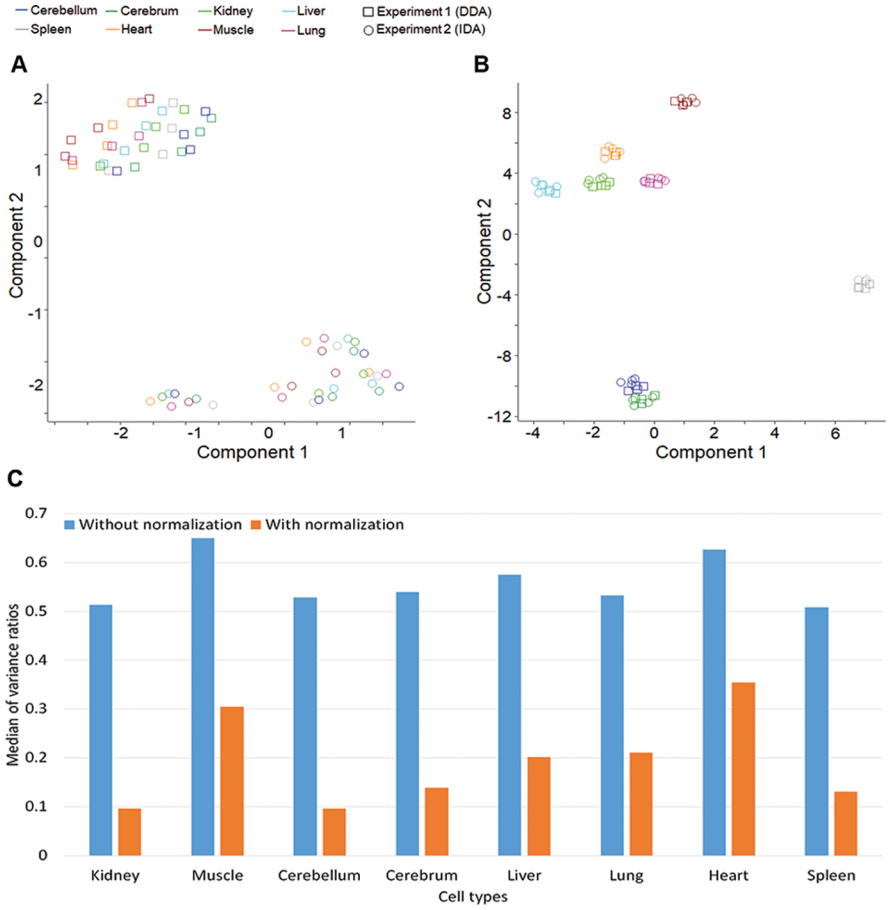


Figure 4. Effect of weighted median normalization. (A) Result of UMAP analysis on the dataset from Bailey et al. analyzed with MaxQuant without normalization. The different colors designate tissue types while the squares and circles denote the acquisition method. (B) Same as (A) but after applying normalization. (C) Each bar represents the median over all protein groups of the ratio of within tissue group variances to total variances. Blue and orange bars stand for before and after normalization.

different acquisition methods have been removed. The impact of the normalization can also be seen by comparing within tissue group variances divided by total variances before and after normalization. Figure 4C shows the media of these over the protein groups. There is a strong decrease in the within group variance brought about by the normalization.

When performing UMAP analysis to the acquisition modes (IDA and DDA) separately, the clustering is also not by tissues but by the TMT 8-plexes (Figure 5A,C). The grouping into TMT-multiplexes coincides with the individual mice, but we assume here that the separation is due to the labeling multiplexes. Independent of what causes the separation, using weighted median normalization removes this batch effect and results in clustering by tissues for both acquisition methods (Figure 5B,D).

Normalization with Reference Channels

In many isobaric labeling experiments, one or more channels in each *n*-plex are filled with a common reference sample, typically a mixture of the samples measured in the study or very similar samples. Here, we show the benefits of weighted median normalization for data with such reference channels. Similar to the previous section, we perform MaxQuant analysis once without and once with normalization. In both cases, we remove the reference channel and perform UMAP analysis. Without applying normalization, the samples are grouped based on the TMT 10-plex in the UMAP plot (Figure 6A). Using weighted median normalization produces two clusters mainly based on the type of mouse (Figure 6B). Most of the data points not following the clustering by the mouse type are the samples from 0 days post infection (dpi). It may be due to the fact that the infection is not active yet at 0 dpi. Hence, we also performed the UMAP analyses of the subset excluding the samples of 0 dpi (Figure 6C). The samples are separated according to WT and *Peli1* knock-out except for one data point. Performing UMAP analysis only for the samples from 20 dpi, the normalized data are completely classified by the type of mice (Figure 6D).

MaxQuant and Perseus Plug-Ins for Isobaric Labeling

All of the developed features for isobaric labeling data analysis were integrated into MaxQuant version 1.6.12.0. Reference channels, normalization methods, and the exponent for the weights of the PSM level ratio normalization can be assigned in the graphical user interface (Supporting Information Figure S2). In order to perform the downstream analysis for the isobaric labeling proteomics data, some newly created plug-ins were integrated in the current version of Perseus (1.6.12.0). All the plug-in activities for isobaric labeling are listed under the heading "isobaric labeling" in Perseus. These activities are "annotation from a file" which can assign names to and group the samples in isobaric labeling *n*-plexes based on the information contained in user-defined categorical rows (Supporting Information AnnotationDataset_1.txt and Anno-

https://dx.doi.org/10.1021/acs.jproteome.0c00209
J. Proteome Res. 2020, 19, 3945−3954

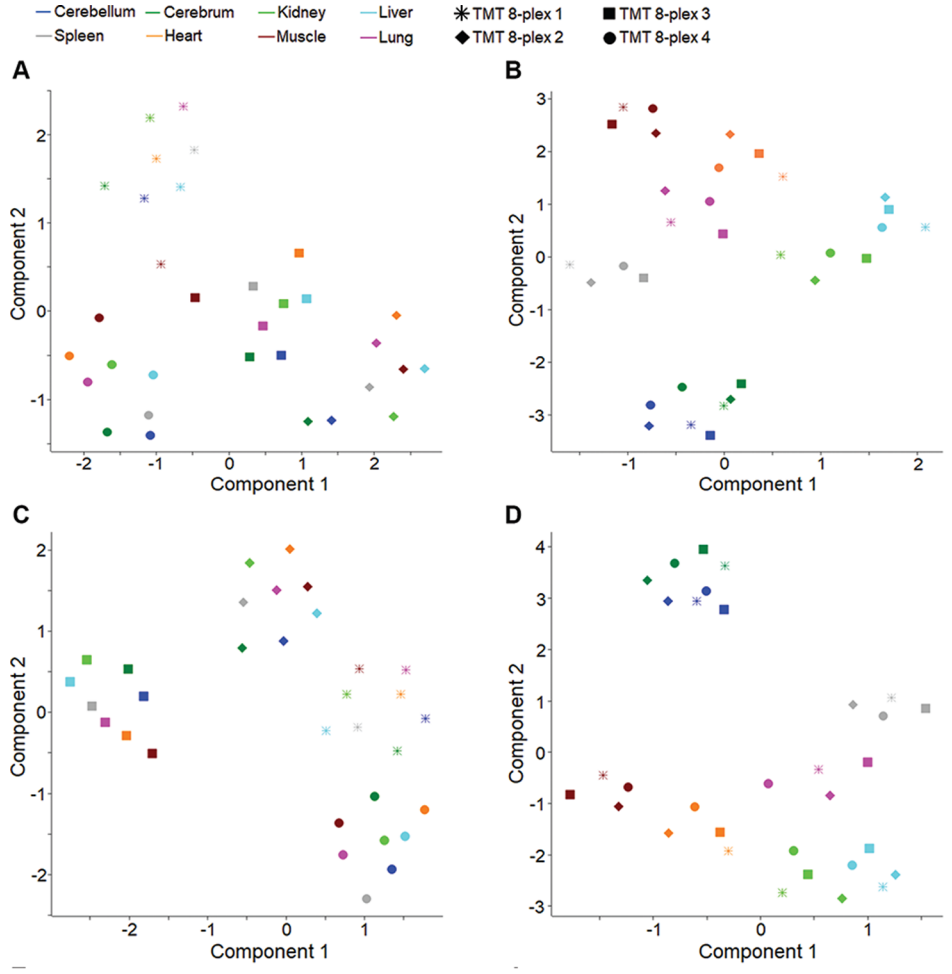


Figure 5. Separate analysis of acquisition methods. (A) Separate UMAP analysis of the IDA data without normalization colors represents different tissues, while the symbol types represent different TMT 8-plexes. The samples cluster by 8-plex. (B) Same as (A) but with normalization. The samples cluster by the tissue. (C) Same as (A) for DDA, (D) Same as (B) for DDA.

tationDataset_2.txt). It can also be used to highlight empty channels, in case they are kept as controls in the further analysis. "Remove channels" is for removing the specific channels, for instance those that have been used as a common reference or carrier channel. Moreover, UMAP and t-SNE are also integrated into Perseus (located in "clustering") for dimension reduction and classification by using PluginInterop and PerseusR.[34−37] The options of the plug-ins can be found in Supporting Information Figures S5 and S6.

CONCLUSIONS

Two novel approaches for isobaric labeling data analysis were presented, which were integrated into MaxQuant: IMBR and PSM-level weighted median ratio normalization. They achieve higher precision and fewer missing values in protein quantification. In addition, a collection of Perseus plugins useful for the downstream analysis of the MaxQuant output for isobaric labeling data was introduced. PSM-level normalization efficiently removes batch effects for the subsequent analysis. It is particularly a flexible method because it does not require to specify what the factor(s) of interest in the dataset are but works in an unsupervised manner in that respect. Our approaches can also be applied to data with prefractionation prior to LC−MS analysis. In that case, the IMBR will only connect features between samples within same or neighboring fractions. Furthermore, it is compatible with MS2 and MS3 methods. Based on results shown in this study, the combination of tools and algorithms in MaxQuant and Perseus is a useful gear for the analysis of isobaric labeling MS data.

An alternative method for recovering more MS/MS spectra for TMT quantification beyond the primarily identified PSMs was presented in the context of the single-cell proteomics technology SCoPE-MS.[38] In this complementary approach, the gain in PSMs was achieved by a Bayesian update of the posterior error probability of low-confidence identifications. The combination of this approach with ours could be of interest and will be the subject of further investigations.

Although isobaric labeling with applying IMBR can decrease the number of missing values significantly, not all of the missing values will be replaced with numbers. Hence, imputation is still an important issue that needs to be addressed. Many data analysis methods require a complete data matrix or show potential benefits from imputation. Numerous studies and tools for optimizing imputation have been published and released.[12,39,40] For the remaining missing values in isobaric labeling, we propose the same treatment as we recommend for LFQ data, which is imputed by drawing from a left-shifted random number distribution, as done by default in Perseus.

https://dx.doi.org/10.1021/acs.jproteome.0c00209
J. Proteome Res. 2020, 19, 3945−3954

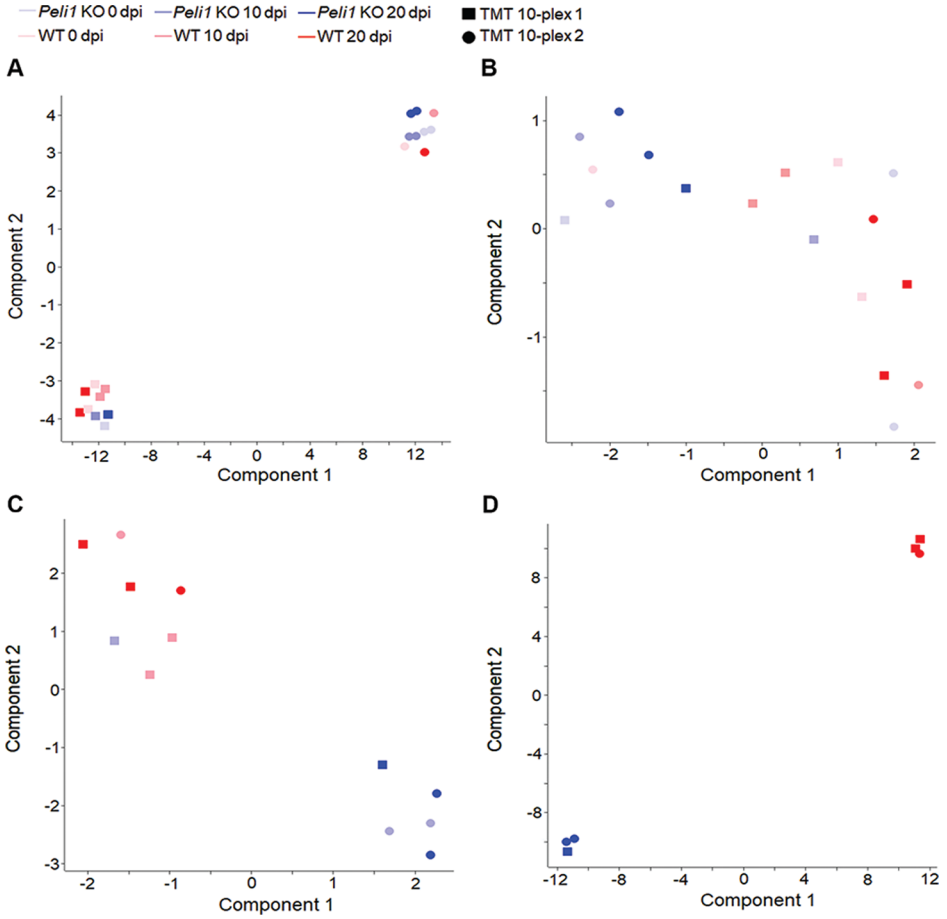


Figure 6. Normalization with the reference channel. (A) Result of UMAP analysis for the dataset by Lereim et al. analyzed with MaxQuant without normalization. The different colors designate conditions and time points while the symbol shapes indicate the TMT batches. (B) Same as (A) but with normalization. (C) Same as (B) but omitting time point 0. (D) Same as (B) but using only the last time point.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.0c00209.

An example for the procedures of weighted ratio normalization; settings in MaxQuant for isobaric labeling data; comparison between the distributions of samples, empty channels, and imputation; screenshot of a workflow in Perseus; plug-ins for isobaric labeling data in Perseus; dimensionality reduction methods in Perseus; statistics of missing values within and between TMT sets; and intensity histograms for identified and matched peptides (PDF)

AnnotationDataset_1.txt: Annotation file for the dataset from Bailey et al. (TXT)

AnnotationDataset_2.txt: Annotation file for the dataset from Lereim et al. (TXT)

■ AUTHOR INFORMATION

Corresponding Author

Jürgen Cox − *Computational Systems Biochemistry, Max-Planck Institute of Biochemistry, Martinsried 82152, Germany; Department of Biological and Medical Psychology, University of Bergen, Bergen 5009, Norway;* orcid.org/0000-0001-8597-205X; Email: cox@biochem.mpg.de

Authors

Sung-Huan Yu − *Computational Systems Biochemistry, Max-Planck Institute of Biochemistry, Martinsried 82152, Germany;* orcid.org/0000-0001-7955-8645

Pelagia Kyriakidou − *Computational Systems Biochemistry, Max-Planck Institute of Biochemistry, Martinsried 82152, Germany;* orcid.org/0000-0002-1139-715X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.0c00209

Author Contributions

S.-H.Y. and J.C. planned and performed the research, developed the software, and wrote the manuscript. S.-H.Y., P.K., and J.C. performed the data analysis.

Notes

The authors declare no competing financial interest.

The MS proteomics data have been deposited to the ProteomeXchange consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD019880 and PXD019881.

■ ACKNOWLEDGMENTS

We thank all members of the Computational Systems Biochemistry research group for helpful discussions. This work has been made possible in part by grant number 2019-202671 from the Chan Zuckerberg Foundation.

REFERENCES

(1) Oda, Y.; Huang, K.; Cross, F. R.; Cowburn, D.; Chait, B. T. Accurate Quantitation of Protein Expression and Site-Specific Phosphorylation. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 6591−6596.

(2) Hsu, J.-L.; Huang, S.-Y.; Chow, N.-H.; Chen, S.-H. Stable-Isotope Dimethyl Labeling for Quantitative Proteomics. *Anal. Chem.* **2003**, *75*, 6843−6852.

(3) Merrill, A. E.; Hebert, A. S.; MacGilvray, M. E.; Rose, C. M.; Bailey, D. J.; Bradley, J. C.; Wood, W. W.; El Masri, M.; Westphall, M. S.; Gasch, A. P.; et al. NeuCode Labels for Relative Protein Quantification. *Mol. Cell. Proteomics* **2014**, *13*, 2503−2512.

(4) Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol. Cell. Proteomics* **2002**, *1*, 376−386.

(5) Zhang, B.; VerBerkmoes, N. C.; Langston, M. A.; Uberbacher, E.; Hettich, R. L.; Samatova, N. F. Detecting Differential and Correlated Protein Expression in Label-Free Shotgun Proteomics. *J. Proteome Res.* **2006**, *5*, 2909−2918.

(6) Cox, J.; Hein, M. Y.; Luber, C. A.; Paron, I.; Nagaraj, N.; Mann, M. Accurate Proteome-Wide Label-Free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol. Cell. Proteomics* **2014**, *13*, 2513−2526.

(7) Thompson, A.; Schäfer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Hamon, C. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Anal. Chem.* **2003**, *75*, 1895−1904.

(8) Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; et al. Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-Reactive Isobaric Tagging Reagents. *Mol. Cell. Proteomics* **2004**, *3*, 1154−1169.

(9) Rauniyar, N.; Yates, J. R. Isobaric Labeling-Based Relative Quantification in Shotgun Proteomics. *J. Proteome Res.* **2014**, *13*, 5293−5309.

(10) Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; et al. Multiplexed Protein Quantitation in *Saccharomyces Cerevisiae* Using Amine-Reactive Isobaric Tagging Reagents. *Mol. Cell. Proteomics* **2004**, *3*, 1154−1169.

(11) Albrecht, D.; Kniemeyer, O.; Brakhage, A. A.; Guthke, R. Missing Values in Gel-Based Proteomics. *Proteomics* **2010**, *10*, 1202−1211.

(12) Lazar, C.; Gatto, L.; Ferro, M.; Bruley, C.; Burger, T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J. Proteome Res.* **2016**, *15*, 1116−1125.

(13) O'Connell, J. D.; Paulo, J. A.; O'Brien, J. J.; Gygi, S. P. Proteome-Wide Evaluation of Two Common Protein Quantification Methods. *J. Proteome Res.* **2018**, *17*, 1934−1942.

(14) Cox, J.; Mann, M. MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification. *Nat. Biotechnol.* **2008**, *26*, 1367−1372.

(15) Leprevost, F. V.; Lima, D. B.; Crestani, J.; Perez-Riverol, Y.; Zanchin, N.; Barbosa, V. C.; Carvalho, P. C. Pinpointing Differentially Expressed Domains in Complex Protein Mixtures with the Cloud Service of PatternLab for Proteomics. *J. Proteomics* **2013**, *89*, 179−182.

(16) Sinitcyn, P.; Tiwary, S.; Rudolph, J.; Gutenbrunner, P.; Wichmann, C.; Yılmaz, Ş.; Hamzeiy, H.; Salinas, F.; Cox, J. MaxQuant Goes Linux. *Nat. Methods* **2018**, *15*, 401.

(17) Tyanova, S.; Temu, T.; Cox, J. The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics. *Nat. Protoc.* **2016**, *11*, 2301−2319.

(18) Sinitcyn, P.; Rudolph, J. D.; Cox, J. Computational Methods for Understanding Mass Spectrometry−Based Shotgun Proteomics Data. *Annu. Rev. Biomed. Data Sci.* **2018**, *1*, 207−234.

(19) Prianichnikov, N.; Koch, H.; Koch, S.; Lubeck, M.; Heilig, R.; Brehmer, S.; Fischer, R.; Cox, J. MaxQuant Software for Ion Mobility Enhanced Shotgun Proteomics. *Mol. Cell. Proteomics* **2020**, *19*, 1058−1069.

(20) Cox, J.; Matic, I.; Hilger, M.; Nagaraj, N.; Selbach, M.; Olsen, J. V.; Mann, M. A Practical Guide to the MaxQuant Computational Platform for SILAC-Based Quantitative Proteomics. *Nat. Protoc.* **2009**, *4*, 698−705.

(21) Ritchie, M. E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C. W.; Shi, W.; Smyth, G. K. Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* **2015**, *43*, No. e47.

(22) Robinson, M. D.; Oshlack, A. A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data. *Genome Biol.* **2010**, *11*, R25.

(23) Plubell, D. L.; Wilmarth, P. A.; Zhao, Y.; Fenton, A. M.; Minnier, J.; Reddy, A. P.; Klimek, J.; Yang, X.; David, L. L.; Pamir, N. Extended Multiplexing of Tandem Mass Tags (TMT) Labeling Reveals Age and High Fat Diet Specific Proteome Changes in Mouse Epididymal Adipose Tissue. *Mol. Cell. Proteomics* **2017**, *16*, 873−890.

(24) Maes, E.; Hadiwikarta, W. W.; Mertens, I.; Baggerman, G.; Hooyberghs, J.; Valkenborg, D. CONSTANd: A Normalization Method for Isobaric Labeled Spectra by Constrained Optimization. *Mol. Cell. Proteomics* **2016**, *15*, 2779−2790.

(25) Kim, P. D.; Patel, B. B.; Yeung, A. T. Isobaric Labeling and Data Normalization without Requiring Protein Quantitation. *J. Biomol. Tech.* **2012**, *23*, 11−23.

(26) D'Angelo, G.; Chaerkady, R.; Yu, W.; Hizal, D. B.; Hess, S.; Zhao, W.; Lekstrom, K.; Guo, X.; White, W. I.; Roskos, L.; et al. Statistical Models for the Analysis of Isobaric Tags Multiplexed Quantitative Proteomics. *J. Proteome Res.* **2017**, *16*, 3124−3136.

(27) Khan, S. Y.; Ali, M.; Kabir, F.; Renuse, S.; Na, C. H.; Talbot, C. C.; Hackett, S. F.; Riazuddin, S. A. Proteome Profiling of Developing Murine Lens Through Mass Spectrometry. *Invest. Ophthalmol. Visual Sci.* **2018**, *59*, 100−107.

(28) O'Brien, J. J.; O'Connell, J. D.; Paulo, J. A.; Thakurta, S.; Rose, C. M.; Weekes, M. P.; Huttlin, E. L.; Gygi, S. P. Compositional Proteomics: Effects of Spatial Constraints on Protein Quantification Utilizing Isobaric Tags. *J. Proteome Res.* **2018**, *17*, 590−599.

(29) Tyanova, S.; Cox, J. Perseus A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research. *Methods Mol. Biol.* **2018**, *1711*, 133−148.

(30) Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M. Y.; Geiger, T.; Mann, M.; Cox, J. The Perseus Computational Platform for Comprehensive Analysis of (Prote)Omics Data. *Nat. Methods* **2016**, *13*, 731−740.

(31) Bailey, D. J.; McDevitt, M. T.; Westphall, M. S.; Pagliarini, D. J.; Coon, J. J. Intelligent Data Acquisition Blends Targeted and Discovery Methods. *J. Proteome Res.* **2014**, *13*, 2152−2161.

(32) Lereim, R. R.; Oveland, E. The Brain Proteome of the Ubiquitin Ligase Peli1 Knock-Out Mouse during Experimental Autoimmune Encephalomyelitis. *J. Proteomics Bioinf.* **2016**, *9*, 209−219.

(33) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; et al. The PRIDE Database and Related Tools and Resources in 2019: Improving Support for Quantification Data. *Nucleic Acids Res.* **2019**, *47*, D442−D450.

(34) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **2018**, arXiv:1802.03426.

(35) Becht, E.; McInnes, L.; Healy, J.; Dutertre, C.-A.; Kwok, I. W. H.; Ng, L. G.; Ginhoux, F.; Newell, E. W. Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP. *Nat. Biotechnol.* **2018**, *37*, 38−44.

(36) Rudolph, J. D.; Cox, J. A Network Module for the Perseus Software for Computational Proteomics Facilitates Proteome Interaction Graph Analysis. *J. Proteome Res.* **2019**, *18*, 2052−2064.

https://dx.doi.org/10.1021/acs.jproteome.0c00209
J. Proteome Res. 2020, 19, 3945−3954

(37) van der Maaten, L. Accelerating t-SNE Using Tree-Based Algorithms. *J. Mach. Learn. Res.* **2014**, *15*, 3221−3245.
(38) Budnik, B.; Levy, E.; Harmange, G.; Slavov, N. SCoPE-MS: Mass Spectrometry of Single Mammalian Cells Quantifies Proteome Heterogeneity during Cell Differentiation. *Genome Biol.* **2018**, *19*, 161.
(39) Lazar, C. *imputeLCMD: A Collection of Methods for Left-Censored Missing Data Imputation*, 2015.
(40) Wei, R.; Wang, J.; Su, M.; Jia, E.; Chen, S.; Chen, T.; Ni, Y. Missing Value Imputation Approach for Mass Spectrometry-Based Metabolomics Data. *Sci. Rep.* **2018**, *8*, 663.

https://dx.doi.org/10.1021/acs.jproteome.0c00209
J. Proteome Res. 2020, 19, 3945−3954