Nucleic Acids Research, 2020 **1**
doi: 10.1093/nar/gkaa498

NAguideR: performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses

Shisheng Wang[1], Wenxue Li[2], Liqiang Hu[1], Jingqiu Cheng[1], Hao Yang[1,*] and Yansheng Liu[2,3,*]

[1]West China-Washington Mitochondria and Metabolism Research Center; Key Lab of Transplant Engineering and Immunology, MOH, Regenerative Medicine Research Center, West China Hospital, Sichuan University, Chengdu 610041, China, [2]Yale Cancer Biology Institute, Yale University, West Haven, CT 06516, USA and [3]Department of Pharmacology, Yale University School of Medicine, New Haven, CT 06520, USA

Received February 18, 2020; Revised April 20, 2020; Editorial Decision May 31, 2020; Accepted June 08, 2020

ABSTRACT

Mass spectrometry (MS)-based quantitative proteomics experiments frequently generate data with missing values, which may profoundly affect downstream analyses. A wide variety of imputation methods have been established to deal with the missing-value issue. To date, however, there is a scarcity of efficient, systematic, and easy-to-handle tools that are tailored for proteomics community. Herein, we developed a user-friendly and powerful stand-alone software, *NAguideR*, to enable implementation and evaluation of different missing value methods offered by 23 widely used missing-value imputation algorithms. *NAguideR* further evaluates data imputation results through classic computational criteria and, unprecedentedly, proteomic empirical criteria, such as quantitative consistency between different charge-states of the same peptide, different peptides belonging to the same proteins, and individual proteins participating protein complexes and functional interactions. We applied *NAguideR* into three label-free proteomic datasets featuring peptide-level, protein-level, and phosphoproteomic variables respectively, all generated by data independent acquisition mass spectrometry (DIA-MS) with substantial biological replicates. The results indicate that *NAguideR* is able to discriminate the optimal imputation methods that are facilitating DIA-MS experiments over those sub-optimal and low-performance algorithms. *NAguideR* further provides downloadable tables and figures supporting flexible data analysis and interpretation. *NAguideR* is freely available at http://www.omicsolution.org/wukong/NAguideR/ and the source code: https://github.com/wangshisheng/NAguideR/.

INTRODUCTION

Mass spectrometry (MS)-based quantitative proteomics provides a versatile approach for profiling thousands of peptides, proteins and proteoforms between different experimental conditions and disease specimens (1–3). The successful applications of quantitative proteomics, however, has been entangled with the lack of high reproducibility and consistency, which is often manifested as data missing values being generated between different technical replicates, experimental batches, biological replicates, and research groups. These missing values [or, *not available (NA)* data points] also frequently and negatively affect the subsequent analysis of proteomic data (4,5), such as hypothesis testing, principal component analysis and hierarchical clustering analysis, which routinely require complete data matrix as input.

The missing values in proteomic datasets (6) were previously discussed and ascribed to three types of causality: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR), based on NA frequency and signal-to-noise patterns (7,8). The missing value issue can be profound, especially in traditional shotgun proteomics where only a fraction of ionized peptides is selected for identification (9–11). Nevertheless, in the last few years, quantitative proteomics underwent a remarkable evolution, yielding a significant increase of protein detection consistency (12,13). This is due to the development of new MS methods and workflows, such as large-scale SRM/PRM measurement (14,15), retention time or spectral library-based MS1 alignment (16–20), multiplexed tandem mass tag labeling (e.g. TMT) (21), and more recently, data independent acquisition mass spectrometry (DIA-MS)

*To whom correspondence should be addressed. Tel: +1 203 737 3853; Fax: +1 203 737 7335; Email: yansheng.liu@yale.edu
Correspondence may also be addressed to Hao Yang. Email: yanghao@scu.edu.cn

exemplified by SWATH-MS (22). For example, researchers have shown consistent detection of thousands of proteins between multiple clinical samples using TMT (1,2), and samples in even larger cohort sizes (i.e. >100–1000s) using library-based MS1 alignment (23) and DIA-MS (24,25). The increased consistency of sensitivity essentially translates to much fewer missing values in the resultant sample vs. protein (or peptide) data matrix, reducing not only MCAR, MAR, but also certain MNAR occurrences.

Although the protein-level missing values has been significantly reduced with the state-of-the-art methodological developments such as DIA-MS approach, NAs are not eliminated in the data. The reasons include, e.g. (a) the scoring of peptide identification in DIA do not always reach statistical significance in every sample (even if the peptide peak group is present) (26), (b) the retention time alignment between a large number of samples might fail due to the LC variations, spray instability, etc. (27,28) and (c) protein false discovery rate (FDR) becomes much more challenging to be controlled when multiple samples are combined (29). (d) Furthermore, post-translational modifications (PTM) oriented proteomic datasets normally feature much more prevalent missing values than the bulk-protein quantification due to additional analytical difficulties. Taking phosphoproteomics as an example, phosphorylated proteins are often low abundant, biologically dynamic, and their quantitative changes are frequently subtle and site specific. Notably, the localization of phosphosite in a peptide sequence requires fragment ions carrying the particular PTM site to be detected, scored and confidently assigned, which is even more challenging for multiple samples. Therefore, missing value imputation is still indispensable for handling proteomics and phosphoproteomic datasets, even if they are generated by DIA-MS. Unfortunately, studies addressing NA imputation for such DIA datasets are currently lacking.

Herein, we aim to provide an efficient, systematic, and easy-to-handle tool that is tailored for quantitative proteomics to deal with NA imputation. Many imputation methods have been developed for omics datasets (30), such as the global approach (e.g. singular value decomposition based imputation (SVD) (31)), local approach (e.g. *k*-nearest neighbours imputation (KNN) (31)), hybrid (e.g. LinCmb (5)) and knowledge assisted approach (32). Furthermore, relevant software packages such as MSnbase (33), IMDE (34), missMS (35), ANPELA (36,37) have been available. These options being available, the bottom-up proteomic quantification is nevertheless based on the measurement of ionized peptide precursors and their fragments derived from a given protein in a biological sample where the proteins are functionally connected. However, none of the available tools have made the usage of such empirical, uniform principle of proteomics to guide the method selection for NA imputation. Other limitations of these tools may include, e.g. the lack of graphic user-friendly interface, the lack of multiple evaluation criteria for the imputed results (38), and the lack of flexibility of handle data structure of proteomics.

In this study, we present an online tool, *NAguideR*, which integrates up to 23 commonly used missing value imputation methods, namely, zero (8), minimum (3,19), column median (39), row median (39), BPCA (38), SVD (31), KNN (31), Seq-KNN (40), trKNN (41), Mice-norm (42), Mice-cart (42), MLE (43), QR (44), Mindet (45), Minprob (45), LLS (46), Impseq (47), Impseqrob (48), IRM (49), RF (50), PI (51), GRR (52), GMS (53) (see Methods and Supplementary Table S1). Most importantly, *NAguideR* provides two categories of evaluation criteria (four classic computational criteria and four empirical proteomics criteria) to assess the imputation performance of various methods. We processed three DIA-MS datasets extensively as examples to exhibit the originality and utility of this software in analyzing phosphoproteomic, peptide and protein level results. Furthermore, we include sufficient biological replicates ($N = 10$ for each study), so that NA evaluation can be performed by referring to the full datasets, benchmarking the robustness of the imputation results in those stimulated datasets with a limited number of replicates (e.g. $N = 3$). Altogether, we found that *NAguideR* recognizes the uniform knowledge in bottom-up proteomics and is helpful in guiding missing value imputation, filling a gap in the pipeline for automated analysis of massive proteomic datasets.

MATERIALS AND METHODS

Data collection and acquisition

Three case-study datasets acquired by DIA-MS (or SWATH-MS) (12,22) were included for testing the availability and capability of *NAguideR*. All the three datasets followed an experimental sampling schema of 10 versus 10 biological replicates. This number of replicates is much more than those in a routine proteomic experiment (e.g. $N = 3$), enabling the estimation reference for experiments with a much smaller number of replicates.

Dataset 1. Phosphoproteomic DIA-MS quantitative dataset for nocodazole treated cells (or 'PhosDIA' in short). The MS samples injected in a previous study (54) for developing the IPF, an algorithm for identifying post-translationally modified peptides, were herein re-measured by a new, powerful Orbitrap Lumos DIA platform (55) for this study. Briefly, for the experimental condition, U2OS cells (about 3–4 million cells per plate) were treated with nocodazole (Sigma-Aldrich), an anti-mitotic drug, at a final concentration of 100 ng/ml for 18 h, which inhibits microtubule dynamics and thus arrests the cell cycle at G_2/M phase. Treated and untreated samples ($N = 10$ replicates, respectively) were collected and processed for protein digestion (54). Phosphopeptide enrichment was performed by using TiO_2 resin (GL Sciences). The final phosphopeptides were desalted using C18 ultramicrospin columns (Nest). Phosphopeptide mixture originating from ∼10% of the starting cell materials per culturing dish was injected for DIA-MS.

The DIA-MS measurements were performed on an established system (55). Briefly, the peptide separation was performed on EASY-nLC 1200 systems (Thermo Scientific) using a self-packed analytical PicoFrit column (New Objective) (75 μm × 30 cm length). The Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Scientific) with a NanoFlex ion source was coupled to the LC platform. Spray voltage was set to be 2000 V and heating capillary

was kept at 275°C. Using the Xcalibur 4.2.47 (Thermo Scientific), the DIA-MS method consisted of a MS1 survey scan and 40 MS2 scans of variable windows. The MS1 scan range was 350–1650 m/z and the MS1 resolution was set to be 120k. The MS1 full scan AGC target value was set to be 2.0E5 and the maximum injection time was 100 ms, The MS2 resolution was set to 30 000 at m/z 200. The MS2 range was set to be 200–1800 m/z and normalized HCD collision energy was 28%. The MS2 AGC was set to be 5.0E5 and the maximum injection time was 50 ms. The default peptide charge state was set to 2. The same samples were also measured by a shotgun analysis following a previously documented method (56).

To analyze the DIA-MS results of 'PhosDIA', Spectronaut software (57,58) was used to generate a spectral library from both shotgun proteomic and DIA acquisitions measuring phosphoproteomic samples. The data was reported by Spectronaut with default settings and a Q value cutoff of 1% at both peptide and protein levels. In particular, the PTM localization score were strictly kept at above 0.75 (59) to ensure the phosphosites are localized with a certainty similar to Class I confidence (59–61). The averaged phosphopeptide enrichment efficiency were determined to be 65.07 ± 5.32% among 20 samples. All the peptide level data were quantified by Spectronaut with default settings and were subjected for NA imputation analysis.

Dataset 2. Protein-level SWATH-MS quantitative dataset for formaldehyde (FA) treated cells (or 'ProtSWATH' in short). This dataset was published in a previous study in which HeLa Kyoto cells were treated with or without 200 μM FA for 5 h (62). The SWATH-MS measurement was performed on a SCIEX 5600 plus TripleTOF instrument. The proteome changes induced by FA treatment was demonstrated to be minimal and specific, with <1% of the detected proteins showed statistically significant reductions ($P < 0.05$, Benjamini–Hochberg adjusted), presenting a challenging case for relative label free quantification at the protein level, for which the proteomic analysis has been already matured (26). To analyze 'ProtSWATH' dataset, a spectral library containing mass spectrometric assays for 10 000 human proteins (63) and 1% peptide and 1% protein-FDR (29) were applied to Spectronaut based analysis (57,58). In particular, the MS2 peak area of the top3 most abundant peptides were averaged and summarized for protein quantification. No PTM score was needed.

Dataset 3. Peptide-level SWATH-MS quantitative dataset for formaldehyde (FA) treated cells (or, 'pepSWATH' in short). This dataset is identical to 'ProtSWATH' but were summarized at the peptide precursor level with 1% FDR. However, to compare the imputed missing values by algorithms of *NAguideR* to imputation from the 'Requantification' option in OpenSWATH (27,64), the peptide precursor level quantities were reported after OpenSWATH analysis with 'Requantification' enabled, which infers the peak boundaries from the closest neighboring run after retention time alignment and quantify the fragment-ion signal within those boundaries (27), as reported before (62).

Missing value imputation methods

To embrace multiple choices for users, a total of 23 published methods for NA imputation were integrated in *NAguideR*. Based on the implementation algorithm of these methods, they can be classified into three types (8,45): (i) *single value methods (SV methods)*, including zero, minimum, column median, row median, Mindet, Minprob, PI, which features replacing missing values by a constant or a randomly selected value; (ii) *global structure methods (GS methods)*, including SVD, BPCA, MLE, Impseq, Impseqrob, which decompose the data matrix or minimize the determinant of the covariance and then iteratively reconstruct the missing values; (iii) *local similarity methods (LS methods)*, including KNN, Seq-KNN, trKNN, LLS, QR, IRM, GRR, GMS, Mice-norm, Mice-cart, RF, which exploit local similarity structure based on the expression profiles of those objects (etc. peptides, proteins) in the data. Additionally, to facilitate the selection, these methods can be also classified into fast (zero, minimum, column median, row median, Mindet, Minprob, PI, SVD, MLE, Impseq, Impseqrob, KNN, Seq-KNN, LLS, QR, GRR) and slow ones (BPCA, trKNN, IRM, GMS, Mice-norm, Mice-cart, RF), based on the calculation of their practical time cost (Supplementary Figure S1). Detailed descriptions about all these 23 algorithms and their implementation of can be found in Supplementary Table S1.

Evaluation Criteria

Four classic criteria and four empirical proteomic criteria are available for evaluating the performance of every imputation method that are implemented independently in *NAguideR*.

Four classic criteria

(a1) Normalized root mean square error (NRMSE) (38). This criterion can evaluate the differences between original values and imputed values and calculated using the following formula:

$$\mathrm{NRMSE} = \sqrt{\frac{\mathrm{mean}(y_o - y_i)^2}{\mathrm{variance}\,(y_o)}} \qquad (1)$$

where y_o means original values and y_i means imputed values. The smaller NRMSE value indicates that the method has better performance for imputation.

(a2) NRMSE based sum of ranks (SOR) (44,52). This criterion is a robust nonparametric measurement, which calculates the rank of NRMSE to compare different imputation methods:

$$\mathrm{SOR} = \sum_{i=1}^{n} \mathrm{Rank}_i\,(\mathrm{NRMSE}) \qquad (2)$$

where Rank_i(NRMSE) indicates the NRMSE ranks of different imputation methods in ith missing variable, n means the total number of missing variables.

(a3) Average correlation coefficient between the original and imputed values (ACC/ACC_OI) (30). By default, the Pearson correlation coefficient is calculated for measuring

how strong a relationship is between the original and imputed values.

(a4) Procrustes statistical shape analysis (PSS) (52,65). This criterion is typically used to assess the similarity of two input matrix through the sum of squared differences. Herein the principal component matrix is extracted from principal component analysis (PCA) as the unsupervised input matrix for evaluating the space alteration of the original sample distribution and the imputed sample distribution.

Four proteomic criteria

(b1) Average correlation coefficient within the different charge states of each peptide (ACC_Charge) (66). This criterion can be deduced by:

$$\text{Peptide}_k = \frac{\sum_{i=1,\ j=2,i\neq j}^{m} \text{cor}\left(\text{Charge}_i,\ \text{Charge}_j\right)}{m} \quad (3)$$

$$\text{ACC\_Charge} = \frac{\sum_{k=1}^{n} \text{Peptide}_k}{n} \quad (4)$$

where the kth peptide has m charge states ($m > 1$), then we calculate the average correlation between every two charge states (Charge_i, Charge_j) of the kth peptide, n means the total number of peptides with multiple charges.

(b2) Average correlation coefficient within the different peptides of each protein (ACC_PepProt) (67). This criterion can be calculated as below:

$$\text{Protein}_k = \frac{\sum_{i=1,\ j=2,i\neq j}^{m} \text{cor}\left(\text{Peptide}_i,\ \text{Peptide}_j\right)}{m} \quad (5)$$

$$\text{ACC\_PepProt} = \frac{\sum_{k=1}^{n} \text{Protein}_k}{n} \quad (6)$$

where the kth protein has m peptides ($m > 1$), then we calculate the average correlation between every two peptides (Peptide_i, Peptide_j) of the kth protein, n means the total number of proteins with multiple peptides.

(b3) Average correlation coefficient within every protein complex based on CORUM database (ACC_CORUM) (68). This criterion can be calculated as below:

$$\text{Complex}_k = \frac{\sum_{i=1,\ j=2,i\neq j}^{m} \text{cor}\left(\text{Protein}_i,\ \text{Protein}_j\right)}{m} \quad (7)$$

$$\text{ACC\_CORUM} = \frac{\sum_{k=1}^{n} \text{Complex}_k}{n} \quad (8)$$

where the kth protein complex has m proteins ($m > 1$), then we calculate the average correlation between every two proteins (Protein_i, Protein_j) of the kth protein complex, n means the total number of complexes with multiple proteins that can be matched in users' proteomics data.

(b4) Average correlation coefficient within each cluster of protein-protein interaction network based on hu.MAP database (ACC_PPI) (69). This criterion can be calculated by:

$$\text{Cluster}_k = \frac{\sum_{i=1,\ j=2,i\neq j}^{m} cor\left(\text{Protein}_i,\ \text{Protein}_j\right)}{m} \quad (9)$$

$$\text{ACC\_PPI} = \frac{\sum_{k=1}^{n} \text{Cluster}_k}{n} \quad (10)$$

where the kth cluster has m proteins ($m > 1$), then we calculate the average correlation between every two proteins (Protein_i, Protein_j) of the kth cluster, n means the total number of clusters with multiple proteins that can be matched in users' proteomics data.

All correlation coefficients in the proteomic criteria were calculated with Pearson method by default. And a larger value, in general, indicates that the imputation method under evaluation has a better performance. After obtaining all values based on each criterion, we divided them by corresponding maximum value and returned their ranks respectively. Four classic criteria can be enabled for different data types (e.g. genomics data, proteomics data, metabolomics data, etc.), while the four proteomic criteria can be particularly applied for proteomics data. Every criterion processes its own distinctive performance evaluation of various imputation methods. Moreover, all imputation results, assessment results and figures are interactively displayed on the web panel, and downloadable for end users. More detailed information can be found in Supplementary Notes.

Tool Implementation

All functions in *NAguideR* were compiled in R (Version 3.6.1, https://www.r-project.org/) (70), and the graphical user interface (GUI) was developed in Shiny (Version 1.2.0, https://github.com/rstudio/shiny). The web tool was deployed on a server with 64GB RAM and Genuine Intel(R) CPU E2687WV running the CentOS Linux release 7.6.1810 (Core) operating system. Users can access and process their own data freely in *NAguideR* without any login requirement through some popular web browsers, such as Google Chrome, Mozilla Firefox, Safari (Supplementary Table S2). In addition, the source codes of *NAguideR* are available on the GitHub repository: https://github.com/wangshisheng/NAguideR/ under the MIT license. Users can choose to operate this tool on their own computers, where the local GUI is working exactly the same as the online version. The detailed installation and operation manual can be found in Supplementary Notes.

Differential expression data stimulation and analysis

Differential expression analysis of two sample groups in each dataset was performed using full datasets or randomly selected observations: (i) We used the full data (10 biological replicates in each group) to construct 'Gold Standard' of differentially expressed proteins/peptides. Furthermore, we randomly selected (ii) five biological replicates ('Random 5') and (iii) three biological replicates ('Random 3') in each group to implement differential expression, and then repeated this process 100 times. Then, the total 100 results for every protein/peptide in the (ii) and (iii) situations were used to infer the biological data fidelity after NA imputation, by comparing to 'Gold Standard' results. To generate volcano plots we used the median values of these 100 results. The statistical significance was tested by two-tailed Student's t-test and the P values were corrected for multiple testing with the Benjamini–Hochberg (BH) method

(71). Proteins/peptides with BH-adjusted $P < 0.05$ and the absolute value of logarithmic fold changes with base 2 ($|Log_2(FCs)| > 0.585$ (i.e. a relative fold change of 1.5 folds) were considered to be differentially expressed. The PTM motif enrichment analysis of differentially regulated phosphosites was performed with motifeR (72), using the comparison between first three samples (according their actual acquisition time which is random in each group) and the 'Gold Standard' result.

Data availability

The new mass spectrometry data of PhosDIA for this study (40 raw files) and all the spectral libraries used have been deposited to the ProteomeXchange Consortium via the PRIDE (73) partner repository with the dataset identifier PXD017476.

RESULTS

Overview of data analysis procedure of *NAguideR*

Basically, there are four main steps in the data analysis process of *NAguideR* (Figure 1 and Supplementary Figure S2): (i) Data upload. In this step, users should upload the original intensity data matrix with NAs (peptides, peptides with certain PTMs, or protein identities in the rows, and sample names in the columns, Figure 1A). (ii) Initial data filtration (optional). Based on the user's choice, these proteins/peptides with excessively high proportion of NA and large coefficient of variation (CV) can be discarded in this step (Figure 1B, and see Supplementary Figure S3 for all the three example datasets). Note these criteria can be optimized iteratively upon the user's trial with *NAguideR* so that satisfactory results can be achieved. Here *NAguideR* also provides a summary note of input data quality regarding completeness before and after the filtration step (Supplementary Notes 3.3). (iii) Missing value imputation. Users can execute and obtain the matrix results of 23 imputation methods from this step (Figure 1C) with a few clicks and minimal parameter selections (Supplementary Notes 4). (iv) Result evaluation. The classic criteria and proteomic empirical criteria (see below) are applied to evaluate every result from step 3. Two comprehensive evaluation tables with ranks of each imputation method are provided to help users select suitable algorithm for their own data. Moreover, for this step, *NAguideR* implements three additional optional functions that are all at user's discretion (Supplementary Notes 5): (a) it enables users to customize the criteria and set relative weightings for specific experimental designs (e.g., if a mixture of protein standards is measured in which no *in-vivo* protein complex formation or interactions are expected); (b) it provides warning messages for users to review if the final imputation results end up with indiscriminate scores across each imputation method following classic or proteomic criteria (as a 'Final check' report); and (c) it allows users to directly visualize the results of a particular peptide or protein item (i.e. spiked-in standard peptides, proteins, or known housekeeping proteins like beta-actin, etc.) before and after imputation (as a 'Targeted check' option, Supplementary Notes 5). All results and figures in all above steps can be downloaded in the format of csv or pdf. Detailed descriptions of each step are shown in Supplementary Figure S2 and Supplementary Notes.

Display of data completeness in three DIA datasets

To gauge the frequency of NA in DIA-MS dataset, we visualized the number of peptides/proteins quantified and missing values in each dataset using plots generated by *NAguideR* (Supplementary Figure S3). Collectively, >50 000 phosphopeptides and peptides were profiled among 20 samples in the two peptide precursor-level datasets (i.e. PhosDIA and PepSWATH) and ∼5000 proteins in ProtSWATH dataset, respectively (Supplementary Table S3). However, both of PhosDIA and PepSWATH had a large proportion of peptides with missing value in at least one sample (76.3% in PhosDIA and 55.1% in PepSWATH). The reason for high missing values in PepSWATH might be stemmed from the large, human proteome-wide assay library being used for peptide identification (63). The highest prevalence of NAs in PhosDIA could be ascribed to the significantly rewired phosphoproteome after nocodazole treatment (54) as well as the extra scoring step of phosphosite localization after peptide identification (59), presenting a most challenging case for NA imputation among the three DIA-MS datasets. In the protein-level example dataset (i.e. ProtSWATH), 4797 proteins were detected and quantified in any of the 20 samples, of which 20.4% contained at least one missing value, suggesting that the missing value problem is partially compromised by protein assignment process (e.g. Top3 summarization, see Methods), Altogether, these results from biological replicates demonstrate a pressing need for missing value imputation for proteomic experiments, even when DIA-MS is used.

Evaluation of imputation methods by correlation-centric analysis

According to our test runs on ProtSWATH dataset, we estimated the time consumption of each NA method and thus chose 16 out of 23 methods that requires less computational procession as default methods in *NAguideR*. The seven methods left, namely BPCA, trKNN, IRM, Mice-norm, Mice-cart, GMS and RF, can be enabled upon the small lists and fast internet speed (Supplementary Figure S1, and Supplementary Notes).

To assess the NA imputation result following each method, for each of the three datasets, we firstly only extracted the complete data matrix from the original datasets and generated random missing values on it with a similar proportion of missing values existed in the original data matrix. This strategy ensured that every imputed data point will have a real reference (i.e. the original value), facilitating the comparison between imputation methods as well as the comparison between evaluation standards. All 23 imputation methods (Supplementary Table S1) were conducted on all datasets. After imputations, we compared the original values and imputed values with Pearson correlation analysis and density plots (Figure 2A and Supplementary Figure S4 for PhosDIA data, Supplementary Figure S5A for PepSWATH data, Supplementary Figure S6A for ProtSWATH data). We found that

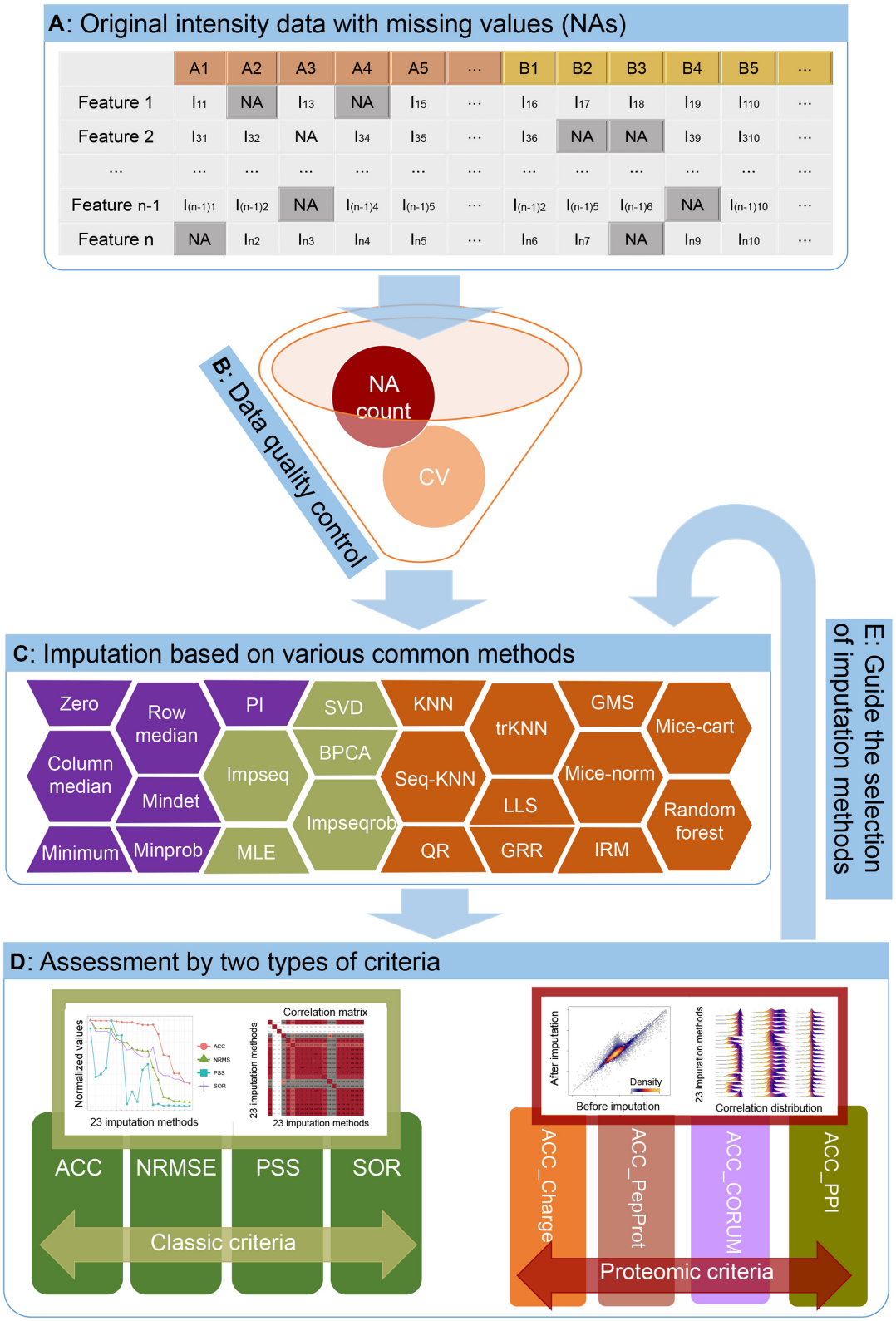

Figure 1. The overall workflow of *NAguideR*. (**A**) Uploading of original proteomics data with missing values (NAs). (**B**) Optional data quality control step for removing proteins/peptides with high proportion of NAs or large CV. (**C**) Missing value imputation based on the embedded methods. (**D**) Performance evaluation by multiple criteria (four classic criteria and four proteomic criteria). (**E**) The selection of well-performed imputation methods guided by the classic criteria and proteomic criteria.

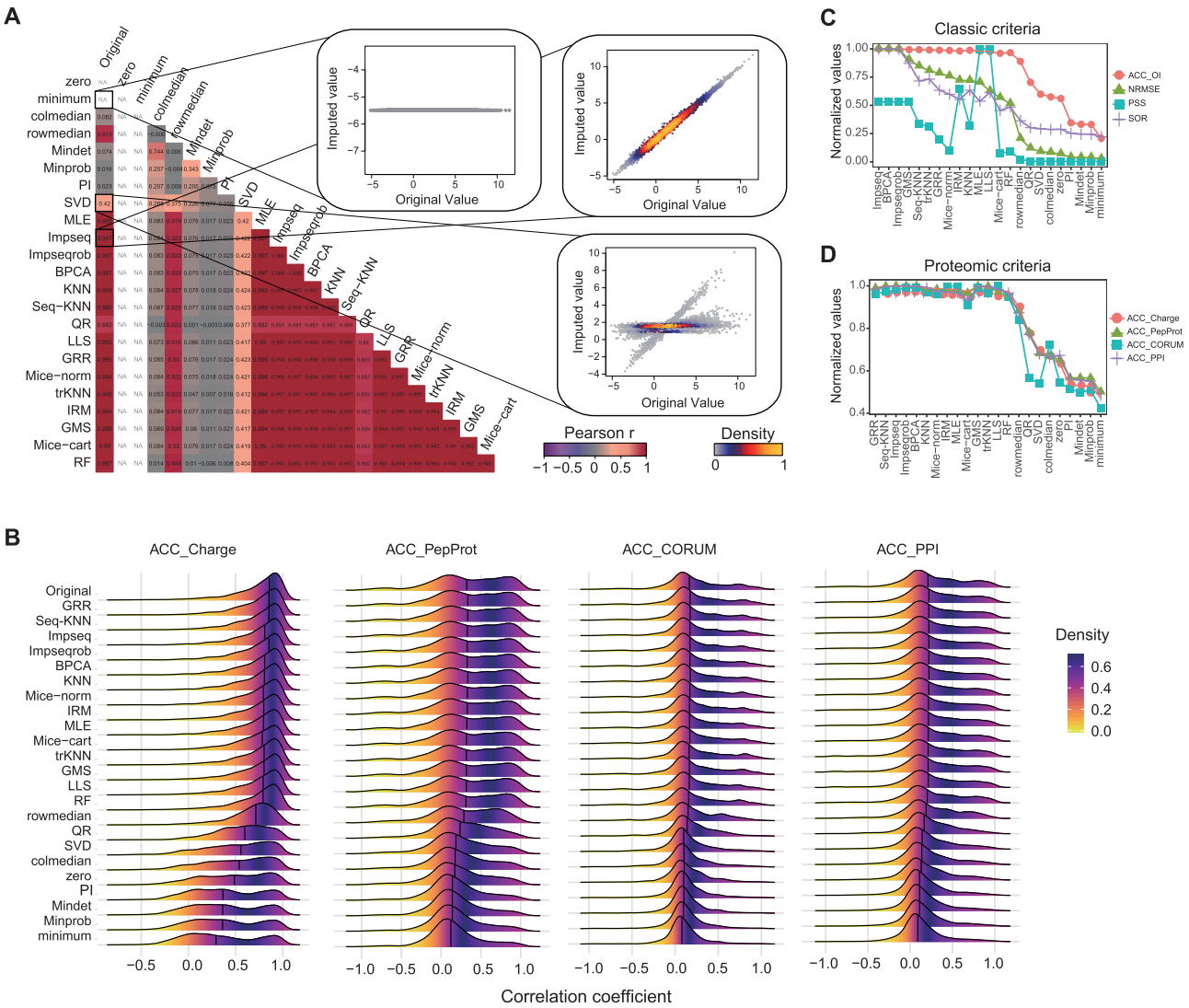

Figure 2. Systematic evaluation analysis of PhosDIA dataset. (**A**) Pearson correlation analysis of the original intensities and imputed intensities based on 23 methods. Density plots illustrate the correlation in detail between the original values and imputed values from minimum, SVD, and Impseq respectively as examples. NA in the correlation matrix means 'No Result' because the standard deviations of imputed values from zero and minimum method are equal to 0, and hence the cor function returns NA. (**B**) Comparison of the distribution of the correlation coefficient among original values and 23 imputation methods under the four proteomic criteria. The comprehensive scores distribution of 23 imputation methods under the four classic criteria (**C**) and four proteomic criteria (**D**). 'Normalized values' here means every score is divided by the corresponding maximum value.

certain imputation algorithms (e.g. Impseq, BPCA, Seq-KNN and GRR) could obtain higher correlation coefficients than others (e.g. Minprob, minimum, zero and PI) across all three datasets, suggesting that certain NA methods maybe preferable for proteomic datasets, based on the classic correlation profiling between original and imputed values.

Empirical proteomic principles facilitate the selection of NA algorithm

Following, we asked if the empirical bottom-up proteomic principles can be employed to inspect the imputation outcome. We extracted and tested correlations between peptide or protein entries based on quantitative consistency between different charge-states of the same peptide, different peptides belonging to the same proteins, and individual proteins participating functional complexes and interactions (see the example of correlation between charge states in Supplementary Figure S7 and Methods). Similar but more discriminative results compared to the correlation matrix above were obtained, which was based on the correlation coefficient distribution following proteomic criteria, supporting that the distributions from such as GRR, Seq-KNN, Impseq, BPCA were more similar to the original results (Figure 2B for PhosDIA data, Supplementary Figure S5B for PepSWATH data and Supplementary Figure S6B for ProtSWATH data). We then compared the proteomic criteria to the four classic computational criteria for NA imputation, namely NRMSE, SOR, ACC_OI and PSS, using the ranked normalized scores for each method (Figure 2C, D for PhosDIA data, Supplementary Figure S5C-S5D

for PepSWATH data, and Supplementary Figure S6C, D for ProtSWATH data). Table 1 lists the corresponding ranking results in all three datasets, showing a consistent result that Impseq, Impseqrob, BPCA, trKNN, Seq-KNN and GRR were top-ranked. Interestingly, although both criteria were able to recognize a few favorable NA imputation algorithms in each dataset, the proteomic criteria based on varied mass spectrometric or biological rules generated more consistent evaluation between criteria than the classic criteria metrics. Considering the additional benefits of applying proteomic criteria, such as direct application, easy biological interpretation, and the facilitated communication between researchers, we deem the proteomic criteria efficiently help the user to select NA imputation method. Therefore, we have included both classic criteria and the above four proteomic criteria results in *NAguideR*.

Furthermore, we assessed the robustness of proteomic criteria in evaluating the outcome of NA imputation. To do this, we randomly produced missing values on the three datasets from the proportion of 5–70% at the whole data matrix level, and in step of 5%. At each step, we repeated the imputation and evaluation process. The results (Supplementary Figure S8A) suggested that the top-ranked methods performed well and consistently across all varying missing proportions in PhosDIA data, and that the differences of scores resulted from different methods under the four proteomic criteria became larger as missing proportions increased. Similar results were obtained from PepSWATH (Supplementary Figure S8B) and ProtSWATH datasets (Supplementary Figure S8C), which are both deemed less challenging than PhosDIA dataset considering their NA prevalence (Supplementary Figure S3). Thus, proteomic criteria are robust to datasets with different extent of NA prevalence.

In a typical proteomic experimental design, a limited number of biological replicates such as $N = 3$ is frequently used. Hence, in all three datasets, we evaluated the robustness of *NAguideR* results by simply using the first three samples of each group (injected with a random order, thus presenting a stopping point for a routine $N = 3$ proteomic investigation). As shown in Figure 3, we found the four proteomic criteria used in *NAguideR* can provide a discriminant score estimation for each imputation method in the '3 versus 3' datasets. And the resultant score distribution between different NA methods largely agrees to the results from '10 versus 10' datasets. Additionally, both classic and proteomic criteria yielded similar ranking results (Table 1, Supplementary Figure S9). Altogether, these results support the feasibility of *NAguideR* and its proteomic criteria in dealing with experiments with limited biological replicates.

In summary, we introduced effective, proteomic principle-derived criteria for estimating the performance of different NA imputation methods, which shows robustness in datasets of varied NA prevalence and limited biological replicates.

Direct application of *NAguideR* facilitates relative proteomic quantification

The above analysis based on referencing the *in-silico* deleted original values demonstrated the usage of *NAguideR* and its proteomic criteria, paving the way to address technical and biological questions in label-free quantification. Previously, the 'Requantification' step from OpenSWATH (27,64) and TRIC algorithms was used to impute the missing data at MS2 level for DIA-MS. 'Requantification' essentially infers the MS2 peak boundaries from the closest neighboring run after retention time alignment and quantifies the fragment-ion signal within those boundaries (27). We therefore compared 'Requantification' to the 23 NA imputation methods supported in *NAguideR*. To survey whether the imputed data points added more variation to the quantification, we plotted intra charge-state correlation of a given peptide precursor quantified across samples before and after NA imputation, following the direct application of different imputation methods including 'Requantification' (Figure 4). This across sample correlation can be also assessed by its average variability for all peptides (Supplementary Figure S10). The results interestingly indicate that the 'Requantification' method only ranked in the middle among the established 23 imputation methods (Figure 4 and Supplementary Figure S10). Thus, alternative NA imputation algorithms should be considered for DIA-MS, such as those provided by *NAguideR* under the four proteomic criteria.

We next address how different imputation methods impact differential expression analysis. Applying *NAguideR* on all three DIA-MS datasets, we obtained completed data matrixes which can be then analyzed by the standard student t-tests between experimental and control groups. Herein we focused on the PhosDIA dataset, which profiled the phosphoproteome following the cell cycle arrest that was well-studied (54,74). Volcano plots in Figure 5A–F and Supplementary Figure S11 illustrated that, the *P*-values between groups for the same original dataset (i.e. PhosDIA) but imputed by different methods can be distinctive. Accordingly, the volcano shapes derived from top-ranked imputation methods (e.g. Impseq, Seq-KNN) were similar to that from 'Gold Standard' (i.e. the original full dataset without NA imputed). In stark contrast, the shapes from low-ranked methods (e.g. minimum) revealed significantly skewed *P*-values and therefore reduced efficiency in determining differential expression. Subsequently, from $N = 10$ replicates per group in PhosDIA, we randomly selected five or three observations by 100 times per group (i.e. 'Random 5' and 'Random 3') and performed the same statistical test. As expected, less-individuals per group reduce statistical significance (Figure 5A–F, median from 100 selections). We noticed that the worst NA imputations (such as minimum) often presented bifurcate and applanate volcano patterns. To further depict the differences in the number of differentially expressed peptides we stimulated the differential phosphopeptide lists based on all 'Random 5' and 'Random 3' selections respectively (Figure 5G). The results intriguingly indicate (a) both the inefficient NA imputation methods and the low biological replicates could weaken the capacity and power of detecting differential expression peptides; (b) the non-suitable imputation methods can significantly impair the differential expression analysis, even with $N = 10$ replicates (e.g. by reporting <15% significant phosphopeptide identities between groups). (c) With an ideal NA imputation, five biological replicates ($N = 5$ per group) may

Table 1. Evaluation ranks of 23 imputation methods on the basis of the classic criteria and the proteomic criteria applied to the three example datasets (i.e. PhosDIA, PepSWATH, and ProtSWATH)

Datasets	PhosDIA				PepSWATH				ProtSWATH			
Groups	10 vs. 10		3 vs. 3		10 vs. 10		3 vs. 3		10 vs. 10		3 vs. 3	
Methods	Rank 1*	Rank 2	Rank 1	Rank 2	Rank 1	Rank 2	Rank 1	Rank 2	Rank 1	Rank 2	Rank 1	Rank 2
Impseq	2.375	3.5	4.25	1.75	6	2.75	4.875	2.625	1.25	4.5	5.625	5.5
Impseqrob	2.875	4	4.5	1.5	5	3.25	4.375	4.25	2.375	2.25	2.625	8
BPCA	3.375	4.5	4	2.75	5.5	2.25	4.125	2.125	2.875	5.5	6.875	5.5
GMS	4.375	10.75	7.75	11	5.25	7.25	6.125	8	6.375	10.5	7.625	10
Seq-KNN	6	2.5	5	4	5.25	3.75	6.875	5	6.125	7	7.375	3
trKNN	6.75	10.5	8.875	9.375	3	6.25	6.5	9	3.5	10	8.375	14.5
GRR	8.25	1.5	6.5	5	6.75	8.5	4.625	5.25	7.75	5.5	3.875	6
IRM	8.25	8.75	6.5	8.75	10.25	13.5	12.375	11.5	14	8	10.875	15
KNN	8.5	5	10.75	11	7.75	6.25	11.125	8	7.25	4.75	8.625	3
LLS	8.625	13	4.75	7.875	10	10.25	9.625	10.75	10.75	12	9.625	15.5
MLE	8.875	9.5	7.25	7.5	12	14	11.625	13.25	12.25	8.5	11.125	1
Mice-norm	9.75	7.25	16.5	13.75	10.5	6.75	11	13	12.25	9.5	16.875	13
RF	13.25	14	7.875	7.5	8.5	11.5	4.875	6.75	8	10.5	5.875	8.5
Mice-cart	13.75	10.25	15.25	13.5	11.75	9.75	13.25	10	14.75	8	17.125	21.5
rowmedian	15	15	14.25	14.75	12.75	14.25	8.625	11	10.5	13.5	4.375	13
QR	16.25	16	16.25	16.75	15.75	15.75	16	16.25	16	16	12.125	9.5
SVD	17.5	17.5	17.75	16.25	19.5	19.5	18.5	16.25	19	18	18.5	20.5
colmedian	17.75	18	17.75	18.5	17.75	18.75	18.5	18.25	18	18	17.5	20
zero	18.5	18.5	17.25	18.5	17.75	18.25	18	18.75	18	19.5	18	22
PI	20	20	19.25	21	19.75	20.75	19.75	20	21.75	19.5	21.125	18
Mindet	21.25	21.25	20.5	21.25	20.75	19.5	21.375	22	20.25	21.25	20.125	14.5
Minprob	21.75	21.75	21	20.75	21.75	20.25	21.125	21	20.25	21.75	19.875	13.5
minimum	23	23	22.25	23	22.75	23	22.75	23	22.75	22	21.875	15

* Rank 1 indicates the mean rank of 23 imputation methods based on four classic criteria; Rank 2 indicates the mean rank of 23 imputation methods based on four proteomic criteria. '10 vs. 10' means there are 10 replicates in each group, and '3 vs. 3' means there are 3 replicates in each group. The top 5 methods are marked with different colors: first, second, third, fourth, fifth.

be already sufficient in this PhosDIA dataset tested, because the $N = 5$ comparison reported similar number of significant identities compared to the $N = 10$ scenario. Finally, because of the extensively investigated phosphoproteomic change following nocodazole treatment, we compared the motif enriched in the lists of differential phosphopeptides. Similar motifs to a previous study (74) were identified. Herein, this motif enrichment analysis is helpful to discern if those best NA methods are too aggressive in reporting regulated phosphopeptides. We found that 'Seq-KNN' (an example of the best NA methods) essentially identified 1494 (i.e. >35%) more phosphopeptide as the significantly regulated hits than 'zero' identified (an example of the worst NA methods), if we use the data of first three biological replicate samples based on MS injection time for both methods. Nevertheless, the separated motif analysis of these additional 35% phosphospeptides by 'Seq-KNN' yielded 13 motifs, 12 of which can be identified by $N = 10$ replicates in 'Golden standard' dataset (Supplementary Figure S12). This result thus suggests that those better NA methods supported by *NAguideR* can efficiently facilitate differential expression analysis and biological research.

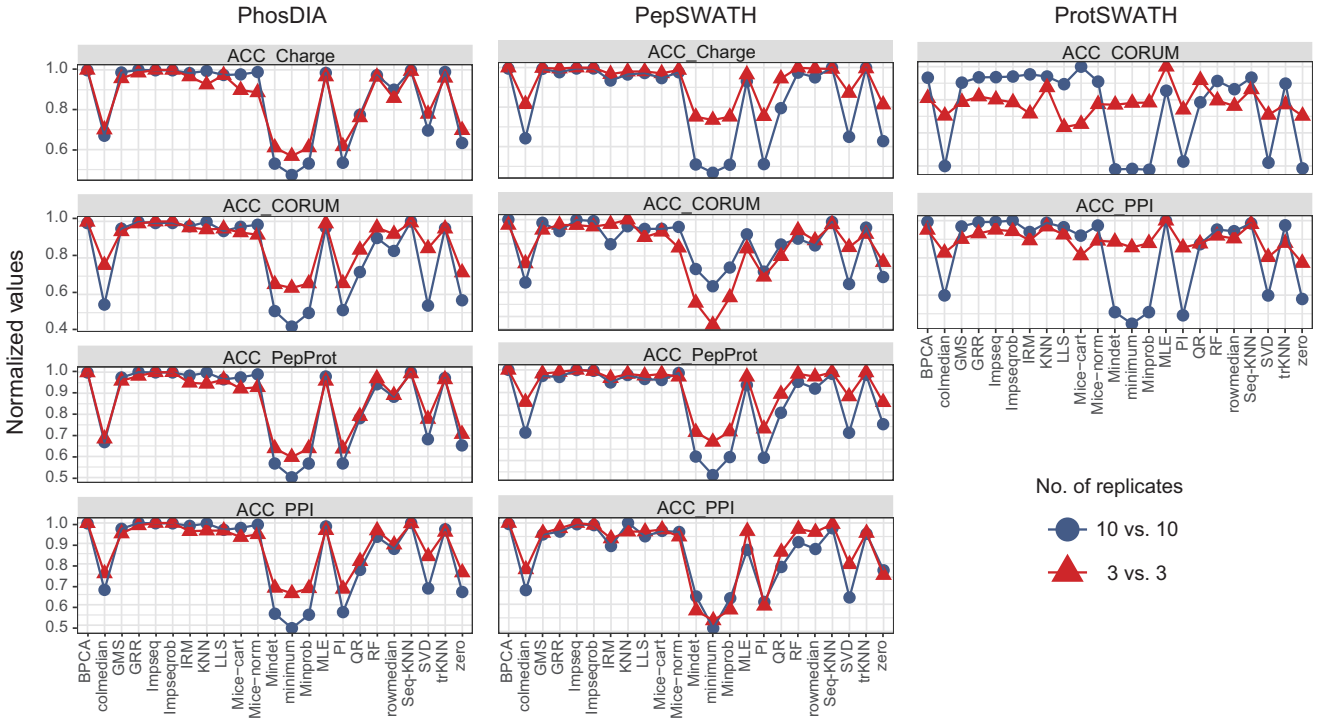

Figure 3. The score distribution of every imputation method based on the proteomic criteria in the three proteomics datasets with different biological replicates. Left panel: PhosDIA, middle: PepSWATH, right: ProtSWATH. 'Normalized values' denotes that every score is divided by corresponding maximum value. '10 versus 10' means that there are 10 replicates in each group (marked with darkblue color), and '3 versus 3' means that there are three replicates in each group (marked with red color).

In summary, the direct application of *NAguideR* promotes prioritizing both NA imputation method (such as 'Requantification') and protein/peptide candidates with real biological regulations.

DISCUSSION

Reproducibility is a cornerstone for scientific research. The MS-based proteomics, however, often generates missing-value datasets between samples and conditions. Despite of the recent technical developments such as DIA-MS, missing values still present a major problem especially in MS datasets profiling protein PTMs. To date, efficient tools that are tailored for proteomics community are rather limited in this regard. In this study, we developed an open source and user-friendly toolkit, *NAguideR*, which implemented 23 missing value imputation methods that are frequently used and eight evaluation criteria, aiming to help scientists select the most appropriate imputation methods during data analysis. We made *NAguideR* to be conveniently accessed through both web tool and stand-alone software version, depending on the data size and internet speed.

There are two main aspects that we consider when choosing these imputation methods: First, all methods should be commonly applied and implemented in many peer-reviewed packages, e.g. MSnbase (33), impute (31), GMSimpute (53); second, as the missing values in proteomics data are generated following different complex mechanisms, these methods should include various families of imputation procedures, which can be potentially used for diverse types of missing values. For example, kNN (40), MLE (43) were proposed to be functional in imputing MCAR/MAR values; MinDet (45) and MinProb (45) were designed initially for handling MNAR values, while GMS (53) does not require specific designation of missing values pattern. In addition, different methods may have their advantages and disadvantages. For example, *SV methods* are relatively simple and fast for large-cohort experiments, but they may introduce severe bias in data and fail to meet certain hypotheses of statistical tests. On the other hand, *GS methods* and *LS methods* generally perform better, but *GS methods* assume the existence of a global covariance structure among all samples or objects (i.e. proteins/peptides/genes) and *LS methods* assume that a strong local correlation exists between objects in the expression matrix. Thus, when the assumptions are not appropriate, their imputation may become less accurate (Supplementary Table S1). Users of *NAguideR* can adjust the method selection based on these advantages and disadvantages. Of note, the practical proteomic datasets could be highly heterogenous between users due to the different sample types, quality, experimental designs, mass spectrometers used and etc. There is unlikely a one-fits-all solution for imputing NAs in all variable datasets. Thus, besides a simple 'Input data check' of data quality such as NA prevalence and data variation as well as a 'Final check' about result heterogeneity, *NAguideR* implements the 23 NA imputation methods without preference. Users can therefore compare the imputation results of different algorithm through

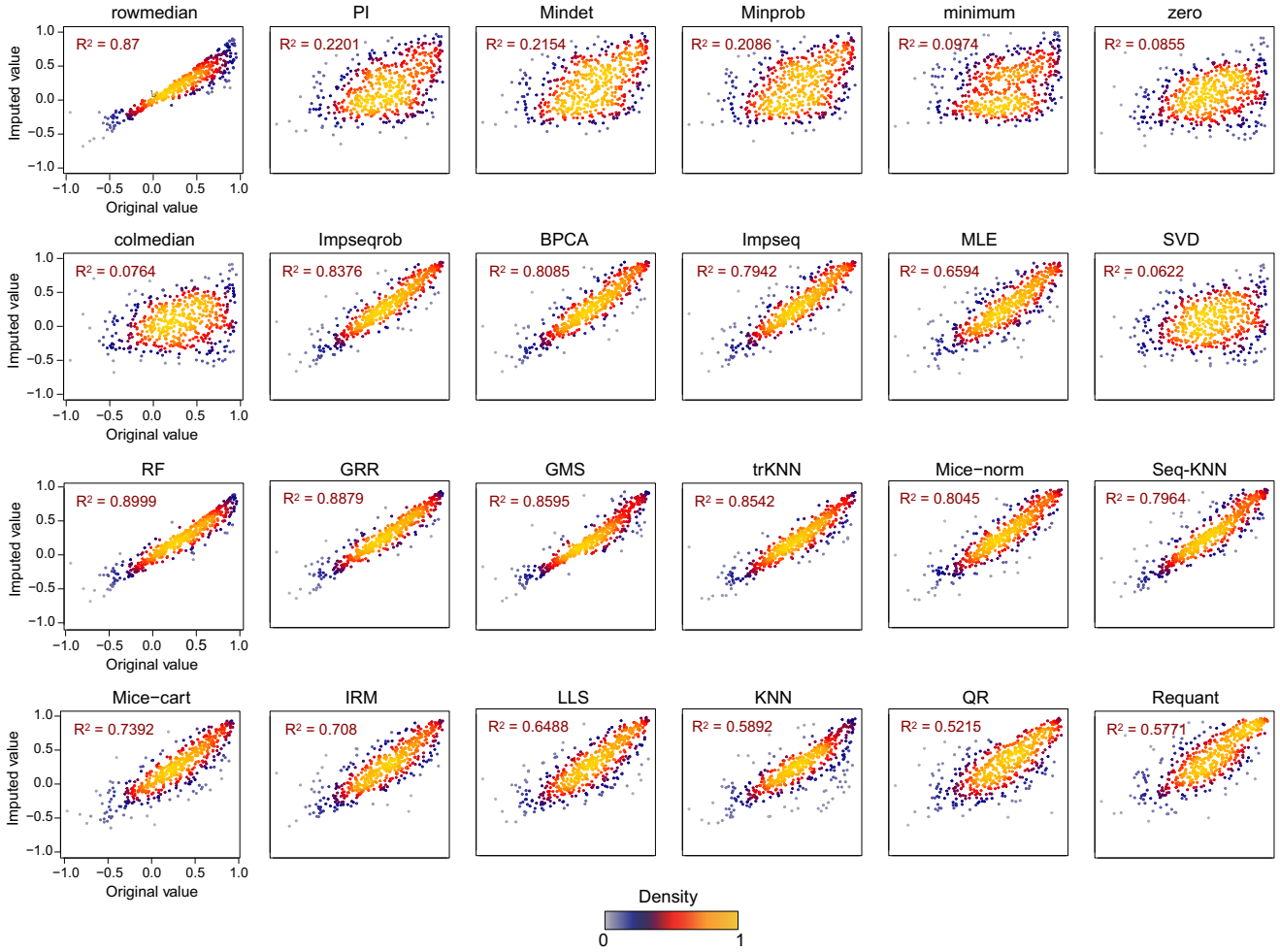

Figure 4. Across sample, quantitative correlation coefficients obtained by different NA imputation methods. Comparisons of original values and imputed values of the quantitative correlation coefficients are shown which are derived under ACC_Charge criterion by the 23 imputation methods and 'Requantification' method for the pepSWATH dataset. The adjusted *R* squared (R^2) of each result was also obtained by 'lm' function and shown for every imputation method.'Requant' denotes 'Requantification' method in OpenSWATH software.

global correlation scores as well as individual data inspection (e.g. through 'Targeted check' option) for selecting a method preferable to their own data.

Besides algorithm integration and flexible implementation, the evaluation step of *NAguideR* is a significant added value compared other solutions such as the individual R-packages, because this step uniquely guides users to select NA imputation method using common rules of bottom-up proteomics by visualizing their own data before and after imputation. The four proteomic criteria were found to generate moderate correlation coefficients that are potentially more discriminative than the extremely skewed correlations between original and imputed results (e.g. those in Figure 2A). The four proteomic criteria can be directly applied and inspected to an entire dataset, avoiding the potential bias of those computational evaluations focusing on those peptide/protein entries (with no NAs at all) whose concentrations tend to be more abundant in the human proteome. Moreover, two peptide-level criteria (i.e. correlation between different charge-states of the same peptide and between different peptides belonging to the same proteins) generated quite consistent results to the two protein-level metrics (i.e., correlation coefficient within each protein complex and within cluster of protein–protein interaction network) in our tested datasets, suggesting *NAguideR* could generate reliable results in selecting NA methods for both peptide- and protein-level data (Table 1). In addition, the usage of *NAguideR* was evaluated to be robust in data with high NA prevalence and with limited numbers of biological replicates (Figure 3 and Supplementary Figure S8), and may facilitate the biological investigation involving differential proteomic and phosphoproteomic measurements (Figure 5G and Supplementary Figure S12). Interestingly, several NA imputation methods such as Seq-KNN, Impseqrob, and Impseq offered better results than those sub-optimal and low-performance algorithms in all DIA datasets (including the simulated datasets with limited biological replicates), underscoring their value in future proteomic analysis. It should be stressed that, because of the multiple NA method integration, NAguideR provides the opportunity to reveal if there are certain methods that are mutually comparable, but significantly better than others.

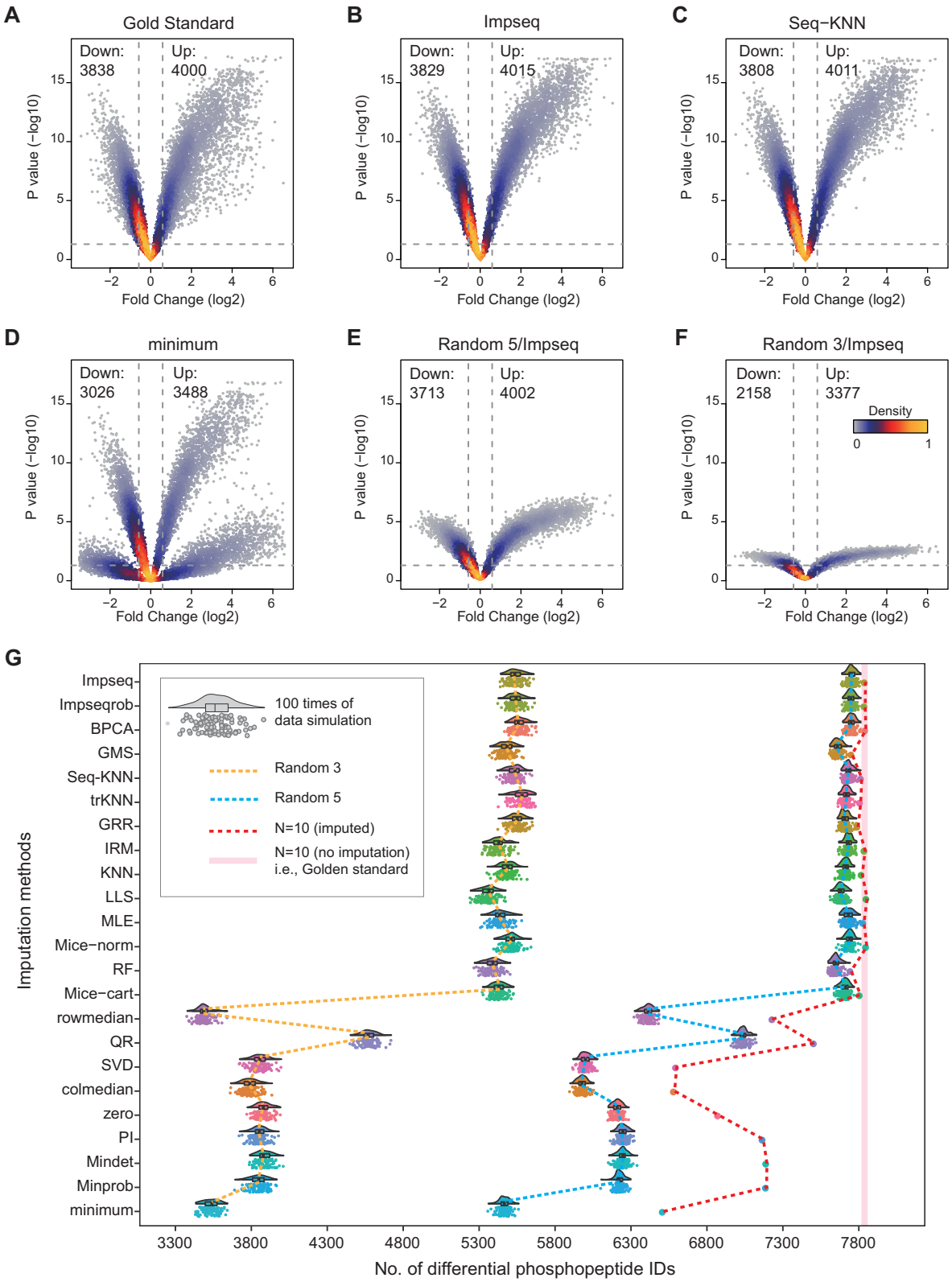

Figure 5. Differential expression and simulation analysis of PhosDIA dataset. Volcano plots of original full data (labelled as 'Gold Standard') (**A**), imputed data from Impseq method (**B**), Seq-KNN method (**C**), minimum method (**D**), imputed data of randomly selected five biological replicates (labelled as 'Random 5') (**E**) and 3 biological replicates (labeled as 'Random 3') (**F**) in each group from Impseq method. ('Down' means down-regulated phosphopeptides, 'Up' means up-regulated phosphopeptides). (**G**) Cloud-rain plots indicating the number of differentially expressed peptides for the 100 randomly selected datasets by 'Random 5' and 'Random 3'. Solid pink line means the number of differentially expressed peptides from gold standard samples. Dashed lines of red, blue and yellow indicate the distribution of the numbers of differentially expressed peptides from each imputation method with all, Random 5 and Random 3 samples, respectively.

This might implicate a fact that applying one of these favorable NA methods could be sufficient for many datasets. Finally, *NAguideR* reserves the potential of updating current methods, integrating additional methods and assessment criteria in the future, and can be useful in aiding the bioinformatic efforts developing new NA algorithms.

We anticipate that *NAguideR* could greatly facilitate the multi-omics studies especially the proteomic research in dealing with NA issue and assist biologists or clinicians with less computational background in analyzing samples at a high throughput.

DATA AVAILABILITY

NAguideR is an open source platform, which initiative available from: https://github.com/wangshisheng/NAguideR under the MIT license. The detailed tutorial about this tool can also be found here: https://github.com/wangshisheng/NAguideR/blob/master/NAguideR_Manual.pdf.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We gratefully thank Dr Chengpin Shen for configuring the network server and Dr Hongwen Zhu for feedback on the manuscript and helpful discussions.

FUNDING

National Natural Science Foundation of China [81871475 to H.Y.]; 1.3.5 project for disciplines of excellence, West China Hospital, Sichuan University [ZYGD18014], Sichuan, China; Y.L. was supported by a pilot grant from Cancer Systems Biology@Yale (CaSB@Yale) and a pilot grant from Yale Cancer Center in the Yale University, CT, USA. Funding for open access charge: Yale University.
Conflict of interest statement. None declared.

REFERENCES

1. Clark,D.J., Dhanasekaran,S.M., Petralia,F., Pan,J., Song,X., Hu,Y., da Veiga Leprevost,F., Reva,B., Lih,T.M., Chang,H.Y. *et al.* (2019) Integrated proteogenomic characterization of clear cell renal cell carcinoma. *Cell*, **179**, 964–983.
2. Gao,Q., Zhu,H., Dong,L., Shi,W., Chen,R., Song,Z., Huang,C., Li,J., Dong,X., Zhou,Y. *et al.* (2019) Integrated proteogenomic characterization of HBV-related hepatocellular carcinoma. *Cell*, **179**, 561–577.
3. Jiang,Y., Sun,A., Zhao,Y., Ying,W., Sun,H., Yang,X., Xing,B., Sun,W., Ren,L., Hu,B. *et al.* (2019) Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature*, **567**, 257–261.
4. Moorthy,K., Saberi Mohamad,M. and Deris,S. (2014) A review on missing value imputation algorithms for microarray gene expression data. *Curr. Bioinformatics*, **9**, 18–22.
5. Jornsten,R., Wang,H.Y., Welsh,W.J. and Ouyang,M. (2005) DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics*, **21**, 4155–4161.
6. Stead,D.A., Paton,N.W., Missier,P., Embury,S.M., Hedeler,C., Jin,B., Brown,A.J. and Preece,A. (2008) Information quality in proteomics. *Brief Bioinform.*, **9**, 174–188.
7. Karpievitch,Y.V., Dabney,A.R. and Smith,R.D. (2012) Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics*, **13**(Suppl.16), S5.
8. Lazar,C., Gatto,L., Ferro,M., Bruley,C. and Burger,T. (2016) Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *J. Proteome Res.*, **15**, 1116–1125.
9. Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.
10. Bell,A.W., Deutsch,E.W., Au,C.E., Kearney,R.E., Beavis,R., Sechi,S., Nilsson,T., Bergeron,J.J. and Group,H.T.S.W. (2009) A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat. Methods*, **6**, 423–430.
11. Domon,B. and Aebersold,R. (2010) Options and considerations when selecting a quantitative proteomics strategy. *Nat. Biotechnol.*, **28**, 710–721.
12. Aebersold,R. and Mann,M. (2016) Mass-spectrometric exploration of proteome structure and function. *Nature*, **537**, 347–355.
13. Collins,B.C., Hunter,C.L., Liu,Y., Schilling,B., Rosenberger,G., Bader,S.L., Chan,D.W., Gibson,B.W., Gingras,A.C., Held,J.M. *et al.* (2017) Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat. Commun.*, **8**, 291.
14. Picotti,P. and Aebersold,R. (2012) Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat. Methods*, **9**, 555–566.
15. Kusebauch,U., Campbell,D.S., Deutsch,E.W., Chu,C.S., Spicer,D.A., Brusniak,M.Y., Slagel,J., Sun,Z., Stevens,J., Grimes,B. *et al.* (2016) Human SRMAtlas: a resource of targeted assays to quantify the complete human proteome. *Cell*, **166**, 766–778.
16. Meier,F., Geyer,P.E., Virreira Winter,S., Cox,J. and Mann,M. (2018) BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat. Methods*, **15**, 440–448.
17. Shen,X., Shen,S., Li,J., Hu,Q., Nie,L., Tu,C., Wang,X., Poulsen,D.J., Orsburn,B.C., Wang,J. *et al.* (2018) IonStar enables high-precision, low-missing-data proteomics quantification in large biological cohorts. *PNAS*, **115**, E4767–E4776.
18. Cox,J., Hein,M.Y., Luber,C.A., Paron,I., Nagaraj,N. and Mann,M. (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics*, **13**, 2513–2526.
19. Johansson,A., Enroth,S., Palmblad,M., Deelder,A.M., Bergquist,J. and Gyllensten,U. (2013) Identification of genetic variants influencing the human plasma proteome. *PNAS*, **110**, 4673–4678.
20. Pasa-Tolic,L., Masselon,C., Barry,R.C., Shen,Y. and Smith,R.D. (2004) Proteomic analyses using an accurate mass and time tag strategy. *BioTechniques*, **37**, 621–624.
21. Thompson,A., Schafer,J., Kuhn,K., Kienle,S., Schwarz,J., Schmidt,G., Neumann,T., Johnstone,R., Mohammed,A.K. and Hamon,C. (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.*, **75**, 1895–1904.
22. Gillet,L.C., Navarro,P., Tate,S., Rost,H., Selevsek,N., Reiter,L., Bonner,R. and Aebersold,R. (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**, O111.016717.
23. Niu,L., Geyer,P.E., Wewer Albrechtsen,N.J., Gluud,L.L., Santos,A., Doll,S., Treit,P.V., Holst,J.J., Knop,F.K., Vilsboll,T. *et al.* (2019) Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease. *Mol. Syst. Biol.*, **15**, e8793.
24. Bruderer,R., Muntel,J., Muller,S., Bernhardt,O.M., Gandhi,T., Cominetti,O., Macron,C., Carayol,J., Rinner,O., Astrup,A. *et al.* (2019) Analysis of 1508 plasma samples by capillary flow data-independent acquisition profiles proteomics of weight loss and maintenance. *Mol. Cell Proteomics*, **18**, 1242–1254.
25. Liu,Y., Buil,A., Collins,B.C., Gillet,L.C., Blum,L.C., Cheng,L.Y., Vitek,O., Mouritsen,J., Lachance,G., Spector,T.D. *et al.* (2015) Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.*, **11**, 786.
26. Navarro,P., Kuharev,J., Gillet,L.C., Bernhardt,O.M., MacLean,B., Rost,H.L., Tate,S.A., Tsou,C.C., Reiter,L., Distler,U. *et al.* (2016) A multicenter study benchmarks software tools for label-free proteome quantification. *Nat. Biotechnol.*, **34**, 1130–1136.

27. Rost,H.L., Liu,Y., D'Agostino,G., Zanella,M., Navarro,P., Rosenberger,G., Collins,B.C., Gillet,L., Testa,G., Malmstrom,L. *et al.* (2016) TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. *Nat. Methods*, **13**, 777.
28. Gupta,S., Ahadi,S., Zhou,W. and Rost,H. (2019) DIAlignR provides precise retention time alignment across distant runs in DIA and targeted proteomics. *Mol. Cell. Proteomics*, **18**, 806–817.
29. Rosenberger,G., Bludau,I., Schmitt,U., Heusel,M., Hunter,C.L., Liu,Y., MacCoss,M.J., MacLean,B.X., Nesvizhskii,A.I., Pedrioli,P.G.A. *et al.* (2017) Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. *Nat. Methods*, **14**, 921–927.
30. Liew,A.W., Law,N.F. and Yan,H. (2011) Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Brief Bioinform*, **12**, 498–513.
31. Troyanskaya,O., Cantor,M., Sherlock,G., Brown,P., Hastie,T., Tibshirani,R., Botstein,D. and Altman,R.B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.
32. Xiang,Q., Dai,X., Deng,Y., He,C., Wang,J., Feng,J. and Dai,Z. (2008) Missing value imputation for microarray gene expression data using histone acetylation information. *BMC Bioinformatics*, **9**, 252.
33. Gatto,L. and Lilley,K.S. (2012) MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, **28**, 288–289.
34. Chiu,C.-C. and Wu,W.-S. (2014) In: *11th IEEE International Conference on Control & Automation (ICCA)*. IEEE, pp. 511–514.
35. O'Brien,J.J., Gunawardena,H.P., Paulo,J.A., Chen,X., Ibrahim,J.G., Gygi,S.P. and Qaqish,B.F. (2018) The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. *Ann. Appl. Stat.*, **12**, 2075.
36. Tang,J., Fu,J., Wang,Y., Luo,Y., Yang,Q., Li,B., Tu,G., Hong,J., Cui,X. and Chen,Y. (2019) Simultaneous improvement in the precision, accuracy and robustness of label-free proteome quantification by optimizing data manipulation chains. *Mol. Cell Proteomics*, **RA118**, 001169.
37. Tang,J., Fu,J., Wang,Y., Li,B., Li,Y., Yang,Q., Cui,X., Hong,J., Li,X. and Chen,Y. (2020) ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Brief Bioinform*, **21**,621–636.
38. Oba,S., Sato,M.-a., Takemasa,I., Monden,M., Matsubara,K.-i. and Ishii,S. (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**, 2088–2096.
39. Dimitriadou,E., Hornik,K., Leisch,F., Meyer,D. and Weingessel,A. (2008) Misc functions of the Department of Statistics (e1071), TU Wien. *R Package*, **1**, 5–24.
40. Kim,K.-Y., Kim,B.-J. and Yi,G.-S. (2004) Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinformatics*, **5**, 160.
41. Shah,J.S., Rai,S.N., DeFilippis,A.P., Hill,B.G., Bhatnagar,A. and Brock,G.N. (2017) Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical metabolomics studies. *BMC Bioinformatics*, **18**, 114.
42. Buuren,S.v. and Groothuis-Oudshoorn,K. (2010) mice: multivariate imputation by chained equations in R. *J. Stat. Softw.*, **45**, doi:10.18637/jss.v045.i03.
43. Ibrahim,J.G., Chen,M.-H., Lipsitz,S.R. and Herring,A.H. (2005) Missing-data methods for generalized linear models: a comparative review. *J. Am. Statist. Assoc.*, **100**, 332–346.
44. Wei,R., Wang,J., Su,M., Jia,E., Chen,S., Chen,T. and Ni,Y. (2018) Missing value imputation approach for mass spectrometry-based metabolomics data. *Sci. Rep.*, **8**, 663.
45. Webb-Robertson,B.-J.M., Wiberg,H.K., Matzke,M.M., Brown,J.N., Wang,J., McDermott,J.E., Smith,R.D., Rodland,K.D., Metz,T.O. and Pounds,J.G. (2015) Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J. Proteome Res.*, **14**, 1993–2001.
46. Kim,H., Golub,G.H. and Park,H. (2004) Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, **21**, 187–198.
47. Verboven,S., Branden,K.V. and Goos,P. (2007) Sequential imputation for missing values. *Comput. Biol. Chem.*, **31**, 320–327.
48. Branden,K.V. and Verboven,S. (2009) Robust data imputation. *Comput. Biol. Chem.*, **33**, 7–13.
49. Templ,M., Kowarik,A. and Filzmoser,P. (2011) Iterative stepwise regression imputation using standard and robust methods. *Comput. Stat. Data Anal.*, **55**, 2793–2806.
50. Kokla,M., Virtanen,J., Kolehmainen,M., Paananen,J. and Hanhineva,K. (2019) Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. *BMC Bioinformatics*, **20**, 492.
51. Tyanova,S., Temu,T., Sinitcyn,P., Carlson,A., Hein,M.Y., Geiger,T., Mann,M. and Cox,J. (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods*, **13**, 731–740.
52. Wei,R., Wang,J., Jia,E., Chen,T., Ni,Y. and Jia,W. (2018) GSimp: a Gibbs sampler based left-censored missing value imputation approach for metabolomics studies. *PLoS Comput. Biol.*, **14**, e1005973.
53. Li,Q., Fisher,K., Meng,W., Fang,B., Welsh,E., Haura,E.B., Koomen,J.M., Eschrich,S.A., Fridley,B.L. and Chen,Y.A. (2020) GMSimpute: a generalized two-step Lasso approach to impute missing values in label-free mass spectrum analysis. *Bioinformatics*, **36**, 257–263.
54. Rosenberger,G., Liu,Y., Rost,H.L., Ludwig,C., Buil,A., Bensimon,A., Soste,M., Spector,T.D., Dermitzakis,E.T., Collins,B.C. *et al.* (2017) Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS. *Nat. Biotechnol.*, **35**, 781–788.
55. Mehnert,M., Li,W., Wu,C., Salovska,B. and Liu,Y. (2019) Combining rapid data independent acquisition and CRISPR gene deletion for studying potential protein functions: a case of HMGN1. *Proteomics*, **19**, 1800438.
56. Li,W., Chi,H., Salovska,B., Wu,C., Sun,L., Rosenberger,G. and Liu,Y. (2019) Assessing the relationship between mass window width and retention time scheduling on protein coverage for data-independent acquisition. *J. Am. Soc. Mass. Spectrom.*, **30**, 1396–1405.
57. Bruderer,R., Bernhardt,O.M., Gandhi,T., Miladinovic,S.M., Cheng,L.Y., Messner,S., Ehrenberger,T., Zanotelli,V., Butscheid,Y., Escher,C. *et al.* (2015) Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteomics*, **14**, 1400–1410.
58. Bruderer,R., Bernhardt,O.M., Gandhi,T., Xuan,Y., Sondermann,J., Schmidt,M., Gomez-Varela,D. and Reiter,L. (2017) Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Mol. Cell. Proteomics*, **16**, 2296–2309.
59. Bekker-Jensen,D.B., Bernhardt,O.M., Hogrebe,A., Martinez-Val,A., Verbeke,L., Gandhi,T., Kelstrup,C.D., Reiter,L. and Olsen,J.V. (2020) Rapid and site-specific deep phosphoproteome profiling by data-independent acquisition without the need for spectral libraries. *Nat. Commun.*, **11**, 787.
60. Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**, 1367–1372.
61. Olsen,J.V., Blagoev,B., Gnad,F., Macek,B., Kumar,C., Mortensen,P. and Mann,M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, **127**, 635–648.
62. Tan,S.L.W., Chadha,S., Liu,Y., Gabasova,E., Perera,D., Ahmed,K., Constantinou,S., Renaudin,X., Lee,M., Aebersold,R. *et al.* (2017) A class of environmental and endogenous toxins induces BRCA2 haploinsufficiency and genome instability. *Cell*, **169**, 1105–1118.
63. Rosenberger,G., Koh,C.C., Guo,T., Rost,H.L., Kouvonen,P., Collins,B.C., Heusel,M., Liu,Y., Caron,E., Vichalkovski,A. *et al.* (2014) A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data*, **1**, 140031.
64. Rost,H.L., Rosenberger,G., Navarro,P., Gillet,L., Miladinovic,S.M., Schubert,O.T., Wolski,W., Collins,B.C., Malmstrom,J., Malmstrom,L. *et al.* (2014) OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotech*, **32**, 219–223.
65. Peres-Neto,P.R. and Jackson,D.A. (2001) How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia*, **129**, 169–178.

66. Li,S., Arnold,R.J., Tang,H. and Radivojac,P. (2011) On the accuracy and limits of peptide fragmentation spectrum prediction. *Anal. Chem.*, **83**, 790–796.
67. Schwarz,E., Levin,Y., Wang,L., Leweke,F.M. and Bahn,S. (2007) Peptide correlation: a means to identify high quality quantitative information in large-scale proteomic studies. *J. Sep. Sci.*, **30**, 2190–2197.
68. Ruepp,A., Brauner,B., Dunger-Kaltenbach,I., Frishman,G., Montrone,C., Stransky,M., Waegele,B., Schmidt,T., Doudieu,O.N., Stumpflen,V. *et al.* (2008) CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.*, **36**, D646–D650.
69. Drew,K., Lee,C., Huizar,R.L., Tu,F., Borgeson,B., McWhite,C.D., Ma,Y., Wallingford,J.B. and Marcotte,E.M. (2017) Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Mol. Syst. Biol.*, **13**, 932.
70. Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Statist.*, **5**, 299–314.
71. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B. Met.*, **57**, 289–300.
72. Wang,S., Cai,Y., Cheng,J., Li,W., Liu,Y. and Yang,H. (2019) motifeR: an integrated web software for identification and visualization of protein post-translational modification motifs. *Proteomics*, 1900245.
73. 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.* (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.*, **47**, D442–D450.
74. Dephoure,N., Zhou,C., Villen,J., Beausoleil,S.A., Bakalarski,C.E., Elledge,S.J. and Gygi,S.P. (2008) A quantitative atlas of mitotic phosphorylation. *PNAS*, **105**, 10762–10767.