

Supporting Manual for:

NAGuideR: performing and prioritizing missing value imputations

for consistent bottom-up proteomic analyses

Shisheng Wang¹, Wenxue Li², Liqiang Hu¹, Jingqiu Cheng¹, 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, West China Hospital, Sichuan University, Chengdu, 610041, China

² Yale Cancer Biology Institute, Yale University, West Haven, CT, 06516, USA

³ Department of Pharmacology, Yale University School of Medicine, New Haven, CT, 06520, USA

Corresponding Author

*Email address: yanghao@scu.edu.cn; yansheng.liu@yale.edu.

Table 1. Description of 20 missing value imputation methods.

| Class | Abbreviation | Manipulation Method | Remark | Function | Package/References |
|-------|--------------|--|--|--------------|--------------------|
| Fast | zero | zero | Replaces the missing values by 0. | 0 | base (1) |
| | minimum | minimum | Replaces the missing values by the smallest non-missing value in the data. | min | base (2) |
| | colmedian | Column median | Replaces the missing values by the median of non-missing value in each column. | impute | e1071 (3) |
| | rowmedian | Row median | Replaces the missing values by the median of non-missing value in each row. | impute | e1071 (3) |
| | SVD | Singular value decomposition imputation | Initializes all missing elements with zero then estimate them as a linear combination of the k most significant eigen-variables iteratively until reaches certain convergence threshold. | svdPca | pcaMethods (4) |
| | KNN | K Nearest Neighbors imputation | K-nearest neighbors in the space of peptides/proteins to impute missing expression values. | impute.knn | impute (4) |
| | Seq-KNN | Sequential K-nearest neighbor | Imputes the missing values sequentially from the peptide/protein having least missing values based on KNN method, and uses the imputed values for the later imputation. | SeqKNN | SeqKnn (5) |
| | LLS | Local least squares imputation | K variables (peptides/ proteins) are selected by Pearson, spearman or Kendall correlation coefficients. Then missing values are imputed by a linear combination of the k selected variables. The optimal combination is found by LLS regression. | llsImpute | pcaMethods (6) |
| | QR | Quantile regression imputation of left-censored data | A missing data imputation method that performs the imputation of left-censored missing data using random draws from a truncated | impute.QRILC | imputeLCMD (7) |

| | | | | | |
|--|---------|---|---|--------------------------------------|----------------|
| | | | distribution with parameters estimated using quantile regression. | | |
| | MLE | Imputation based on maximum likelihood estimation | Maximum likelihood-based imputation method using the EM algorithm. | prelim.norm, em.norm, imp.norm | norm (8) |
| | Mindet | Deterministic minimum imputation | Perform the imputation of left-censored missing data using a deterministic minimal value approach. Considering an expression data with n samples and p features, for each sample, the missing entries are replaced with a minimal value observed in that sample. The minimal value observed is estimated as being the q-th quantile of the observed values in that sample. | impute.MinDet | imputeLCMD (9) |
| | Minprob | Probabilistic minimum imputation | Performs the imputation of left-censored missing data by random draws from a Gaussian distribution centred to a minimal value. Considering an expression data matrix with n samples and p features, for each sample, the mean value of the Gaussian distribution is set to a minimal observed value in that sample. The minimal value observed is estimated as being the q-th quantile of the observed values in that sample. The standard deviation is estimated as the median of the feature standard deviations. | impute.MinProb | imputeLCMD (9) |
| | Impseq | Sequential imputation of missing values | Estimates sequentially the missing values in an incomplete observation by minimizing the determinant of the covariance of the augmented data matrix. Then the observation is added to the complete data matrix and | impSeq | rrcovNA (10) |

| | | | | | |
|------|-------------|---|--|------------------------|--------------------|
| | | | the algorithm continues with the next observation with missing values. | | |
| | Impseqrob | Robust sequential imputation of missing values | Similar to Impseq, but improved by plugging in robust estimators of location and scatter. | impSeqRob | rrcovNA (11) |
| | Mice-normal | Multivariate Imputation by Chained Equations-Bayesian linear regression | Generates multiple imputations for incomplete multivariate data by Gibbs sampling. Missing data can occur anywhere in the data. The algorithm imputes an incomplete column (the target column) by generating 'plausible' synthetic values given other columns in the data. Each incomplete column must act as a target column, and has its own specific set of predictors. The default set of predictors for a given target consists of all other columns in the data. For predictors that are incomplete themselves, the most recently generated imputations are used to complete the predictors prior to imputation of the target column. The imputation method depends on Bayesian linear regression. | mice (method='normal') | mice (12) |
| Slow | BPCA | Bayesian PCA missing value estimation | An iterative method using a Bayesian model to handle missing values. | bpca | pcaMethods (13) |
| | trKNN | Truncation k-nearest neighbors imputation | Applies a Newton-Raphson (NR) optimization to estimate the truncated mean and standard deviation. Then, Pearson correlation was calculated based on standardized data followed by correlation-based kNN imputation. | sim_trKNN_wrapper | Imput_funcs.R (14) |
| | IRM | Iterative | In each step of the iteration, one | irmi | VIM (15) |

| | | | | | |
|--|-----------|--|---|-------------------------|-----------------|
| | | robust model-based imputation | variable is used as a response variable and the remaining variables serve as the regressors. | | |
| | Mice-cart | Multivariate Imputation by Chained Equations-classification and regression trees | Generates multiple imputations for incomplete multivariate data by Gibbs sampling. Missing data can occur anywhere in the data. The algorithm imputes an incomplete column (the target column) by generating 'plausible' synthetic values given other columns in the data. Each incomplete column must act as a target column, and has its own specific set of predictors. The default set of predictors for a given target consists of all other columns in the data. For predictors that are incomplete themselves, the most recently generated imputations are used to complete the predictors prior to imputation of the target column. The imputation method depends on classification and regression trees. | mice (method='cart') | mice (12) |
| | RF | Random forest | Imputes missing values particularly in the case of mixed-type data based on a random forest. It can be used to impute continuous and/or categorical data including complex interactions and nonlinear relations. It yields an out-of-bag (OOB) imputation error estimate. | missForest | missForest (16) |

Supplementary notes

NAGuideR integrates up to 20 common missing value imputation methods (described in Table S1) and provides two categories of evaluation criteria (four classic computational criteria and four common knowledge-based proteomics criteria) to assess the imputation performance of various methods. Here we present the detailed introduction and operation of NAGuideR, users can follow this manuscript to analyze their own data freely and conveniently.

Users can visit this site: <http://www.omicsolution.org/wukong/NAGuideR>. Then the website homepage can be shown like this:



Basically, there are four main steps in NAGuideR:

1. Uploading proteomics expression data and sample information data;
2. Data quality control;
3. Missing value imputation;
4. Performance evaluation;

After this, NAGuideR can provide valuable guidance for users to select one proper method for their own data based on the evaluation results. Detailed introduction can be found in the **Help** part.

Finally, NAGuideR is developed by R shiny (Version 1.3.2), and is free and open to all users with no login requirement. It can be readily accessed by all popular web browsers including Google Chrome, Mozilla Firefox, Safari and Internet Explorer 10 (or later), and so on. We would highly appreciate that if you could send your feedback about any bug or feature request to Shisheng Wang at wssdandan2009@outlook.com.

^_^ Enjoy yourself in NAGuideR ^_^

1. Data Preparation

NAGuideR supports four basic file formats (.csv, .txt, .xlsx, .xls). Before analysis, users should prepare two required data: (1) Proteomics expression data and (2) Sample information data. The data required here could be readily generated based on results of several popular tools such as MaxQuant, PEAKS, Spectronaut, and so on. Then can upload the two data into NAGuideR with right formats respectively and start subsequent analysis.

1.1 Expression data

There are four types of proteomics expression data supported in NAGuideR, among which the main differences are the first few columns.

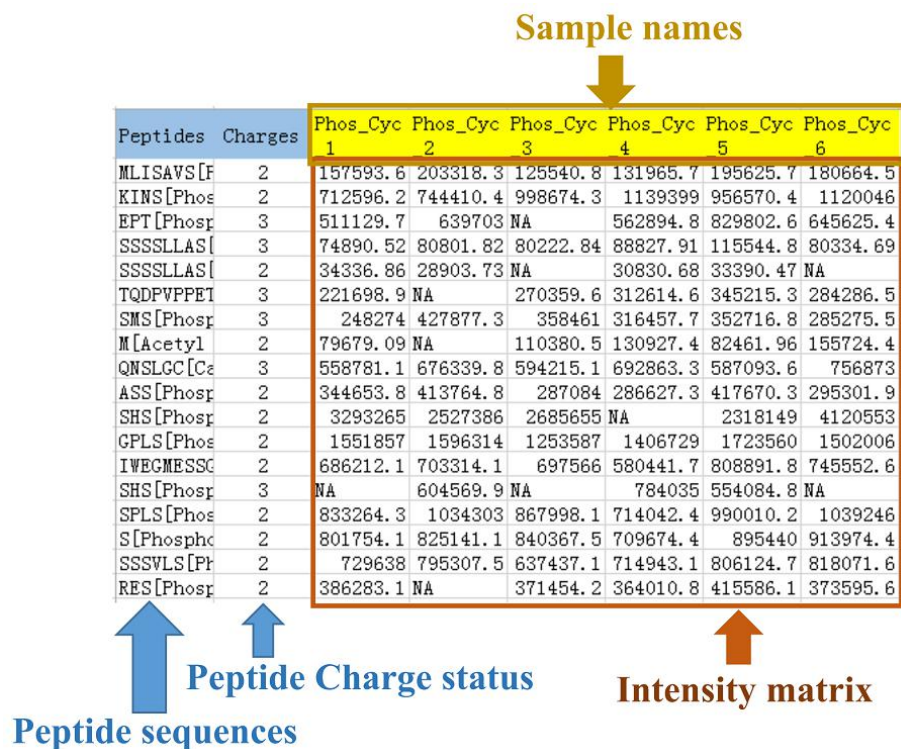
1.1.1 Expression data with peptide sequences, peptide charge status, and protein ids

In this situation, peptide sequences, peptide charge status, and protein ids are sequentially provided in the first three columns of input file. Peptide sequences in the first column can be peptides with post-translational modification (PTM) or stripped peptides (without PTM). The second column is peptide charge status. The protein ids in the third column should be UniProt ids. From the fourth column on, they are peptides/proteins expression intensity in every sample. The data structure is shown as below:

| Peptides | Charges | Uniprot IDs | Phos_Cyc 1 | Phos_Cyc 2 | Phos_Cyc 3 | Phos_Cyc 4 | Phos_Cyc 5 | Phos_Cyc 6 |
|-------------------|---------|-------------|------------|------------|------------|------------|------------|------------|
| MLISAVS[Phospho (| 2 | A0AVK6 | 157593.6 | 203318.3 | 125540.8 | 131965.7 | 195625.7 | 180664.5 |
| KINS[Phospho (STY | 2 | A0AVK6 | 712596.2 | 744410.4 | 998674.3 | 1139399 | 956570.4 | 1120046 |
| EPT[Phospho (STY) | 3 | A0FGR8 | 511129.7 | 639703 | NA | 562894.8 | 829802.6 | 645625.4 |
| SSSSLLAS[Phospho | 3 | A0FGR8 | 74890.52 | 80801.82 | 80222.84 | 88827.91 | 115544.8 | 80334.69 |
| SSSSLLAS[Phospho | 2 | A0FGR8 | 34336.86 | 28903.73 | NA | 30830.68 | 33390.47 | NA |
| TQDPVPPETPSDS[Pho | 3 | A0JLT2 | 221698.9 | NA | 270359.6 | 312614.6 | 345215.3 | 284286.5 |
| SMS[Phospho (STY) | 3 | A0JNW5 | 248274 | 427877.3 | 358461 | 316457.7 | 352716.8 | 285275.5 |
| M[Acetyl (Protein | 2 | A1KXE4 | 79679.09 | NA | 110380.5 | 130927.4 | 82461.96 | 155724.4 |
| QNSLGC[Carbamidom | 3 | A1L020 | 558781.1 | 676339.8 | 594215.1 | 692863.3 | 587093.6 | 756873 |
| ASS[Phospho (STY) | 2 | A1L170 | 344653.8 | 413764.8 | 287084 | 286627.3 | 417670.3 | 295301.9 |
| SHS[Phospho (STY) | 2 | A1L390 | 3293265 | 2527386 | 2685655 | NA | 2318149 | 4120553 |
| GPLS[Phospho (STY | 2 | A1L390 | 1551857 | 1596314 | 1253587 | 1406729 | 1723560 | 1502006 |
| IWEGMESSGS[Phosp | 2 | A1L390 | 686212.1 | 703314.1 | 697566 | 580441.7 | 808891.8 | 745552.6 |
| SHS[Phospho (STY) | 3 | A1L390 | NA | 604569.9 | NA | 784035 | 554084.8 | NA |
| SPLS[Phospho (STY | 2 | A1L390 | 833264.3 | 1034303 | 867998.1 | 714042.4 | 990010.2 | 1039246 |
| S[Phospho (STY)]P | 2 | A1L390 | 801754.1 | 825141.1 | 840367.5 | 709674.4 | 895440 | 913974.4 |
| SSSVLS[Phospho (S | 2 | A1L390 | 729638 | 795307.5 | 637437.1 | 714943.1 | 806124.7 | 818071.6 |
| RES[Phospho (STY) | 2 | A1L390 | 386283.1 | NA | 371454.2 | 364010.8 | 415586.1 | 373595.6 |

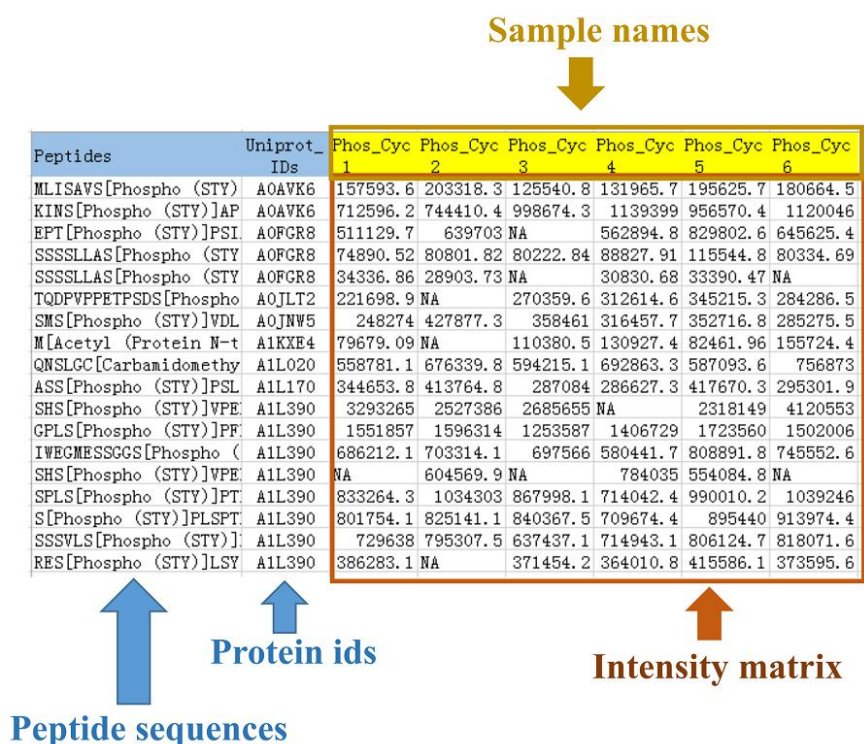
1.1.2 Expression data with peptide sequences and peptide charge status

Similar to the above situation, peptide sequences and peptide charge status are sequentially provided in the first two columns of input file. Peptide sequences in the first column can be peptides with post-translational modification (PTM) or stripped peptides (without PTM). The second column is peptide charge status. From the third column on, they are peptides/proteins expression intensity in every sample. The data structure is shown as below:



1.1.3 Expression data with peptide sequences, and protein ids

Under this circumstance, peptide sequences, and protein ids are sequentially provided in the first two columns of input file. Peptide sequences in the first column can be peptides with post-translational modification (PTM) or stripped peptides (without PTM). The protein ids in the second column should be UniProt ids. From the third column on, they are peptides/proteins expression intensity in every sample. The data structure is shown as below:



1.1.4 Expression data with protein ids

In this situation, protein ids are provided in the first two columns of input file. The protein ids here should be UniProt ids. From the second column on, they are peptides/proteins expression intensity in every sample. The data structure is shown as below:

Sample names

↓

| Uniprot IDs | Phos_Cyc_1 | Phos_Cyc_2 | Phos_Cyc_3 | Phos_Cyc_4 | Phos_Cyc_5 | Phos_Cyc_6 |
|-------------|------------|------------|------------|------------|------------|------------|
| AOAVK6 | 157593.6 | 203318.3 | 125540.8 | 131965.7 | 195625.7 | 180664.5 |
| AOAVK6 | 712596.2 | 744410.4 | 998674.3 | 1139399 | 956570.4 | 1120046 |
| A0FGR8 | 511129.7 | 639703 | NA | 562894.8 | 829802.6 | 645625.4 |
| A0FGR8 | 74890.52 | 80801.82 | 80222.84 | 88827.91 | 115544.8 | 80334.69 |
| A0FGR8 | 34336.86 | 28903.73 | NA | 30830.68 | 33390.47 | NA |
| A0JLT2 | 221698.9 | NA | 270359.6 | 312614.6 | 345215.3 | 284286.5 |
| A0JNW5 | 248274 | 427877.3 | 358461 | 316457.7 | 352716.8 | 285275.5 |
| A1KXE4 | 79679.09 | NA | 110380.5 | 130927.4 | 82461.96 | 155724.4 |
| A1LO20 | 558781.1 | 676339.8 | 594215.1 | 692863.3 | 587093.6 | 756873 |
| A1L170 | 344653.8 | 413764.8 | 287084 | 286627.3 | 417670.3 | 295301.9 |
| A1L390 | 3293265 | 2527386 | 2685655 | NA | 2318149 | 4120553 |
| A1L390 | 1551857 | 1596314 | 1253587 | 1406729 | 1723560 | 1502006 |
| A1L390 | 686212.1 | 703314.1 | 697566 | 580441.7 | 808891.8 | 745552.6 |
| A1L390 | NA | 604569.9 | NA | 784035 | 554084.8 | NA |
| A1L390 | 833264.3 | 1034303 | 867998.1 | 714042.4 | 990010.2 | 1039246 |
| A1L390 | 801754.1 | 825141.1 | 840367.5 | 709674.4 | 895440 | 913974.4 |
| A1L390 | 729638 | 795307.5 | 637437.1 | 714943.1 | 806124.7 | 818071.6 |
| A1L390 | 386283.1 | NA | 371454.2 | 364010.8 | 415586.1 | 373595.6 |

↑ ↑

Protein ids Intensity matrix

1.2 Samples information data

Sample information here means users should identify sample group information. The sample names are in the first column and their orders are same as those in the expression data. Group information is in the second column. The data structure is shown as below:

Sample names

↓

| Samples | Groups |
|--------------|--------|
| Phos_Cyc_1 | Cyc |
| Phos_Cyc_2 | Cyc |
| Phos_Cyc_3 | Cyc |
| Phos_Cyc_4 | Cyc |
| Phos_Cyc_5 | Cyc |
| Phos_Cyc_6 | Cyc |
| Phos_Cyc_7 | Cyc |
| Phos_Cyc_8 | Cyc |
| Phos_Cyc_9 | Cyc |
| Phos_Cyc_10 | Cyc |
| Phos_Noco_1 | Noco |
| Phos_Noco_2 | Noco |
| Phos_Noco_3 | Noco |
| Phos_Noco_4 | Noco |
| Phos_Noco_5 | Noco |
| Phos_Noco_6 | Noco |
| Phos_Noco_7 | Noco |
| Phos_Noco_8 | Noco |
| Phos_Noco_9 | Noco |
| Phos_Noco_10 | Noco |

↑

Sample groups

1.3 Download example datasets

If users want to download the example datasets to their own computer and check the data format locally, they can download them from here:

Step 1: Upload Original Data ?

1. Expression data :

The first few column types:

Peptides+Charges+Proteins

Download example expression data

Download example sample group data

| Peptides | Charges | Uniprot_IDs | Phos_Cyc_1 | Phos_Cyc_2 | Phos_Cyc_3 | Phos_Cyc_4 |
|--|---------|-------------|-------------|-------------|-------------|-------------|
| 1 MLISAVS[Phospho (STY)]PEIR | 2 | A0AVK6 | 157593.625 | 203318.2969 | 125540.7891 | 131965.7031 |
| 2 (STY)APS[Phospho (STY)]PS[Phospho (STY)]PK | 2 | A0AVK6 | 712596.1675 | 744410.375 | 998674.3125 | 1139399.125 |
| 3 EPT[Phospho (STY)]PSIASDISLPATQELR | 3 | A0FGR8 | 511129.6875 | 639703 | | 562894.8125 |
| 4 SSSSLAS[Phospho (STY)]PGHISVK | 3 | A0FGR8 | 74890.52344 | 80801.82031 | 80222.84375 | 88827.91406 |
| 5 SSSSLAS[Phospho (STY)]PGHISVK | 2 | A0FGR8 | 34336.85938 | 28903.73438 | | 30830.67578 |
| 6 TQDPVPPETPSDS[Phospho (STY)]DHK | 3 | A0JLT2 | 221698.9063 | | 270359.5625 | 312614.5938 |
| 7 SMS[Phospho (STY)]VDLSHIPKDLLFK | 3 | A0JNW5 | 248274.0156 | 427877.25 | 358460.9688 | 316457.7188 |

First, select “Load example data” and the example data will be shown on the right panel interactively. Users can visually observe what the data looks like.

Second, users can download the example data (expression data and sample information data) by clicking the corresponding button. The data are save as .csv format and users can open them in other software, such as Excel.

2. Import data.

This is the first step, users should upload data here or load the example data to learn the data formats. By default, we use the example data to show each result of every step.

2.1 Uploading data. When users prepare their data (expression and sample information data set), they can upload these data from here:

The screenshot displays the NAGuideR web application interface. At the top, there is a navigation bar with the NAGuideR logo and links for Welcome, Import Data, NA Overview, Methods, Results and Assessments, and Help. The main content area is divided into two panels. The left panel, titled '1. Parameters panel', contains 'Step 1: Upload Original Data'. It has two sections: '1. Expression data' and '2. Samples information data'. Each section allows users to load experimental or example data, choose a file format (csv, xls, xlsx), and specify import parameters like 'First row as column names?' and 'First column as row names?'. The right panel, titled '2. Results panel', shows the results of the data upload. It contains two sections: '1. Expression data' and '2. Samples information data'. Both sections display a message: 'NAGuideR detects that you do not upload your data. Please upload the expression data, or load the example data to check first.' The results are presented in a table format with a search bar and pagination controls.

There are two main panels: first, *parameters panel*, users can adjust some parameters here; second, *results panel*, many results after users set the parameters will be shown here and users can also download these results.

In the *parameters panel* of “Import Data”, there are two choices for users:

a. Load experimental data. When users choose this option, they can upload their own data from here. Users should select the right format based on their own data and then click “Browse” button to import the data;

First row as column names: this means whether the first row is column names. If true, you should choose this parameter.

First column as row names: this means whether the first column is row names. If true, you should choose this parameter.

b. Load example data. As described in part 1.3, users can choose this option and download the example data to check them locally.

In the *results panel* of “Import Data”, if users don’t upload their data, here will show “NAGuideR detects that you do not upload your data. Please upload the expression data (or sample information data), or load the example data to check first” to warn users.

Before uploading expression data, users should also recognize which type their data belongs to and choose the right parameter by adjusting the “*The first few column types*”. The instruction of the column types can be found above (part 1).

Step 1: Upload Original Data ?

☒ Load experimental data ☐ Load example data

1. Expression data:

1.1 File format:

☒ .csv/txt ☐ .xls ☐ .xlsx

1.2 Import your data :

Browse...

No file selected

1. Expression data :

The first few column types:

Peptides+Charges+Proteins ▲

Peptides+Charges+Proteins

Peptides+Charges

Peptides+Proteins

Proteins

Others

load you

Showing 1 to 1 of 1 entries

3. NA Overview

Users can check the missing value situation of their own data and filter those data with high proportion of missing value in this step. NA is short for Not Available, which means missing value here.

3.1 Parameters

1. *Missing value type*: what the missing values look like in the expression data, for example, Spectronaut (17,18) software usually export “Filtered” as missing values, so users should change this parameter to “Filtered” if their data contain “Filtered”. NAGuideR will recognize these characters and replace them with NAs.

2. *Count NA by each group or not*: if true, NAGuideR will count the number of missing value by each group and calculate the NA ratio, otherwise, calculate the NA ratio across all groups, for example, as below:

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W |
|-------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Peptides | | | | | | | | | | | | | | | | | | | | | | | |
| Charges | | | | | | | | | | | | | | | | | | | | | | | |
| Uniprot IDs | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |
| NA | | | | | | | | | | | | | | | | | | | | | | | |

There are 2 groups (10 biological replicates in each group) here, if users select this parameter, NAGuideR will calculate 2 NA ratios for this peptide (first group: $1/10=0.1$, second group: $5/10=0.5$), otherwise, only one NA ratio: $6/20=0.3$.

3. *NA ratio*: the threshold of NA ratio. Those peptides/proteins with NA ratio above this threshold will be removed.
4. *Median normalization or not*: if true, NAGuideR will process median normalization for original data.
5. *Log or not*: if true, the data will be logarithmic with base 2.
6. *CV threshold (raw scale)*: the threshold of coefficient of variation. Those peptides/proteins with NA ratio above this threshold will be removed. “raw scale” here means the CV of each peptide/protein is calculate using the data before logarithm transformation.
7. *Height for figure*: users can adjust the height of figures by changing this parameter.

If users set these parameters well, then click “calculate” button, the results will appear on the right panel.

Step 2: NA Overview

1. Missing value type:
NA

2. Count NA by each group or not?
☒

3. NA ratio:
0.5

4. Median normalization or not?
☒

5. Log or not?
☐

6. CV threshold (raw scale):
0.3

Height for figure:
900

NA Distribution

NA Filter

NA data

Plot by column

Plot by row

Calculate

Download

Show 20 entries

Search:

| | Phos_Cyc_1 | Phos_Cyc_2 | Phos_Cyc_3 | Phos_Cyc_4 | Phos_Cyc_5 | Phos_Cyc_6 | Phos_Cyc_7 | Phos_Cyc_8 | Phos_Cyc_9 |
|---|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| MLISA/SPPhospho (STY)PEIR_2_ADA/K6 | 157593.625 | 203318.2969 | 125540.7891 | 131965.7031 | 195625.6563 | 180664.5469 | 148941.4688 | 143790.9375 | 91102.99219 |
| KINS/Phospho (STY)APSP/Phospho (STY)SP/Phospho (STY)PIK_2_ADA/K6 | 712596.1875 | 744410.375 | 998674.3125 | 1139399.125 | 956570.375 | 1120045.625 | 860231.875 | 823408.5625 | |
| EPT/Phospho (STY)PSIASDISLPATQELR_3_ADFGR8 | 511129.6875 | 639703 | | 562894.8125 | 829602.625 | 645625.4375 | | 608932.875 | |
| SSSSLAS/Phospho (STY)PGHISVK_3_ADFGR8 | 74890.52344 | 80801.82031 | 80222.84375 | 88827.91406 | 115544.7813 | 80334.6875 | 80562.07031 | 61538.41406 | 53648.84766 |
| SSSSLAS/Phospho (STY)PGHISVK_2_ADFGR8 | 34336.85938 | 28903.73438 | | 30830.67578 | 33390.47266 | | 31978.69141 | 29228.26758 | |
| TQDPVPETPSDS/Phospho (STY)DHK_3_ADLT2 | 221698.9063 | | 270359.5625 | 312614.5938 | 345215.25 | 284286.4688 | 203317.4063 | 218004.125 | 185125.5156 |
| SMS/Phospho (STY)VDLSHPLKDFLLFK_3_AQJNW5 | 248274.0156 | 427877.25 | 358460.9688 | 316457.7188 | 352716.75 | 285275.5 | 331924.5625 | 174794.2344 | 241767.2344 |
| M[Acetyl] (Protein N-term)NPVYSPGSSGVPPY/Phospho (STY)ANAK_2_A1K0E4 | 79679.09375 | | 110380.5 | 130927.3672 | 82461.96094 | 155724.3594 | 113495.2891 | 136404.2969 | 56171.30859 |

3.2 results

a. *NA Distribution*. This part contains three sub-parts:

a.1 *NA data*. Here shows the result where the “Missing value type” will be replaced with NA and users can click “Download” button to download this result to their own computer:

Download

Show 20 entries

Search:

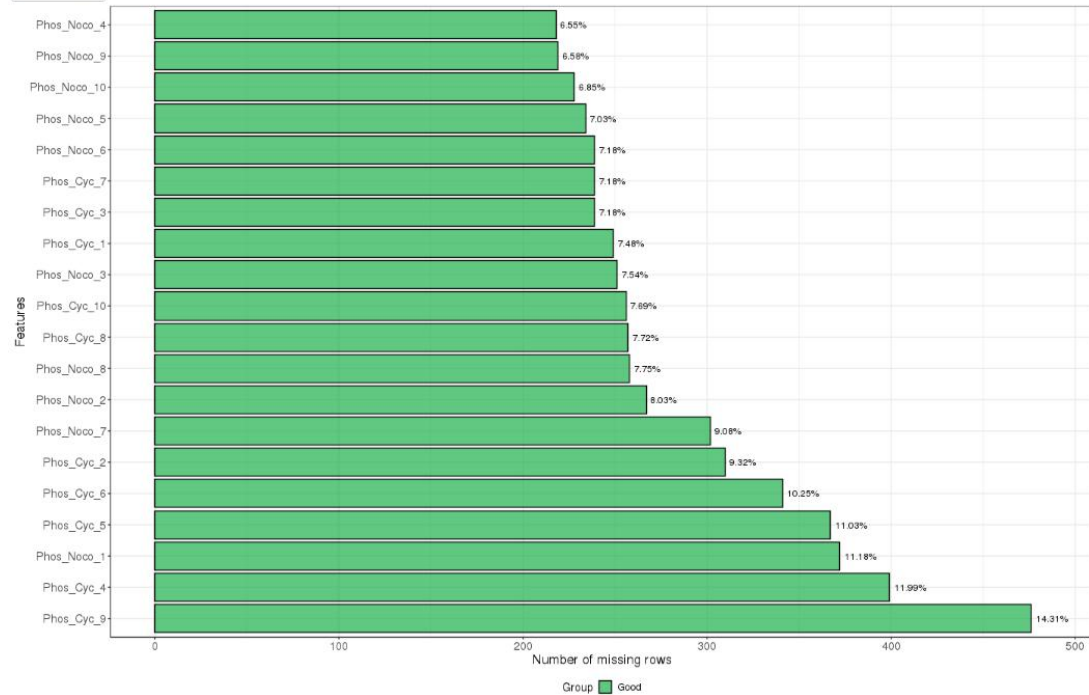
| | Phos_Cyc_1 | Phos_Cyc_2 | Phos_Cyc_3 | Phos_Cyc_4 | Phos_Cyc_5 | Phos_Cyc_6 | Phos_Cyc_7 | Phos_Cyc_8 | Phos_Cyc_9 | Phos_Cyc_10 | Phos_Noco_1 | Phos_Noco_2 |
|--|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| MLISA/SPPhospho (STY)PEIR_2_ADA/K6 | 157593.625 | 203318.2969 | 125540.7891 | 131965.7031 | 195625.6563 | 180664.5469 | 148941.4688 | 143790.9375 | 91102.99219 | 140345.125 | 59488.09766 | 92400.46094 |
| KINS/Phospho (STY)APSP/Phospho (STY)SP/Phospho (STY)PIK_2_ADA/K6 | 712596.1875 | 744410.375 | 998674.3125 | 1139399.125 | 956570.375 | 1120045.625 | 860231.875 | 823408.5625 | | 888177.75 | 509135.625 | 595305.5 |
| EPT/Phospho (STY)PSIASDISLPATQELR_3_ADFGR8 | 511129.6875 | 639703 | | 562894.8125 | 829602.625 | 645625.4375 | | 608932.875 | | 620510.4375 | 323346.5313 | 334969.9063 |
| SSSSLAS/Phospho (STY)PGHISVK_3_ADFGR8 | 74890.52344 | 80801.82031 | 80222.84375 | 88827.91406 | 115544.7813 | 80334.6875 | 80562.07031 | 61538.41406 | 53648.84766 | 65030.57031 | 516738.8125 | 782993.875 |
| SSSSLAS/Phospho (STY)PGHISVK_2_ADFGR8 | 34336.85938 | 28903.73438 | | 30830.67578 | 33390.47266 | | 31978.69141 | 29228.26758 | | 26532.99219 | 333476.9688 | 297875.2188 |
| TQDPVPETPSDS/Phospho (STY)DHK_3_ADLT2 | 221698.9063 | | 270359.5625 | 312614.5938 | 345215.25 | 284286.4688 | 203317.4063 | 218004.125 | 185125.5156 | 245305 | 81982.05469 | 81776.67188 |
| SMS/Phospho (STY)VDLSHPLKDFLLFK_3_AQJNW5 | 248274.0156 | 427877.25 | 358460.9688 | 316457.7188 | 352716.75 | 285275.5 | 331924.5625 | 174794.2344 | 241767.2344 | 284069.4375 | 170259.6094 | 207056.5 |
| M[Acetyl] (Protein N-term)NPVYSPGSSGVPPY/Phospho (STY)ANAK_2_A1K0E4 | 79679.09375 | | 110380.5 | 130927.3672 | 82461.96094 | 155724.3594 | 113495.2891 | 136404.2969 | 56171.30859 | 98299.69531 | | |
| QNSLGC[Carbamidomethyl] (C)IGEC[Carbamidomethyl] (C)IDG/Phospho (STY)GFEAPR_3_A1L020 | 558781.125 | 676339.75 | 594215.0625 | 692863.25 | 587093.5625 | 756873 | 569292.25 | 648059.3125 | | 626625.4375 | 379149.5625 | 361978.75 |

a.2 *Plot by column*. Here shows the result of the NA distribution of every sample.

☐ NA data
 ☒ Plot by column
 ☐ Plot by row

Calculate

Download

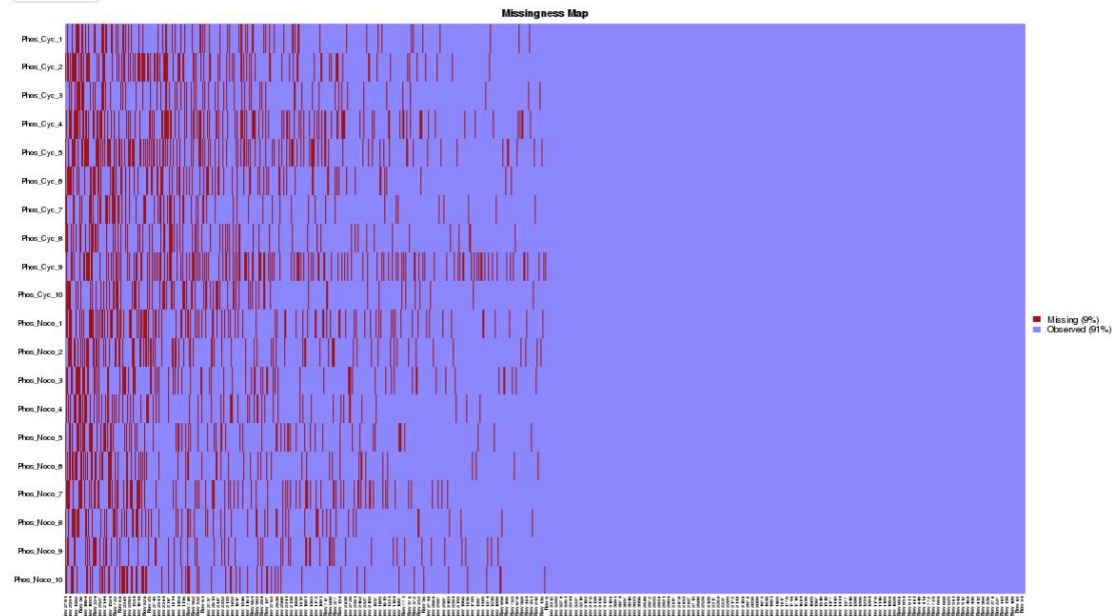


a.2 Plot by row. Here shows the result of the NA distribution of every peptide/protein.

☐ NA data
 ☐ Plot by column
 ☒ Plot by row

Calculate

Download



b. NA filter. This part will show the filtered result.

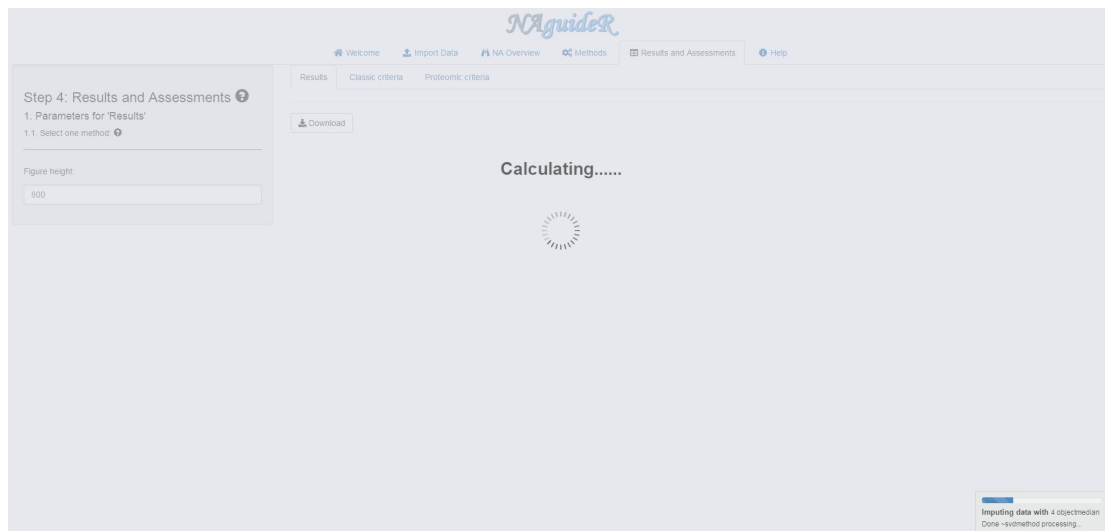
4. Methods

In this step, users can choose or cancel any missing value imputation method. With regard to the running time, we set these fast methods (left part, 15 methods) chosen by default. If users choose those slow methods (right part, 5 methods), that means the running time will be longer. By default, the fast methods are selected. If users want to try these slow methods, they just select the corresponding methods. The detailed information about each method can be found in Table S2. In addition, we also provide the reference for every method just blow each option on the web:

After selecting suitable methods, users need to click 'Calculate' button, and a popup window will be jumped out to show the selected methods, then click 'OK' button and continue:

5. Results and Assessments

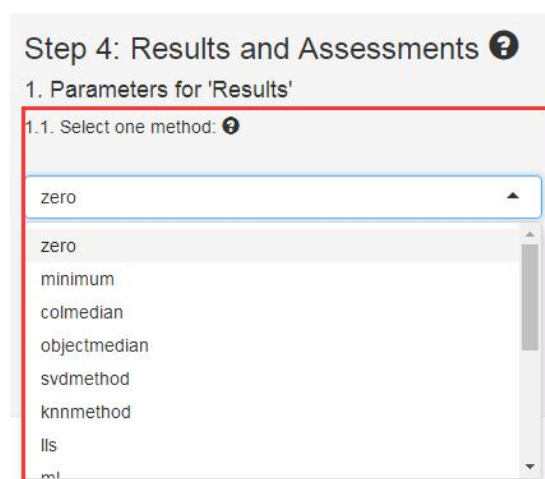
This step will process missing value imputation and performance evaluation of every method that users select in “Methods” step. Click “Results and Assessments”, NAGuideR will start to impute these missing value, a process bar will appear in the bottom right corner to tell users where it goes:



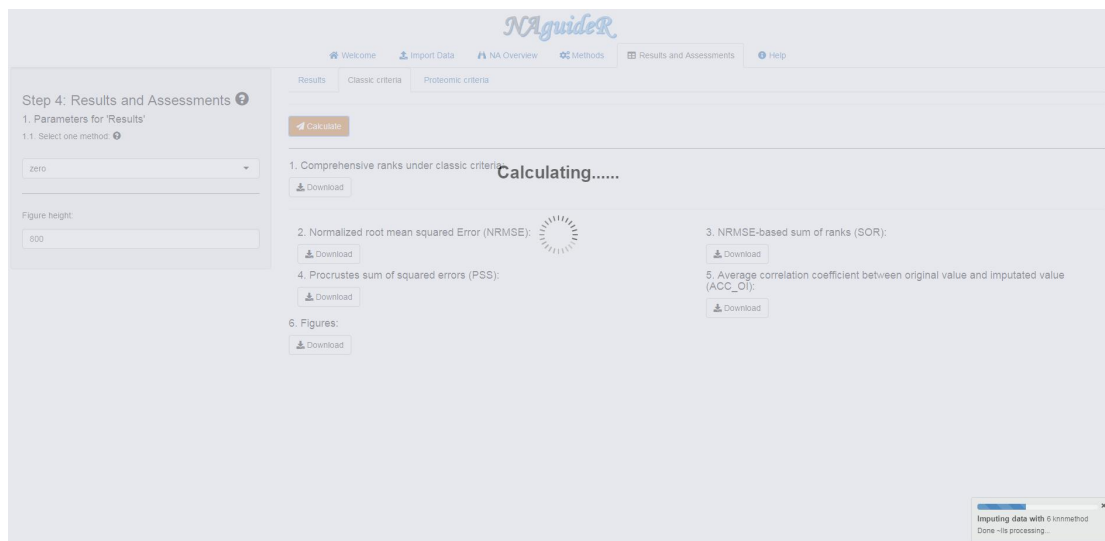
The result from every imputation method will be shown on the “Results” panel:

| | Phos_Cyc_1 | Phos_Cyc_2 | Phos_Cyc_3 | Phos_Cyc_4 | Phos_Cyc_5 | Phos_Cyc_6 | Phos_Cyc_7 |
|---|------------|------------|------------|------------|------------|------------|------------|
| MLISAVSIPhospho (STY)PEIR_2_ADAVK6 | -0.8907 | -0.60469 | -1.19381 | -1.29635 | -0.71933 | -0.88619 | -0.951 |
| KINSIPhospho (STY)APSPHospho (STY)PIK_2_ADAVK6 | 1.28617 | 1.26767 | 1.79805 | 1.81369 | 1.57044 | 1.74598 | 1.57898 |
| EPTIPhospho (STY)PSIASDISLPATQELR_3_ADFGR8 | 0.80678 | 1.04897 | 0 | 0.79635 | 1.36534 | 0.95119 | 0 |
| SSSSLASIPhospho (STY)PGHISVK_3_ADFGR8 | -1.96406 | -1.93597 | -1.83988 | -1.86743 | -1.47898 | -2.05541 | -1.83757 |
| SSSSLASIPhospho (STY)PGHISVK_2_ADFGR8 | -3.08908 | -3.4191 | 0 | -3.39407 | -3.26992 | 0 | -3.17056 |
| TQDPVPPETPSDSIPhospho (STY)DHK_3_ADJLT2 | -0.39831 | 0 | -0.08709 | -0.05213 | 0.10007 | -0.23216 | -0.50201 |
| SMSIPhospho (STY)VDLSHPLKDPFLFK_3_AQJNW5 | -0.23498 | 0.46877 | 0.31985 | -0.0345 | 0.13108 | -0.22715 | 0.20511 |
| M(Acetyl (Protein N-term))NPVYSPGSSGVPIPhospho (STY)ANAK_2_AIKXE4 | -1.87464 | 0 | -1.37948 | -1.30774 | -1.96563 | -1.10051 | -1.34311 |

Users can change the parameter “Select one method” on the left panel to check relative result, for example, if users select “zero”, it will show the result derived from zero method:



Next, click “Classic criteria” and “Calculate” button. NAGuideR will assess every method under the four classic criteria:



The tables and figures are provided here under the four classic criteria.

1. This table shows the comprehensive ranks of every imputation method;
- 2-5, the tables show the scores of every imputation method based on 'Normalized root mean squared Error (NRMSE)', 'NRMSE-based sum of ranks (SOR)', 'Procrustes sum of squared errors (PSS)', and 'Average correlation coefficient between original value and imputed value (ACC_OI)', respectively;
6. Figures here show the normalized scores of every imputation method under the four classic criteria. 'Normalized Values' here means every score divides by corresponding max value.

1. Comprehensive ranks under classic criteria:

Download

Show 20 entries

Search:

| Methods | NRMSE_Rank | SOR_Rank | ACC_OI_Rank | PSS_Rank | Rank_Mean |
|-----------|--------------|----------|-------------|----------|-----------|
| Method 2 | impseq | 1 | 1 | 1 | 1 |
| Method 3 | impseqrob | 2 | 2 | 2 | 2 |
| Method 13 | seqknn | 4 | 3 | 6 | 4 |
| Method 10 | ml | 3 | 6 | 3 | 4.25 |
| Method 4 | knnmethod | 5 | 4 | 5 | 4.5 |
| Method 5 | lts | 6 | 5 | 4 | 5.25 |
| Method 6 | mice-norm | 7 | 7 | 7 | 7 |
| Method 11 | objectmedian | 8 | 8 | 8 | 8 |
| Method 12 | qrlc | 9 | 10 | 11 | 9.75 |
| Method 14 | svdmethod | 10 | 9 | 10 | 10.25 |
| Method 1 | colmedian | 11 | 12 | 11 | 11 |
| Method 15 | zero | 12 | 11 | 9 | 11 |
| Method 7 | mindet | 13 | 13 | 13 | 13 |
| Method 9 | minprob | 14 | 14 | 14 | 14 |
| Method 8 | minimum | 15 | 15 | 15 | 15 |

Showing 1 to 15 of 15 entries

Previous1Next

2. Normalized root mean squared Error (NRMSE):

Download

Show 20 entries

Search:

| | Methods | | NRMSE |
|-----------|--------------|--|---------|
| Method 11 | impseq | | 0.07796 |
| Method 12 | impseqrob | | 0.07814 |
| Method 8 | ml | | 0.10625 |
| Method 15 | seqknn | | 0.11049 |
| Method 6 | knnmethod | | 0.11513 |
| Method 7 | lts | | 0.1237 |
| Method 13 | mice-norm | | 0.16857 |
| Method 4 | objectmedian | | 0.5063 |
| Method 14 | qrlc | | 0.8632 |
| Method 5 | svdmethod | | 0.93162 |
| Method 3 | colmedian | | 1.00393 |
| Method 1 | zero | | 1.08355 |
| Method 10 | mindet | | 2.2209 |
| Method 9 | minprob | | 2.25375 |
| Method 2 | minimum | | 3.28021 |

Showing 1 to 15 of 15 entries

Previous

1

Next

3. NRMSE-based sum of ranks (SOR):

Download

Show 20 entries

Search:

| Methods | SOR | |
|-----------|--------------|-------|
| Method 11 | impseq | 2122 |
| Method 12 | impseqrob | 2142 |
| Method 15 | seqknn | 3536 |
| Method 6 | knnmethod | 3625 |
| Method 7 | lts | 3676 |
| Method 8 | ml | 3696 |
| Method 13 | mice-norm | 4296 |
| Method 4 | objectmedian | 6026 |
| Method 5 | svdmethod | 8030 |
| Method 14 | qrlc | 8110 |
| Method 1 | zero | 8406 |
| Method 3 | colmedian | 8418 |
| Method 10 | mindet | 10135 |
| Method 9 | minprob | 10313 |
| Method 2 | minimum | 11769 |

Showing 1 to 15 of 15 entries

Previous1Next

4. Procrustes sum of squared errors (PSS):

Download

Show 20 entries

Search:

| | Methods | PSS |
|-----------|--------------|---------|
| Method 11 | impseq | 0.00048 |
| Method 12 | impseqrob | 0.00051 |
| Method 8 | ml | 0.00064 |
| Method 7 | lts | 0.00094 |
| Method 6 | knnmethod | 0.00109 |
| Method 15 | seqknn | 0.00129 |
| Method 13 | mice-norm | 0.00556 |
| Method 4 | objectmedian | 0.02591 |
| Method 1 | zero | 0.05313 |
| Method 3 | colmedian | 0.05369 |
| Method 14 | qrlc | 0.05468 |
| Method 5 | svdmethod | 0.06779 |
| Method 10 | mindet | 0.10707 |
| Method 9 | minprob | 0.10904 |
| Method 2 | minimum | 0.13141 |

Showing 1 to 15 of 15 entries

Previous

1

Next

5. Average correlation coefficient between original value and imputed value (ACC_OI):

Download

Show 20 ▾ entries

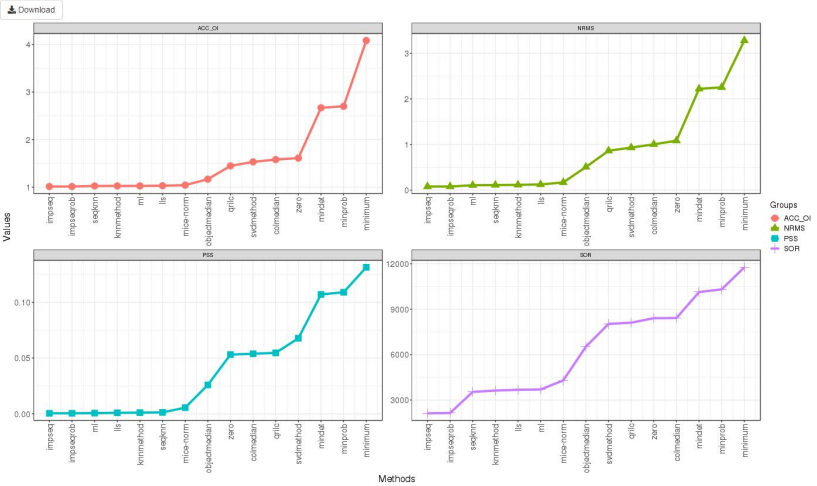
Search:

| | Methods | Cor_mean |
|-----------|--------------|----------|
| Method 11 | impseq | 0.98755 |
| Method 12 | impseqrob | 0.98748 |
| Method 15 | seqknn | 0.9757 |
| Method 6 | knnmethod | 0.975 |
| Method 8 | ml | 0.97447 |
| Method 7 | lts | 0.97116 |
| Method 13 | mice-norm | 0.95947 |
| Method 4 | objectmedian | 0.8567 |
| Method 14 | qrlc | 0.69105 |
| Method 5 | svdmethod | 0.653 |
| Method 3 | colmedian | 0.63258 |
| Method 1 | zero | 0.62062 |
| Method 10 | mindet | 0.37454 |
| Method 9 | minprob | 0.37038 |
| Method 2 | minimum | 0.24487 |

Showing 1 to 15 of 15 entries

Previous1Next

6. Figures:



Then click “Proteomic criteria” and “Calculate” button. NAguideR will assess every method under the four proteomic criteria:



The tables and figures are provided here under the four proteomic criteria.

1. This table shows the comprehensive ranks of every imputation method;
- 2-5, the tables show the scores of every imputation method based on 'Average correlation coefficient between peptides with different charges (ACC_Charge)', 'Average correlation coefficient between peptides in a same protein (ACC_PepProt)', 'Average correlation coefficient between protein complexes (ACC_CORUM)', 'Average correlation coefficient between protein complexes (ACC_PPI)', respectively;
6. Figures here show the correlation coefficient distribution of the original values and the imputed values from every imputation method under the four proteomic criteria.

1. Comprehensive ranks under proteomic criteria:

Download

Show 20 entries

Search:

| | Methods | Charge_Rank | PepProt_Rank | CORUM_Rank | PPI_Rank | Rank_Mean |
|-----------|--------------|-------------|--------------|------------|----------|-----------|
| Method 4 | knmmethod | 2 | 1 | 1 | 2 | 1.5 |
| Method 13 | seqknn | 1 | 2 | 4 | 1 | 2 |
| Method 2 | impseq | 3 | 3 | 2 | 3 | 2.75 |
| Method 3 | impseqrob | 4 | 4 | 3 | 4 | 3.75 |
| Method 5 | its | 5 | 5 | 6 | 5 | 5.25 |
| Method 10 | mi | 6 | 6 | 5 | 6 | 5.75 |
| Method 6 | mice-norm | 7 | 7 | 7 | 7 | 7 |
| Method 11 | objectmedian | 8 | 8 | 8 | 8 | 8 |
| Method 12 | qrlc | 9 | 9 | 9 | 11 | 9.5 |
| Method 14 | svdmethod | 10 | 10 | 10 | 9 | 9.75 |
| Method 1 | colmedian | 11 | 11 | 12 | 10 | 11 |
| Method 15 | zero | 12 | 12 | 11 | 12 | 11.75 |
| Method 7 | mindet | 13 | 13 | 13 | 13 | 13 |
| Method 9 | minprob | 14 | 14 | 14 | 14 | 14 |
| Method 8 | minimum | 15 | 15 | 15 | 15 | 15 |

Showing 1 to 15 of 15 entries

Previous1Next

2. Average correlation coefficient between peptides with different charges (ACC_Charge):

Download

Show 20 entries

Search:

| | Methods | ACC_Charge |
|-----------|--------------|------------|
| Method 15 | seqknn | 0.84803 |
| Method 6 | knmmethod | 0.84666 |
| Method 11 | impseq | 0.84525 |
| Method 12 | impseqrob | 0.84508 |
| Method 7 | its | 0.84018 |
| Method 8 | mi | 0.83723 |
| Method 13 | mice-norm | 0.82966 |
| Method 4 | objectmedian | 0.73897 |
| Method 14 | qrlc | 0.62566 |
| Method 5 | svdmethod | 0.60933 |
| Method 3 | colmedian | 0.59157 |
| Method 1 | zero | 0.58832 |
| Method 10 | mindet | 0.43458 |
| Method 9 | minprob | 0.42645 |
| Method 2 | minimum | 0.35983 |

Showing 1 to 15 of 15 entries

Previous1Next

3. Average correlation coefficient between peptides in a same protein (ACC_PepProt):

Download

Show 20 entries

Search:

| | Methods | ACC_peppro |
|-----------|--------------|------------|
| Method 6 | knmmethod | 0.54688 |
| Method 15 | seqknn | 0.54677 |
| Method 11 | impseq | 0.54602 |
| Method 12 | impseqrob | 0.54588 |
| Method 7 | its | 0.54151 |
| Method 8 | mi | 0.54064 |
| Method 13 | mice-norm | 0.53333 |
| Method 4 | objectmedian | 0.47951 |
| Method 14 | qrlc | 0.40258 |
| Method 5 | svdmethod | 0.38689 |
| Method 3 | colmedian | 0.37715 |
| Method 1 | zero | 0.37693 |
| Method 10 | mindet | 0.27806 |
| Method 9 | minprob | 0.27274 |
| Method 2 | minimum | 0.22728 |

Showing 1 to 15 of 15 entries

Previous1Next

4. Average correlation coefficient between protein complexes (ACC_CORUM):

Download

Show 20 entries

Search:

| | Methods | ACC_CORUM |
|-----------|--------------|-----------|
| Method 6 | knmmethod | 0.30498 |
| Method 11 | impseq | 0.30475 |
| Method 12 | impseqrob | 0.30471 |
| Method 15 | seqknn | 0.30459 |
| Method 8 | mi | 0.29933 |
| Method 7 | its | 0.29666 |
| Method 13 | mice-norm | 0.29583 |
| Method 4 | objectmedian | 0.2485 |
| Method 14 | qrlc | 0.21802 |
| Method 5 | svdmethod | 0.19725 |
| Method 1 | zero | 0.19269 |
| Method 3 | colmedian | 0.18941 |
| Method 10 | mindet | 0.15264 |
| Method 9 | minprob | 0.15054 |
| Method 2 | minimum | 0.127 |

Showing 1 to 15 of 15 entries

Previous1Next

5. Average correlation coefficient between protein complexes (ACC_PPI):

Download

Show 20 entries

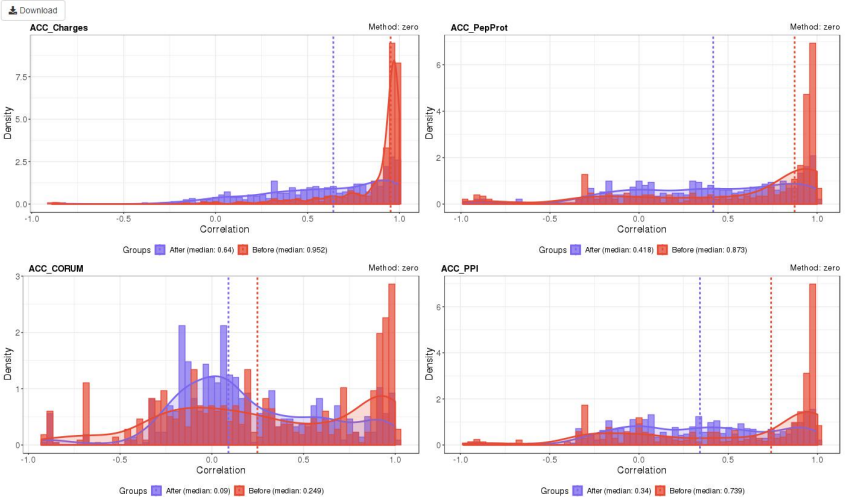
Search:

| | Methods | ACC_PPI |
|-----------|--------------|---------|
| Method 15 | seqknn | 0.48217 |
| Method 6 | knmmethod | 0.48201 |
| Method 11 | impseq | 0.48111 |
| Method 12 | impseqrob | 0.48108 |
| Method 7 | its | 0.4779 |
| Method 8 | mi | 0.47428 |
| Method 13 | mice-norm | 0.46884 |
| Method 4 | objectmedian | 0.41256 |
| Method 5 | svdmethod | 0.35871 |
| Method 3 | colmedian | 0.34504 |
| Method 14 | qrlc | 0.33687 |
| Method 1 | zero | 0.33539 |
| Method 10 | mindet | 0.22936 |
| Method 9 | minprob | 0.22714 |
| Method 2 | minimum | 0.18582 |

Showing 1 to 15 of 15 entries

Previous1Next

6. Figures:



6. Help

This part provides some introduction and operation manual about NAGuideR, so that users can quickly learn what this tool is and how to use this tool.

Detailed description

1. Overview of NAGuideR

2. User manual

1.1 Abstract

Mass-spectrometry (MS) based quantitative proteomics experiments frequently generate data with missing values, which may profoundly affect downstream analyses. A wide variety of missing value 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 web tool, NAGuideR, to enable implementation and evaluation of different missing value methods offered by twenty popular missing-value imputation algorithms. Evaluation of data imputation results can be performed 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 functional protein complexes. We applied NAGuideR into three label-free proteomic datasets featuring peptide-level, protein-level, and phosphoproteomic variables respectively, all generated by data independent 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 web-tool further provides downloadable tables and figures supporting flexible data analysis and interpretation. The flowchart below summarizes the process of data analysis in NOREVA.

A: Original intensity data with missing values (NAs)

| | A1 | A2 | A3 | A4 | A5 | ... | B1 | B2 | B3 | B4 | B5 | ... |
|-------------|-------|-------|-----|-------|-------|-----|-------|-------|-----|-------|-------|-----|
| Feature 1 | I1 | NA | I3 | NA | I5 | ... | I8 | I7 | I9 | I10 | I11 | ... |
| Feature 2 | I1 | I2 | NA | I4 | I6 | ... | I8 | NA | I10 | I11 | I12 | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| Feature n-1 | Ia-10 | Ia-12 | NA | Ia-14 | Ia-16 | ... | Ia-18 | Ia-19 | NA | Ia-21 | Ia-22 | ... |
| Feature n | NA | Ic | Ic | Ic | Ic | ... | Ic | Ic | NA | Ic | Ic | ... |

B: Data quality control

Detailed description

1. Overview of NAGuideR

2. User manual

2.1 Input data preparation

NAGuideR supports four basic file formats (.csv, .txt, .xlsx, .xls). Before analysis, users should prepare two required data: (1) Proteomics expression data and (2) Sample information data. The data required here could be readily generated based on results of several popular tools such as MaxQuant, PEAKS, Spectronaut, and so on. Then can upload the two data into NAGuideR with right formats respectively and start subsequent analysis.

2.1.1 Proteomics expression data

There are four types of proteomics expression data supported in NAGuideR, among which the main differences are the first few columns.

2.1.1.1 Expression data with peptide sequences, peptide charge status, and protein ids

In this situation, peptide sequences, peptide charge status, and protein ids are sequentially provided in the first three columns of input file. Peptide sequences in the first column can be peptides with post-translational modification (PTM) or stripped peptides (without PTM). The second column is peptide charge status. The protein ids in the third column should be UniProt ids. From the fourth column on, they are peptides/proteins expression intensity in every sample. The data structure is shown as below:

Sample names

| Peptides | Charges | Uniprot IDs | Phos_Cyc_1 | Phos_Cyc_2 | Phos_Cyc_3 | Phos_Cyc_4 | Phos_Cyc_5 | Phos_Cyc_6 |
|--------------------|---------|-------------|------------|------------|------------|------------|------------|------------|
| MLISAVS[Phospho (| 2 | A0AVK6 | 1575935.6 | 203318.3 | 125540.8 | 1311955.7 | 1195625.7 | 1190664.0 |
| ELWS[Phospho (STY | 2 | A0AVK6 | 712596.2 | 744410.4 | 598674.3 | 1139399 | 956570.4 | 1120046 |
| EPT[Phospho (STY | 3 | A0PGR8 | 511129.7 | 639703.NA | | 562894.8 | 829802.8 | 645625.4 |
| SSSLLAS[Phospho | 3 | A0PGR8 | 74890.52 | 80801.82 | 80222.84 | 88827.91 | 115544.8 | 80334.60 |
| SSSLLAS[Phospho | 2 | A0PGR8 | 34336.86 | 28903.73 | NA | | 30830.68 | 33390.47 |
| TQSPPTPTSD[Pho | 3 | A0JL12 | 221698.9 | NA | 270359.6 | 312614.6 | 345215.3 | 284286.5 |
| SMS[Phospho (STY | 3 | A0JWF5 | | 248274.4 | 427877.3 | 358461 | 316457.7 | 352716.8 |
| M[acetyl (Protein | 2 | A1KEE4 | 79679.09 | NA | | 110380.5 | 130927.4 | 82461.96 |
| QMSGC[Carbamido | 3 | A1L020 | 558781.1 | 676339.8 | 594215.1 | 692853.3 | 587093.6 | 756873 |
| ASIS[Phospho (STY | 2 | A1L170 | 344653.8 | 413764.8 | 287084 | 286627.3 | 417670.3 | 295301.9 |
| SBS[Phospho (STY | 2 | A1L390 | 3293265 | 2527386 | 2685655 | NA | | 2318149 |
| GFLS[Phospho (STY | 2 | A1L390 | 1551857 | 1596314 | 1253587 | 1406729 | 1723560 | 1502006 |
| IYBMS390S[Phosp | 2 | A1L390 | 586212.1 | 703314.1 | 697566 | 580441.7 | 808891.8 | 745552.6 |
| SBS[Phospho (STY | 3 | A1L390 | NA | 604569.9 | NA | | 784035 | 554084.8 |
| SPLS[Phospho (STY | 2 | A1L390 | 833264.3 | 1034303 | 867998.1 | 714042.4 | 990010.2 | 1039246 |
| S[Phospho (STY) JF | 2 | A1L390 | 801754.1 | 825141.1 | 840367.5 | 709674.4 | | 895440 |
| SSSLS[Phospho (S | 2 | A1L390 | | 728638 | 795307.5 | 637437.1 | 714943.1 | 806124.7 |
| RES[Phospho (STY | 2 | A1L390 | 386263.1 | NA | | 371454.2 | 364010.8 | 415585.1 |

7. How to run this tool locally?

NAGuideR is an open source software for non-commercial use and all codes can be obtained on our GitHub: <https://github.com/wangshisheng/NAGuideR>. If users want to run *NAGuideR* on their own computer, they should operate as below:

7.1 As this tool was developed with R, you may :

- a) Install R. You can download R from here: <https://www.r-project.org/>.
- b) Install RStudio. (Recommendatory but not necessary). You can download RStudio from here: <https://www.rstudio.com/>.
- c) Check packages. After installing R and RStudio, you should check whether you have installed these packages (shiny, shinyBS, shinyjs, shinyWidgets, DT, gdata, ggplot2, ggsci, openxlsx, data.table, DT, raster, Metrics, vegan, tidyverse, ggExtra, cowplot, Amelia, e1071, impute, SeqKnn, pcaMethods, norm, imputeLCMD, VIM, rrcovNA, mice, missForest). You may run the codes below to check them:

```
if(!require(pacman)) install.packages("pacman")
pacman::p_load(shiny, shinyBS, shinyjs, shinyWidgets, DT, gdata, ggplot2, ggsci,
openxlsx, data.table, DT, raster, Metrics, vegan, tidyverse, ggExtra, cowplot,
Amelia, e1071, impute, SeqKnn, pcaMethods, norm, imputeLCMD, VIM, rrcovNA, mice,
missForest)
```

Please note, you may find the SeqKnn package (<https://github.com/cran/SeqKnn>) can not be installed rightly as it has not been updated for a long time. If so, please download this package from here: https://github.com/wangshisheng/NAGuideR/blob/master/SeqKnn_1.0.1.tar.gz. Then you can install this package locally:

```
setwd('path') #path is where the two packages are.
install.packages("SeqKnn_1.0.1.tar.gz", repos = NULL, type="source")
```

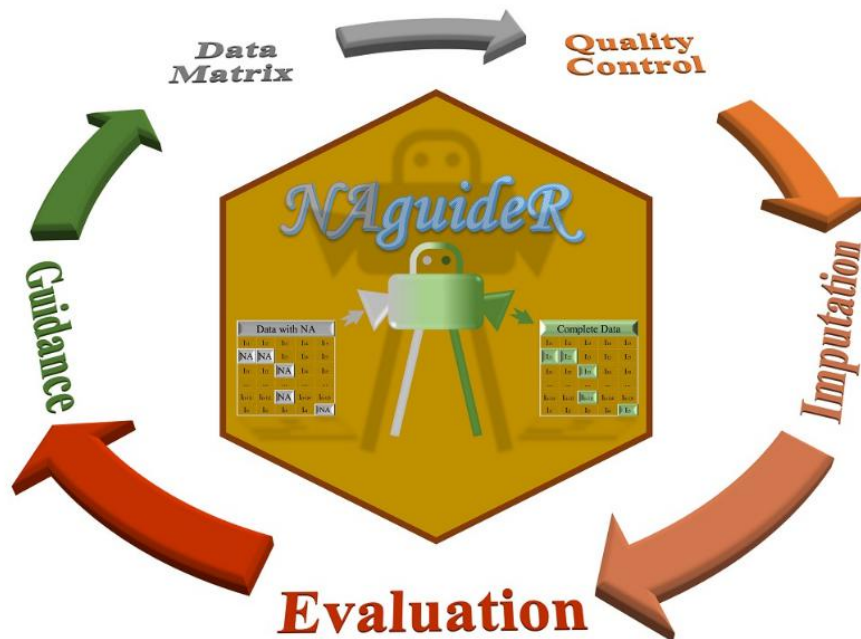
- d) Run this tool locally

```
if(!require(NAGuideR)) devtools::install_github("wangshisheng/NAGuideR")
library(NAGuideR)
NAGuideR_app()
```

Then *NAGuideR* will be started as below, and the detailed operation about *NAGuideR* can be found in the Supplementary Notes part 1-6:

~~ Dear Users, Welcome to NAguideR ~~

NAguideR is a web-based tool, which integrates 20 common missing value imputation methods and provides two categories of evaluation criteria (4 classic criteria and 4 proteomic criteria) to assess the imputation performance of various methods. We hope this tool could help scientists impute the missing values systematically and present valuable guidance to select one proper method for their own data. In addition, this tool supports both online access and local installation.



Basically, there are four main steps in NAguideR:

1. Uploading proteomics expression data and sample information data;
2. Data quality control;
3. Missing value imputation;
4. Performance evaluation;

After this, NAguideR can provide valuable guidance for users to select one proper method for their own data based on the evaluation results. Detailed introduction can be found in the **Help** part.

Finally, NAguideR is developed by R shiny (Version 1.3.2), and is free and open to all users with no login requirement. It can be readily accessed by all popular web browsers including Google Chrome, Mozilla Firefox, Safari and Internet Explorer 10 (or later), and so on. We would highly appreciate that if you could send your feedback about any bug or feature request to Shisheng Wang at wssdandan2009@outlook.com.

^_^ Enjoy yourself in NAguideR ^_^

III. Reference

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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. *Scientific reports*, **8**, 663.
8. Ibrahim, J.G., Chen, M.-H., Lipsitz, S.R. and Herring, A.H. (2005) Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, **100**, 332-346.
9. 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.
10. Verboven, S., Branden, K.V. and Goos, P. (2007) Sequential imputation for missing values. *Computational Biology and Chemistry*, **31**, 320-327.
11. Branden, K.V. and Verboven, S. (2009) Robust data imputation. *Computational Biology and Chemistry*, **33**, 7-13.
12. Buuren, S.v. and Groothuis-Oudshoorn, K. (2010) mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 1-68.
13. 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.
14. 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.
15. Templ, M., Kowarik, A. and Filzmoser, P. (2011) Iterative stepwise regression imputation using standard and robust methods. *Computational Statistics & Data Analysis*, **55**, 2793-2806.
16. 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**, 1-11.
17. Bruderer, R., Bernhardt, O.M., Gandhi, T., Miladinovic, S.M., Cheng, L.Y., Messner, S.,

- Ehrenberger, T., Zanutelli, 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. *Molecular & cellular proteomics : MCP*, **14**, 1400-1410.
18. 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. *Molecular & cellular proteomics : MCP*, **16**, 2296-2309.