NAR Genomics and Bioinformatics, 2024, **6**, lqae071
https://doi.org/10.1093/nargab/lqae071
Advance access publication date: 14 June 2024
Application Notes

omicsMIC: a comprehensive benchmarking platform for robust comparison of imputation methods in mass spectrometry-based omics data

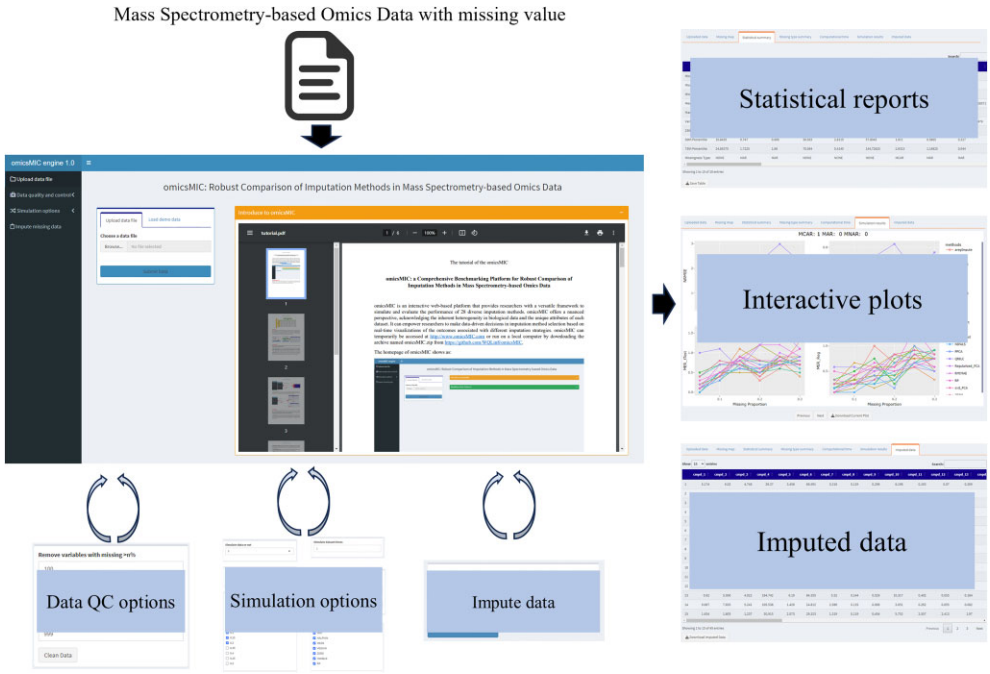
Weiqiang Lin [1], Jiadong Ji [2], Kuan-Jui Su [1], Chuan Qiu [1], Qing Tian [1], Lan-Juan Zhao [1], Zhe Luo [1], Chong Wu [3], Hui Shen [1] and Hongwen Deng [1,*]

[1]Tulane Center for Biomedical Informatics and Genomics, Deming Department of Medicine, School of Medicine, Tulane University, New Orleans, LA 70112, USA
[2]Institute for Financial Studies, Shandong University, Jinan, Shandong 250100, China
[3]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77230, USA
[*]To whom correspondence should be addressed. Tel: +1 504 988 1310; Email: hdeng2@tulane.edu

Abstract

Mass spectrometry is a powerful and widely used tool for generating proteomics, lipidomics and metabolomics profiles, which is pivotal for elucidating biological processes and identifying biomarkers. However, missing values in mass spectrometry-based omics data may pose a critical challenge for the comprehensive identification of biomarkers and elucidation of the biological processes underlying human complex disorders. To alleviate this issue, various imputation methods for mass spectrometry-based omics data have been developed. However, a comprehensive comparison of these imputation methods is still lacking, and researchers are frequently confronted with a multitude of options without a clear rationale for method selection. To address this pressing need, we developed omicsMIC (mass spectrometry-based omics with Missing values Imputation methods Comparison platform), an interactive platform that provides researchers with a versatile framework to evaluate the performance of 28 diverse imputation methods. omicsMIC offers a nuanced perspective, acknowledging the inherent heterogeneity in biological data and the unique attributes of each dataset. Our platform empowers researchers to make data-driven decisions in imputation method selection based on real-time visualizations of the outcomes associated with different imputation strategies. The comprehensive benchmarking and versatility of omicsMIC make it a valuable tool for the scientific community engaged in mass spectrometry-based omics research. omicsMIC is freely available at https://github.com/WQLin8/omicsMIC.

Graphical abstract



Received: October 13, 2023. **Revised:** April 25, 2024. **Editorial Decision:** May 22, 2024. **Accepted:** May 30, 2024

Introduction

The recent advancements in mass spectrometry-based omics, such as proteomics, lipidomics and metabolomics, have ushered in a new era of scientific discovery, advancing our understanding of biological mechanisms (1) and potentially reshaping the discovery of biomarkers (2,3), drug discovery and precision medicine (4) for human complex disorders by providing insights into molecular biology and enabling comprehensive molecular analysis. In particular, the integration of other omics data with mass spectrometry-based omics data has emerged as a potent strategy for advancing our comprehension of complex biological progress (5).

However, one of the main drawbacks of mass spectrometry-based omics data is that they typically contain a large proportion of missing values, even in the range of 30–50% (6,7). Commonly, these missing values can be attributed to either the genuine absence of the compound in the measured sample or the presence of the molecular feature at a concentration below the mass spectrometer's detection limit. Therefore, many studies handled missing values with simple value replacement methods such as zero (8), half of the minimum value (9) and the minimum value (10). However, the issue of missing data is a complex problem that can arise in various situations including the following. (i) Mass spectrometry techniques may encounter technical problems during data acquisition, such as instrument errors, signal interference or instrument malfunctions (11–13). (ii) Prior to mass spectrometry analysis, samples undergo a series of preparation and extraction steps. These steps may involve chemical reactions, sample handling or extraction processes, which introduce variability and uncertainty (11,12,14). (iii) Processing and interpreting mass spectrometry data is a complex process that involves steps such as signal denoising, background correction and quality filtering. During the processing, certain data points may not meet specific quality standards or the limitations of the algorithms. These issues can result in certain data points not being accurately recorded or obtained and thus being treated as missing data (15,16). Inadequately addressing these missing data can introduce bias in the subsequent statistical analysis and interpretation of mass spectrometry-based data, potentially compromising the reliability of the downstream analysis and the accuracy of the results.

Imputation is a common approach to dealing with missing data, which treats missing values using the information that is available from the existing data. So far, numerous imputation methods have been proposed for handling missing values in -omics studies. In this study, we roughly divide incorporated imputation methods into three categories: (i) simple value replacement; (ii) model-based approaches; and (iii) machine learning-based approaches. Simple value replacement is a commonly used technique for handling missing data in various fields such as zero, half of the minimum value and the minimum value. This strategy can quickly fill in missing values and allow for downstream analyses to be conducted on complete datasets. However, it may skew the distribution or underestimate measures of variance and lead to more bias (16,17). To account for this limitation of simple value replacement, many model-based imputation methods have been developed, which leverage statistical and computational models to estimate missing values, taking into account the patterns and relationships present in the data. Examples include Bayesian principal component analysis (BPCA) (18) and singular value decomposition (SVD) imputation (19). Furthermore, machine learning-based approaches have become increasingly popular, as they can handle diverse data distributions and complex relationships. For instance, K-nearest neighbor (KNN) imputations (20) use similarity between samples to predict missing values, and random forest-based methods (21) use the random forest ensemble algorithm to make predictions.

Numerous imputation methods have been proposed, and previous studies suggested that different methods may be required to achieve good performance under different circumstances (15,16,22). However, a comprehensive and systemic comparison of the performance of these imputation methods under different conditions is still lacking. In this study, we develop omicsMIC (mass spectrometry-based omics with Missing values Imputation methods Comparison platform), a user-friendly platform that provides a versatile framework for simulating and evaluating a diverse range of imputation strategies, tailored to users' specific datasets. Our platform can help users determine the most appropriate imputation method for their datasets with specific objectives and allow users to perform the imputation on their datasets once the preferred imputation approach is determined. We anticipate that this platform will be helpful for the community, particularly for researchers without an extensive background in computer science or programming within this biomedical field.

Materials and methods

Overview of the omicsMIC interactive platform

omicsMIC is a comprehensive application that allows advanced comparison of 28 imputation methods (Table 1) for mass spectrometry-based omics data in a user-interactive fashion. The omicsMIC includes Data quality and control, Data simulation and Data imputation. It introduces a potent data simulation feature, armed with seven customizable parameters. The whole workflow of the omicsMIC is shown in Figure 1A.

Missing mechanism

The missing completely at random (MCAR) mechanism (23) describes a process in which missing values cannot be attributed to either the presence or absence of specific molecular features in the samples. In simpler terms, this mechanism is characterized by the random occurrence of missing data, which can be considered as a complete absence of correlation between the missing and observed portions of the data. In the context of omicsMIC, the MCAR missingness will be simulated by randomly removing values from the dataset using the uniform distribution. Different levels of missingness will be simulated by removing varying proportions of the values.

The missing at random (MAR) mechanism (23) describes that the probability of missingness in a variable is related to observed data but not to unobserved data. In other words, the probability of a data point being missing depends on the values of other observed variables in the dataset, but it is not related to the missing variable itself. In omicsMIC, the MAR is modeled where a feature X_1 will lead to the missingness of another feature X_2 in the same sample. The simulation process starts by randomly choosing two different features: X_1 and X_2. The values of X_1 are arranged in ascending order. Next, a cut-off percentage point is randomly sampled from a χ^2 distribution and then normalized by dividing it by 30 (24). This normalized value is employed to determine the proportion of

Table 1. Brief description of imputation methods evaluated in omicsMIC

Category	Methods	Source	Reference
Simple value replacement	Zero		
	Half-min		
	Min		
	Mean		
	Median		
	hotdeck	R VIM package	(27)
	Random replacement		
Model-based imputation methods	BPCA	R pcaMethods package	https://github.com/hredestig/pcamethods
	QRILC	R imputeLCMD package	https://github.com/cran/imputeLCMD
	GSimp	https://github.com/WandeRum/GSimp	(26)
	mice	R mice package	(28)
	mice_cart	R mice package	(28)
	svd_PCA	R pcaMethods package	https://github.com/hredestig/pcamethods
	Regular_PCA	R missMDA package	(29)
	EM_PCA	R missMDA package	(29)
	PPCA	R pcaMethods package	https://github.com/hredestig/pcamethods
	aregImpute	R Hmisc package	https://github.com/harrelfe/Hmisc
	rmiMAE	https://github.com/NishithPaul/missingImputation/blob/main/rmiMAE.R	(30)
	wlsMisImp	https://github.com/NishithPaul/tWLSA	https://github.com/NishithPaul/tWLSA
	NIPALS	R pcaMethods package	https://github.com/hredestig/pcamethods
Machine learning-based imputation methods	mice_RF	R missRanger package	https://github.com/mayer79/missRanger
	Extra_Trees	R missRanger package	https://github.com/mayer79/missRanger
	missForest	R missForest package	(21)
	NLPCA	R pcaMethods package	https://github.com/hredestig/pcamethods
	GD_KNN	R VIM package	(27)
	Cor_KNN	https://github.com/WandeRum/GSimp	(26)
	trunc_KNN	https://github.com/WandeRum/GSimp	(26)
	ED_KNN	https://github.com/WandeRum/GSimp	(26)

the highest X_2 values to be marked as missing, thus simulating MAR missingness. This procedure will be repeated until the desired level of overall missing data for MAR is achieved.

The missing not at random (MNAR) mechanism (23) describes that the missingness of a variable is related to the unobserved data or values that are not included in the dataset. In other words, the probability of missingness depends on the actual value of the variable that is missing, and this missingness is not explained by the observed data alone. In omicsMIC, feature X_1 will be randomly selected, and its values will be sorted in ascending order. Then, a similar process to that of the MAR case will be applied, with the key difference being that missing values will be generated for X_1 when its values fall below the cut-off point. This process is repeated until the intended proportion of missing values is achieved.

Evaluation of the missingness mechanism

MCAR: in omicsMIC, pairwise correlations (Spearman and Kendall) will be computed between the missingness vector of each feature (where present values were replaced with 1 and missing values with 0) and all other features in the dataset. Features will be categorized as MCAR when no statistically significant correlations are detected between any pair of features.

MNAR: the Kolmogorov–Smirnov (KS) goodness-of-fit test (or the Cucconi test with Benjamini–Hochberg correction) will be used on the remaining features to assess whether their distributions exhibit left truncation. Features demonstrating left-truncated distributions in comparison with left-censored normal distributions will be classified as MNAR.

MAR: the remaining features will be considered as MAR.

Performance evaluation

According to different downstream analysis purposes, five matrices will be applied to quantitatively evaluate the performance of different imputation methods:

Normalized root mean square error (NRMSE) (18) is used to comprehensively assess the imputation accuracy of data imputation methods. NRMSE calculates the difference in the estimation between the imputed and original values:

$$NRMSE = \sqrt{\frac{mean\left((X^{orig} - X^{imp})^2\right)}{var\left(X^{orig}\right)}},$$

where X^{orig} represents the original complete data and X^{imp} represents the imputed data. NRMSE ranges between 0 and 1, with a lower value indicating better accuracy. A value of 0 means that the predicted values are a perfect match with the actual values, while a value of 1 means that the predictions have no predictive power and are as accurate as simply using the mean of the actual values.

PCA–Procrustes analysis (PCA-P) (16): PCA will be performed on the original dataset and imputed data, to reduce their dimensionality while retaining the most important pat-

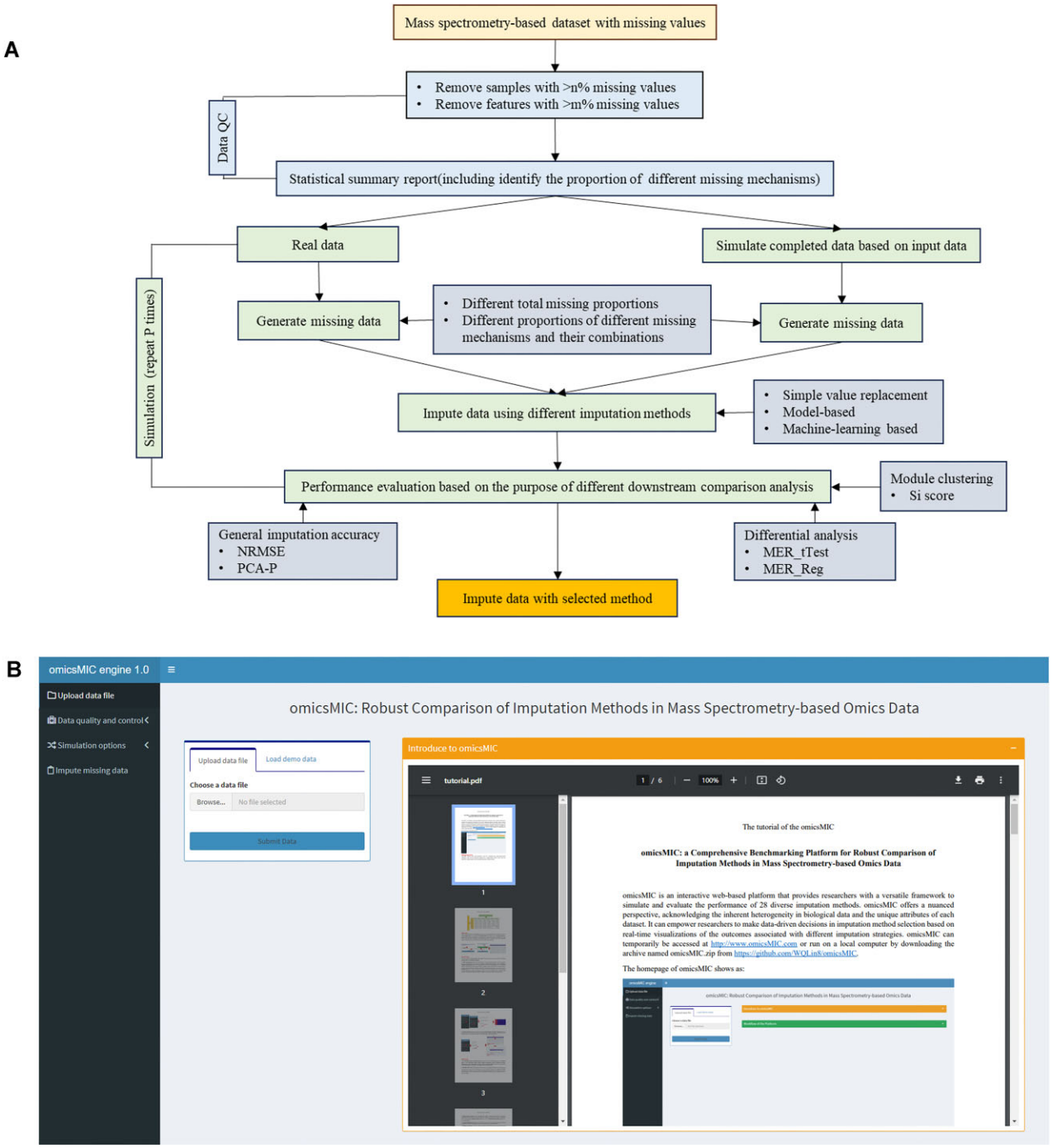


Figure 1. The workflow (**A**) and main interface (**B**) of omicsMIC.

terns in the data and extracting the first two principal components from the PCA results. Subsequently, Procrustes analysis will be applied in combination with PCA to evaluate the effectiveness of each imputation method in preserving the underlying patterns and relationships present in the original data. A smaller PCA–Procrustes value indicates better alignment, signifying that the imputation method is more accurate and effectively preserves the overall structure of the data.

Misclassification error rate (MER) is performed to compare the ability of different imputation methods to correctly recover variables with differences in the imputed data compared with the original data. It is defined as the proportion of incorrectly identified significant differences to the total number of true significant differences:

$$MER = avg\left(\frac{F}{T}\right),$$

where F denotes the total count of false positives for each variable, while T represents the total count of true positives for each variable. A lower MER value enhances the reliability of identifying genuinely significant outcomes while minimizing the potential for false discoveries. Considering the practical requirements for real-world applications, we evaluate the performance of different imputation methods based on t-test (MER_tTest) and regression analyses (MER_Reg).

The silhouette coefficient (*Si* score) (25) is used to evaluate the clustering performance of various imputation methods. It combines cohesion and separation, which can compare the quality of clustering outcomes obtained from different imputation techniques. It is defined as:

$$Si = \frac{b_i - a_i}{\max\left(b_i,\ a_i\right)},$$

$$\Delta Si = avg\left(\left|Si^{imp} - Si^{orig}\right|\right),$$

where a_i signifies the average distance between the i-th sample and all other samples in the same cluster, b_i signifies the average distance between the i-th sample and all samples in each cluster, and Si^{imp} and Si^{orig} signify the *Si* score in the imputed dataset and original dataset separately. In addition weighted gene co-expression network analysis (WGCNA) will be performed to do clustering. The lower the ΔSi score is, the better the clustering recovery performance.

Results and discussion

The main interface of omicsMIC displays the analysis steps on an expandable menu on the left side and shows a tutorial section that includes recommendations for parameter selection and explanations of simulated results based on the demo data on the right side (Figure 1B). OmicsMIC starts with uploading a data file (Figure 2A), containing mass spectrometry-based quantitative data. The file should be organized in a sample–feature format, where each row represents a unique sample, and each column represents a distinct feature. OmicsMIC accommodates CSV format files. Once the data are uploaded, the platform will provide a quick overview of the uploaded dataset, allowing users to verify that the correct data have been uploaded (Figure 2B). We also provide a targeted metabolomics dataset (26) to help users familiarize themselves with the functionalities of the platform (Figure 2C).

Data quality and control

Within the omicsMIC platform, two vital data quality control functionalities are implemented, allowing users to tailor their data preparation according to their specific requirements.

(i) Filter missing data: our platform allows users to control data quality precisely by customizing the threshold for removing samples and features with excessive missing data. Users can specify missing rate thresholds (Figure 2D) so that samples and features with missing rates over the thresholds will be excluded from the subsequent analyses. Upon user-defined criteria, omicsMIC will generate a missing data pattern heatmap (Figure 2E), visually representing the distribution of missing values across samples and features. This heatmap offers valuable insights into the data's completeness and assists users in identifying patterns or trends in missing data.

(ii) Statistical summary: in this section, we present a comprehensive statistical summary after the removal of samples and features surpassing the specified missing data threshold, providing users with essential insights into the characteristics of their dataset alongside the statistical metrics and missing mechanisms. Users have the flexibility to select correlation methods and goodness-of-fit indices (Figure 2F), tailoring the analysis to identify missing mechanisms. For each variable in the dataset, omicsMIC will calculate and report the following statistical metrics: missing number, missing %, min, mean, max, variance, 25th percentile, 50th percentile, 75th percentile and missingness type (Figure 2G). Additionally, it will also provide the user with a summary report of the proportion of different missing mechanisms (Figure 2H).

Data simulation

The omicsMIC platform introduces a potent data simulation feature, armed with seven customizable parameters. These parameters have been thoughtfully crafted to equip users with a versatile toolkit for effectively appraising various imputation methods. Here, we describe these parameters in detail:

(i) *Simulate data or not*: this parameter empowers users to decide whether they wish to generate simulated data based on the characteristics of their uploaded dataset for imputation method assessment. If users want to use real data to carry out downstream simulations, the missing value in the uploaded dataset will be excluded by removing samples and features to obtain a completed dataset (Figure 2I).

(ii) *Simulate dataset times*: if the previous parameters were set to require simulated data, users can define the number of times they wish to generate simulated data. Commonly, multiple iterations introduce diversity into the simulations and thus increase the reliability of the results (Figure 2J).

(iii) *Simulate missingness times*: users can adjust the parameter for missing data generation, which specifies the number of missing data instances created in each completed dataset (Figure 2K).

(iv) *Missing percentage*: users can select pre-defined overall missing proportion (5–50%) and/or customize the overall missing proportion in the simulated dataset (Figure 2L).

(v) *Missing type ratios*: this parameter defines different missing mechanism ratio settings (Figure 2M), pivotal for evaluating imputation method stability under diverse missing mechanisms. In this study, we consider three types of missing mechanisms: missing not at random (MNAR), missing at random (MAR) and missing completely at random (MCAR). In omicsMIC, seven different ratio combinations of these three mechanisms are pre-defined (MCAR, MAR, MNAR): (0.6, 0.2, 0.2), (0.2, 0.6, 0.2), (0.2, 0.2, 0.6), (1/3, 1/3, 1/3), (1, 0, 0), (0, 1, 0) and (0, 0, 1). Users can also customize the proportions of different missing mechanisms. The missing type ratios calculated from the data quality and control step will be included as a scenario for the simulation automatically.

(vi) *Imputation methods*: in the omicsMIC application, a total of 28 imputation methods are included. Imputation methods are categorized into simple value replacement, model-based and machine learning-based techniques, offering users flexibility tailored to their research needs (Figure 2N–P).

(vii) *Evaluation methods*: omicsMIC offers various evaluation metrics for assessing imputation methods: NRMSE, PCA-P, MER_tTest, MER_Reg and *Si* score. Different evaluation metrics will help guide users to select methods that best suit their research objectives (Figure 2Q).

The 'Data Simulation' section presents unparalleled flexibility and control for imputation method assessment. Leveraging these parameters, researchers can conduct rigorous evaluations, make informed decisions and select the most appro-

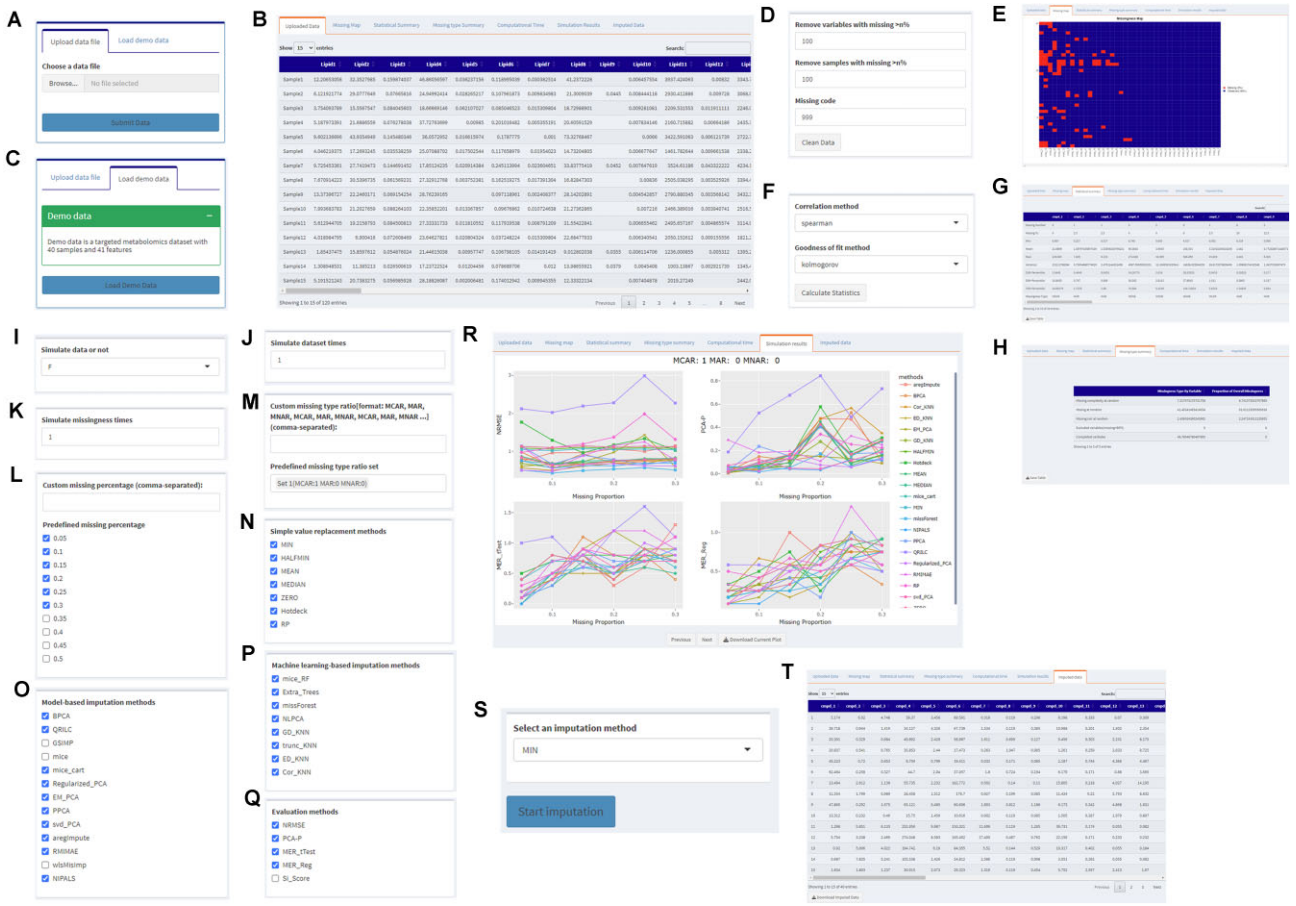


Figure 2. omicsMIC begins by uploading your data (**A**) or loading demo data (**B**) and a brief review will be provided (**C**). Following the data upload, it performs data quality control (QC) through pre-processing based on pre-defined thresholds for missing samples and variables (**D**). A heatmap (**E**) will be generated to visually represent the missing data after applying the filtering process. Subsequently, appropriate correlation methods are selected and goodness-of-fit tests are conducted (**F**) to evaluate different patterns of missingness. A descriptive statistical table is provided (**G**), with a summary table (**H**) summarizing the various patterns of missing data. In the simulation section, it provides users with a control panel featuring seven parameters that allow customization of the simulation scenario: *Simulate data or not* (**I**) provides the user with a choice to use simulation data or real data to perform the simulation. *Simulate dataset times* (**J**) and *Simulate Missingness times* (**K**) configure the times of generating simulated data and times of generating missing data corresponding. *Missing percentage* (**L**) sets up the percentage of missing value. *Missing type ratios* (**M**) provides different missing type combinations. *Imputation methods* (**N–P**) provides 28 imputation methods for comparison analysis. Five matrices from the *Evaluation methods* (**Q**) configure panel can be used to do carry out imputation performance evaluation. The comparison results (**R**) are presented as interactive figures which can be exported as PNG files. Finally, the imputation function (**S**) is provided to use the appropriate method to carry out imputation, and imputed data can be downloaded for further downstream data analysis (**T**).

priate imputation methods, thereby enhancing the quality and trustworthiness of their downstream analysis. The comparison results are presented as interactive figures (Figure 2R) which can be exported as static PNG files.

Data imputation

Based on simulation results, the users can utilize omicsMIC to perform the imputation on their quality control-processed datasets using an appropriate imputation method (Figure 2S) and download the imputed dataset (Figure 2T) for further downstream data analysis.

Conclusion

omicsMIC offers a versatile framework for simulating and evaluating a wide array of imputation strategies for mass spectrometry-based omics data. Given the inherent heterogeneity of biological data, omicsMIC equips users with real-time visualizations of imputation outcomes, facilitating informed and rational method selection. Notably, most imputation strategies incorporated in the omicsMIC platform are not only for mass spectrometry-based omics data but can also be applied to other types of continuous data, and omicsMIC has the potential for future updates to accommodate newly developed methods and additional evaluation criteria. Furthermore, as the source codes of omicsMIC will be publicly shared, omicsMIC is also a valuable tool for the development and evaluation of novel imputation methods.

Data availability

omicsMIC is freely available at https://github.com/WQLin8/omicsMIC and https://doi.org/10.5281/zenodo.10016741.

Funding

This work was partially benefited by grants from the National Institutes of Health [U19AG055373, R01AG061917 and R01AR069055].

Conflict of interest statement

None declared.

References

1. Dai,X. and Shen,L. (2022) Advances and trends in omics technology development. *Front. Medi*, **9**, 911861.
2. Núñez,E., Fuster,V., Gómez-Serrano,M., Valdivielso,J.M., Fernández-Alvira,J.M., Martínez-López,D., Rodríguez,J.M., Bonzon-Kulichenko,E., Calvo,E., Alfayate,A., *et al.* (2022) Unbiased plasma proteomics discovery of biomarkers for improved detection of subclinical atherosclerosis. *EBioMedicine*, **76**, 103874.
3. Tolstikov,V., Moser,A.J., Sarangarajan,R., Narain,N.R. and Kiebish,M.A. (2020) Current status of metabolomic biomarker discovery: impact of study design and demographic characteristics. *Metabolites*, **10**, 224.
4. Clarke,N.J. (2016) Mass spectrometry in precision medicine: phenotypic measurements alongside pharmacogenomics. *Clin. Chem.*, **62**, 70–76.
5. Khan,S., Ince-Dunn,G., Suomalainen,A. and Elo,L.L. (2020) Integrative omics approaches provide biological and clinical insights: examples from mitochondrial diseases. *J. Clin. Invest.*, **130**, 20–28.
6. Hrydziuszko,O. and Viant,M.R. (2012) Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. *Metabolomics*, **8**, S161–S174.
7. Webb-Robertson,B.J., Wiberg,H.K., Matzke,M.M., Brown,J.N., Wang,J., McDermott,J.E., Smith,R.D., Rodland,K.D., Metz,T.O., Pounds,J.G., *et al.* (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.
8. Wang,G., Heijs,B., Kostidis,S., Mahfouz,A., Rietjens,R.G.J., Bijkerk,R., Koudijs,A., van der Pluijm,L.A.K., van den Berg,C.W., Dumas,S.J., *et al.* (2022) Analyzing cell-type-specific dynamics of metabolism in kidney repair. *Nat. Metab.*, **4**, 1109–1118.
9. Hagenbeek,F.A., Pool,R., van Dongen,J., Draisma,H.H.M., Jan Hottenga,J., Willemsen,G., Abdellaoui,A., Fedko,I.O., den Braber,A., Visser,P.J., *et al.* (2020) Heritability estimates for 361 blood metabolites across 40 genome-wide association studies. *Nat. Commun.*, **11**, 39.
10. Talmor-Barkan,Y., Bar,N., Shaul,A.A., Shahaf,N., Godneva,A., Bussi,Y., Lotan-Pompan,M., Weinberger,A., Shechter,A., Chezar-Azerrad,C., *et al.* (2022) Metabolomic and microbiome profiling reveals personalized risk factors for coronary artery disease. *Nat. Med.*, **28**, 295–302.
11. Buonarati,M.H. and Schoener,D. (2019) Investigations beyond standard operating procedure on internal standard response. *Bioanalysis*, **11**, 1669–1678.
12. Fraier,D., Ferrari,L., Heinig,K. and Zwanziger,E. (2019) Inconsistent internal standard response in LC-MS/MS bioanalysis: an evaluation of case studies. *Bioanalysis*, **11**, 1657–1667.
13. Le Blaye,O. (2019) Variations in internal standard response: some thoughts and real-life cases. *Bioanalysis*, **11**, 1715–1725.
14. Bijlsma,S., Bobeldijk,I., Verheij,E.R., Ramaker,R., Kochhar,S., Macdonald,I.A., van Ommen,B. and Smilde,A.K. (2006) Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Anal. Chem.*, **78**, 567–574.
15. Wilson,M.D., Ponzini,M.D., Taylor,S.L. and Kim,K. (2022) Imputation of missing values for multi-biospecimen metabolomics studies: bias and effects on statistical validity. *Metabolites*, **12**, 671.
16. 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.
17. Deng,Y., Chang,C., Ido,M.S. and Long,Q. (2016) Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Sci. Rep.*, **6**, 21689.
18. Oba,S., Sato,M.A., Takemasa,I., Monden,M., Matsubara,K. and Ishii,S. (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**, 2088–2096.
19. 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.
20. Franco-Lopez,H., Ek,A.R. and Bauer,M.E. (2001) Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. *Remote Sens. Environ.*, **77**, 251–274.
21. Stekhoven,D.J. and Bühlmann,P. (2012) MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, **28**, 112–118.
22. Do,K.T., Wahl,S., Raffler,J., Molnos,S., Laimighofer,M., Adamski,J., Suhre,K., Strauch,K., Peters,A., Gieger,C., *et al.* (2018) Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics*, **14**, 128.
23. Rubin,D.B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.
24. 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 Bioinf.*, **20**, 492.
25. Rousseeuw,P.J. (1987) Silhouettes—a graphical aid to the interpretation and validation of cluster-analysis. *J. Comput. Appl. Math.*, **20**, 53–65.
26. 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.
27. Kowarik,A. and Templ,M. (2016) Imputation with the R package VIM. *J. Stat. Softw.*, **74**, 1–16.
28. Buuren,S.v. and Taylor and FrancisTaylor and Francis (2018) In: *Flexible Imputation of Missing Data*. 2nd edn., CRC Press, Boca Raton, FL.
29. Josse,J. and Husson,F. (2016) missMDA: a package for handling missing values in multivariate data analysis. *J. Stat. Softw.*, **70**, 1–31.
30. Kumar,N., Hoque,M.A. and Sugimoto,M. (2021) Kernel weighted least square approach for imputing missing values of metabolomics data. *Sci. Rep.*, **11**, 11108.

Received: October 13, 2023. **Revised:** April 25, 2024. **Editorial Decision:** May 22, 2024. **Accepted:** May 30, 2024