PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

GSimp: A Gibbs sampler based left-censored missing value imputation approach for metabolomics studies

Runmin Wei[1,2☯], Jingye Wang[1☯*], Erik Jia[3], Tianlu Chen[4], Yan Ni[1], Wei Jia[1]

1 Metabolomics Shared Resource, University of Hawaii Cancer Center, Honolulu, Hawaii, United States of America, **2** Department of Molecular Biosciences and Bioengineering, University of Hawaii at Manoa, Honolulu, Hawaii, United States of America, **3** High School, Punahou School, Honolulu, Hawaii, United States of America, **4** Shanghai Key Laboratory of Diabetes Mellitus and Center for Translational Medicine, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai, China

☯ These authors contributed equally to this work.
* jingyew@hawaii.edu

OPEN ACCESS

Citation: Wei R, Wang J, Jia E, Chen T, Ni Y, Jia W (2018) GSimp: A Gibbs sampler based left-censored missing value imputation approach for metabolomics studies. PLoS Comput Biol 14(1): e1005973. https://doi.org/10.1371/journal.pcbi.1005973

Editor: Jens Nielsen, Chalmers University of Technology, SWEDEN

Received: September 12, 2017

Accepted: January 12, 2018

Published: January 31, 2018

Data Availability Statement: All datasets are freely available at: https://github.com/WandeRum/GSimp.

Funding: This work was supported by National Institutes of Health/National Cancer Institute Grant 1U01CA188387-01A1, National Institutes of Health/National Cancer Institute Grant P30 CA071789, and National Natural Science Foundation of China (No. 31501079). The funders had no role in study design, data collection and

Abstract

Left-censored missing values commonly exist in targeted metabolomics datasets and can be considered as missing not at random (MNAR). Improper data processing procedures for missing values will cause adverse impacts on subsequent statistical analyses. However, few imputation methods have been developed and applied to the situation of MNAR in the field of metabolomics. Thus, a practical left-censored missing value imputation method is urgently needed. We developed an iterative Gibbs sampler based left-censored missing value imputation approach (GSimp). We compared GSimp with other three imputation methods on two real-world targeted metabolomics datasets and one simulation dataset using our imputation evaluation pipeline. The results show that GSimp outperforms other imputation methods in terms of imputation accuracy, observation distribution, univariate and multivariate analyses, and statistical sensitivity. Additionally, a parallel version of GSimp was developed for dealing with large scale metabolomics datasets. The R code for GSimp, evaluation pipeline, tutorial, real-world and simulated targeted metabolomics datasets are available at: https://github.com/WandeRum/GSimp.

Author summary

Missing values caused by the limit of detection/quantification (LOD/LOQ) were widely observed in mass spectrometry (MS)-based targeted metabolomics studies and could be recognized as missing not at random (MNAR). MNAR leads to biased parameter estimations and jeopardizes following statistical analyses in different aspects, such as distorting sample distribution, impairing statistical power, etc. Although a wide range of missing value imputation methods was developed for–omics studies, a limited number of methods was designed appropriately for the situation of MNAR currently. To alleviate problems caused by MNAR and to facilitate targeted metabolomics studies, we developed a Gibbs sampler based missing value imputation approach, called GSimp, which is public-

analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

accessible on GitHub. And we compared our method with existing approaches using an imputation evaluation pipeline on both of the real-world and simulated metabolomics datasets to demonstrate the superiority of our method from different perspectives.

Introduction

Missing values are commonly existed in mass spectrometry (MS) based metabolomics datasets. Many statistical methods require a complete dataset, which makes missing data an inevitable problem for subsequent data analysis. Generally speaking, missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR) are three commonly accepted missing types [1,2]. When the probability of a missing value is depended on other observed variables but not the value itself, it is regarded as MAR [1,2] (e.g., false peak matching, deconvolution of co-eluting compounds). MCAR is from completely unexpected missingness without any relationships with other variables (e.g., stochastic fluctuations, random errors from incomplete derivatization or ionization). Targeted metabolomics studies have been widely used for the accurate quantification of specific groups of metabolites. Due to the limit of compound quantifications (LOQ), missing values are usually caused by signal intensities lower than LOQ, also known as left-censored missing, which can be assigned to MNAR.

The processing of missing values has been developed and studied in MS data, which is an indispensable step in the metabolomics data processing pipeline [3]. One simple solution is the substitution of missing by determined values, such as zero, half of the minimum value (HM) or LOQ/c where c denotes a positive integer. Determined value substitutions, although commonly applied for dealing with missing values in metabolomics studies [4–6], can significantly affect the subsequent statistical analyses in different ways (e.g., underestimate variances of the variable, decrease statistical power, fabricate pseudo-clusters among observations, etc.) [1]. Advanced statistical imputation methods have been developed for high-dimensional–omics studies (e.g., k-nearest neighbors (kNN) [7], singular value decomposition (SVD) [8,9], random forest (RF) [10]) that are available to users on several metabolomics data analysis software [11–15]. MetaboAnalyst [15–17] is a popular metabolomics data processing web-tool providing kNN, Probabilistic PCA (PPCA), Bayesian PCA (BPCA), SVD, or substitution by determined values (HM, mean, median, minimum). However, most advanced statistical imputation methods are mainly aiming at imputing MCAR/MAR and not suitable for the situation of MNAR. So far, a limited number of approaches dealing with left-censored missing values were applied by researchers [18,19]. Quantile regression approach for left-censored missing (QRILC) imputes missing data using random draws from a truncated distribution with parameters estimated using quantile regression [18]. Although this imputation keeps the overall distribution of missing parts compared to determined value substitutions, it may produce stochastic imputed values since no extra information is used for the prediction of missing parts. Another imputation method recently developed for MNAR is k-nearest neighbor truncation (kNN-TN) [19]. This approach applies Maximum Likelihood Estimators (MLE) for the means and standard deviations of missing variables based on truncated normal distribution. Then a Pearson correlation based kNN imputation method was implemented on standardized data. Although the author stated that kNN-TN could impute both MNAR and MAR, the imputed values were entirely dependent on the nearest neighbors while no constraint was placed upon the imputation. Thus, this approach might cause an overestimation of MNAR missing values.

To reduce adverse effects caused by missing values in following metabolomics data analyses, we developed a left-censored missing value imputation framework, GSimp, where a prediction model was embedded in an iterative Gibbs sampler. Next, we compared GSimp with HM, QRILC, and kNN-TN on two real-world metabolomics datasets and one simulation dataset to demonstrate the advantages of GSimp regarding imputation accuracy, observation distribution, univariate and multivariate analysis [20], and sensitivity. Our findings indicate that GSimp is a robust method in handling left-censored missing values in targeted metabolomics studies.

Results

Gibbs sampler in GSimp

A variable containing missing elements from free fatty acids (FFA) dataset was randomly selected to track the sequence of corresponding parameters and estimates across the first 500 iterations out of a total of 2000 (100 × 20) iterations using GSimp. From Fig 1, we can observe that both fitted value $\hat{y}$ and sample value $\tilde{y}$ reach to the convergence after several iterations and the standard deviation estimate σ drop to a steady state of small values. In addition, an upper constraint for the distribution of $\tilde{y}$ indicated that it was drawn from a truncated normal distribution.

Imputation comparisons

We evaluated four different MNAR imputation/substitution methods on FFA, bile acids (BA) targeted metabolomics and simulation datasets. First, we measured the imputation performances using label-free approaches. Sum of ranks (SOR) was used to measure the imputation accuracy regarding the imputed values of each missing variable. From the upper panel of Fig 2, we can observe that GSimp has the best performance with the lowest SOR across all varying numbers of missing variables in both FFA and BA datasets. To measure the extent of imputation induced distortion on observation distributions, the PCA-Procrustes analysis was

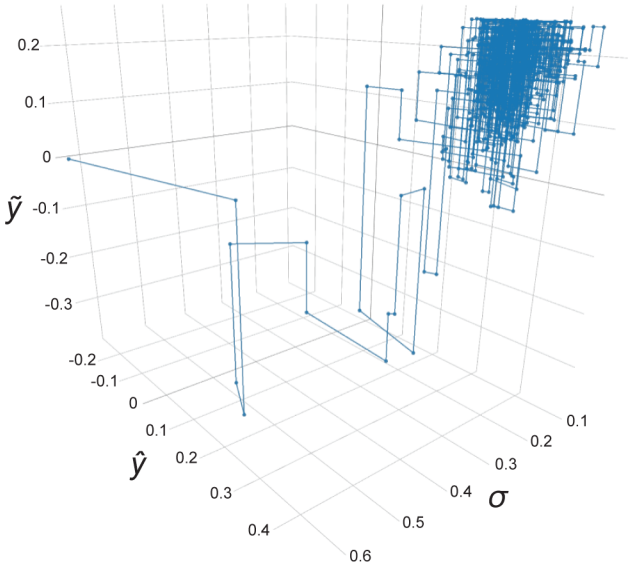

Fig 1. Sequentially parameters updating in GSimp. The first 500 iterations out of a total of 2000 (100×20) iterations using GSimp where $\hat{y}$, $\tilde{y}$ and σ represent fitted value, sample value and standard deviation correspondingly.

https://doi.org/10.1371/journal.pcbi.1005973.g001

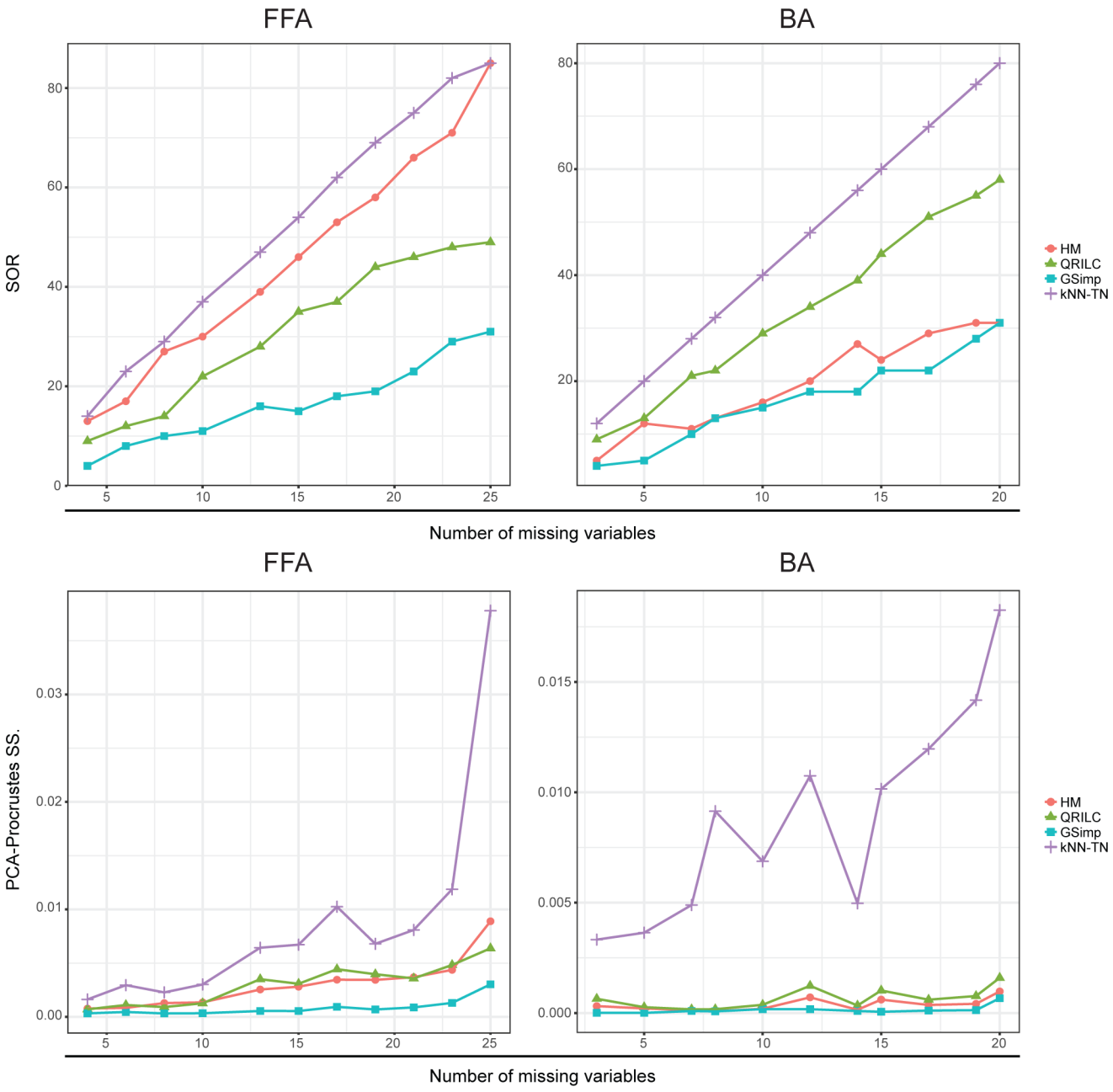

Fig 2. Evaluations of different imputation methods using unlabeled approaches. SOR on FFA dataset (upper left) and BA dataset (upper right) along with different numbers of missing variables based on four imputation methods: HM (red circle), QRILC (green triangle), GSimp (blue square), and kNN-TN (purple cross). PCA-Procrustes sum of squared errors on FFA dataset (lower left) and BA dataset (lower right) along with different numbers of missing variables based on four imputation methods: HM (red circle), QRILC (green triangle), GSimp (blue square), and kNN-TN (purple cross).

https://doi.org/10.1371/journal.pcbi.1005973.g002

conducted between the original data and imputed data. The lower panel of Fig 2 shows that GSimp has the lowest Procrustes sum of squared errors compared to other methods, which means GSimp kept the overall observation distribution of original dataset with the least distortions.

Then, we measured the imputation performances with clinical group information provided. We compared the results of univariate and multivariate analyses for imputed and original

datasets. Since this is a case-control study, student's *t*-tests were applied for univariate analyses. Then we compared the results by calculating Pearson's correlation between log-transformed *p*-values calculated from imputed and original data for missing variables. Again, GSimp performs best with the highest correlations among four methods (upper panel of Fig 3) along with different numbers of missing variables, and it implies GSimp keeps the most original biological

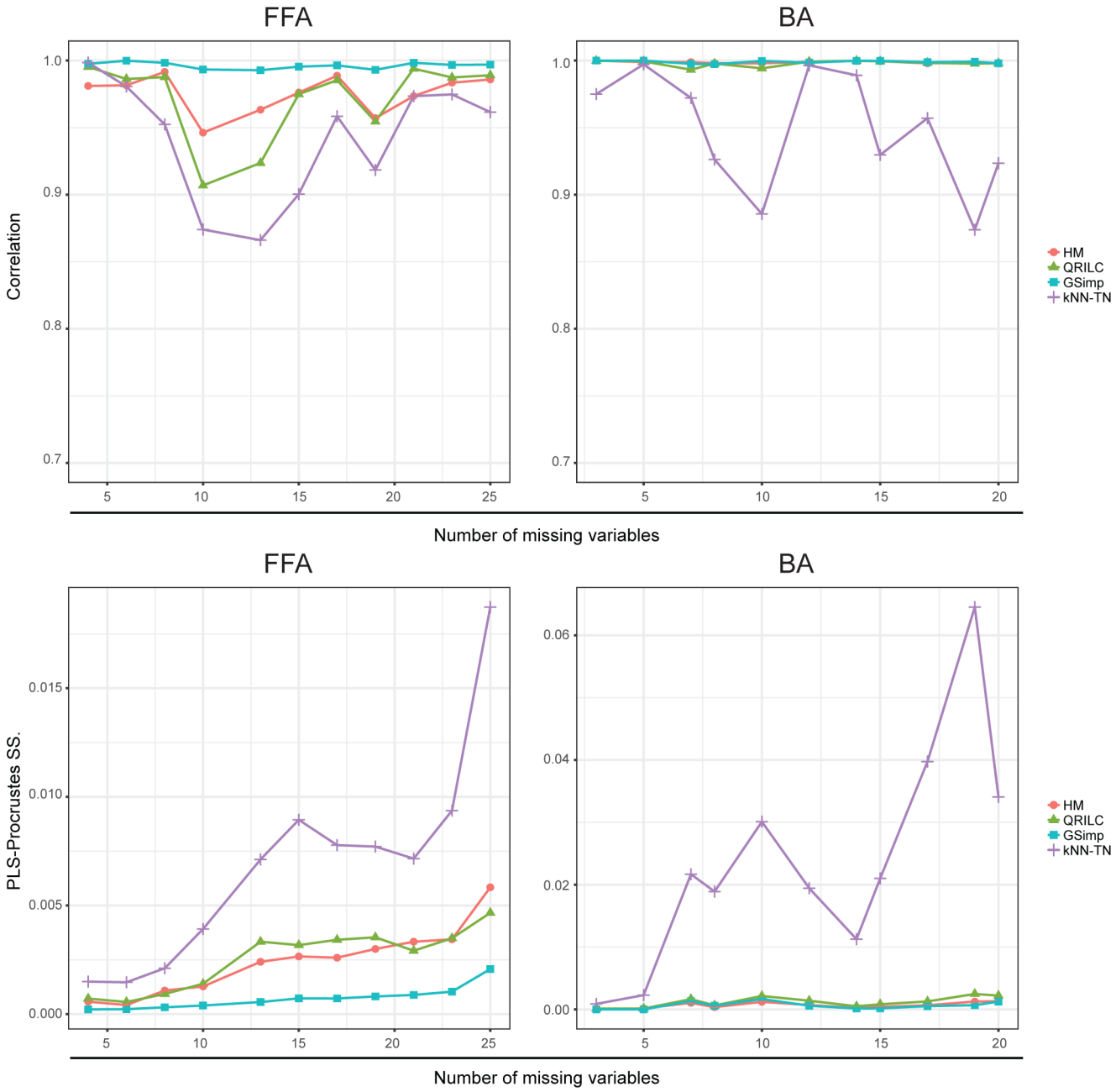

Fig 3. Evaluations of different imputation methods using labeled approaches. Pearson's correlation between log-transformed p-values of student's t-tests on FFA dataset (upper left) and BA dataset (upper right) along with different numbers of missing variables based on four imputation methods: HM (red circle), QRILC (green triangle), GSimp (blue square), and kNN-TN (purple cross). PLS-Procrustes sum of squared errors on FFA dataset (lower left) and BA dataset (lower right) along with different numbers of missing variables based on four imputation methods: HM (red circle), QRILC (green triangle), GSimp (blue square), and kNN-TN (purple cross).

https://doi.org/10.1371/journal.pcbi.1005973.g003

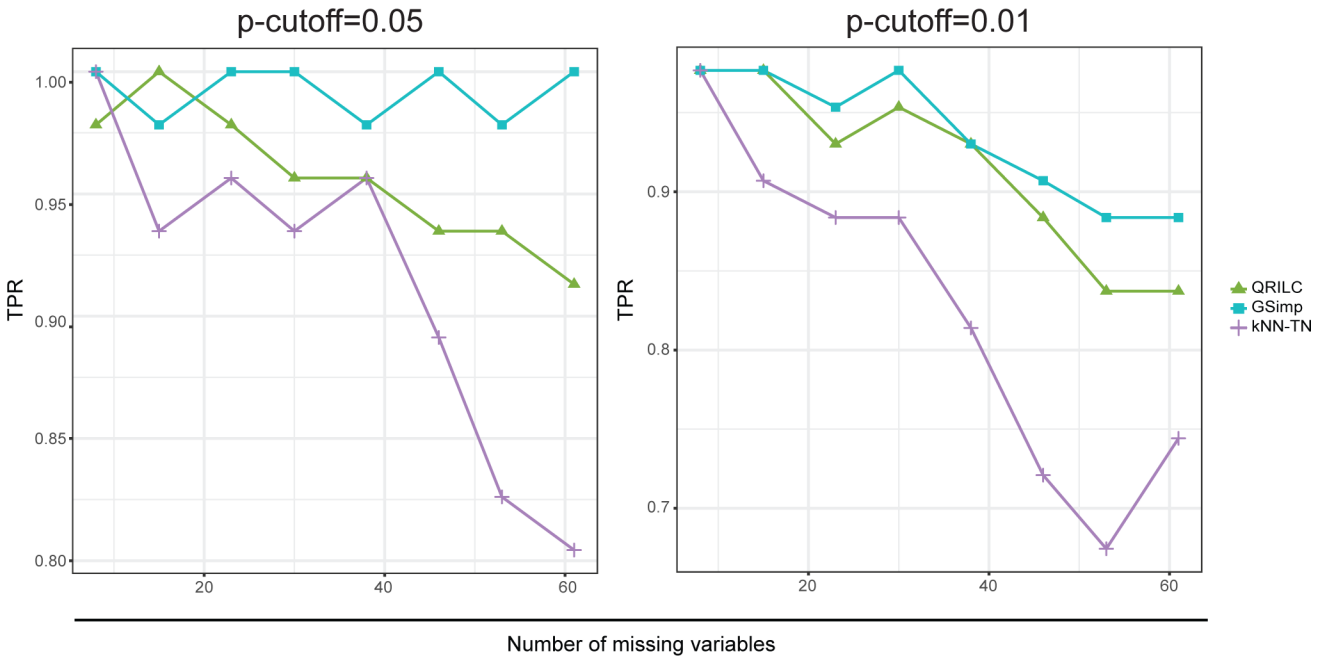

Fig 4. Evaluations of different imputation methods using TPR for various *p*-cutoffs on simulation dataset. *TPR* along with different numbers of missing variables based on three imputation methods: QRILC (green triangle), GSimp (blue square), and kNN-TN (purple cross) among different p-cutoff = 0.05 (left panel), and 0.01 (right panel).

https://doi.org/10.1371/journal.pcbi.1005973.g004

variations regarding the univariate analyses results. For the multivariate analyses, we applied PLS-DA to distinguish the group differences. Similarly, we conducted PLS-Procrustes analysis while PLS was employed as a supervised dimension reduction technique. The lower panel of Fig 3 demonstrates that GSimp preferably restores the original observation distribution with the lowest Procrustes sum of squared errors among four imputation methods.

On the simulation dataset, we compared QRILC, kNN-TN, and GSimp using same approaches. Consistent results were recognized (S1 Fig), and GSimp presents the best performances on the simulation dataset with the lowest SOR and PCA/PLS-Procrustes sum of squared errors and the highest correlation of univariate analysis results. Moreover, to examine the influences of statistical power using different imputation methods, we calculated the true positive rate (*TPR*) as the capacities to detect differential variables on different imputation datasets. Again, with both *p*-cutoff of 0.05 and 0.01, GSimp shows the overall highest *TPR* over different missing numbers (Fig 4). This implies that GSimp impairs the sensitivity to the least extent among three methods, which is reasonable since GSimp also keeps the highest correlation of *p*-values in previous comparisons.

Discussion

The purpose of this study is to develop a left-censored missing value imputation approach for targeted metabolomics data analysis. We evaluated GSimp with other three imputation methods (kNN-TN, QRILC, and HM) and suggested that GSimp was superior to others using different evaluation methods. To illustrate the performance of GSimp, we randomly selected one variable containing missing values from FFA dataset (Fig 5) to compare the imputed values and original values. Although determined value substitution (e.g., HM) were widely used by researchers in the field of metabolomics, our results indicated that HM could severely distort the data distribution (upper left panel of Fig 5), thus impairing subsequent analyses. In

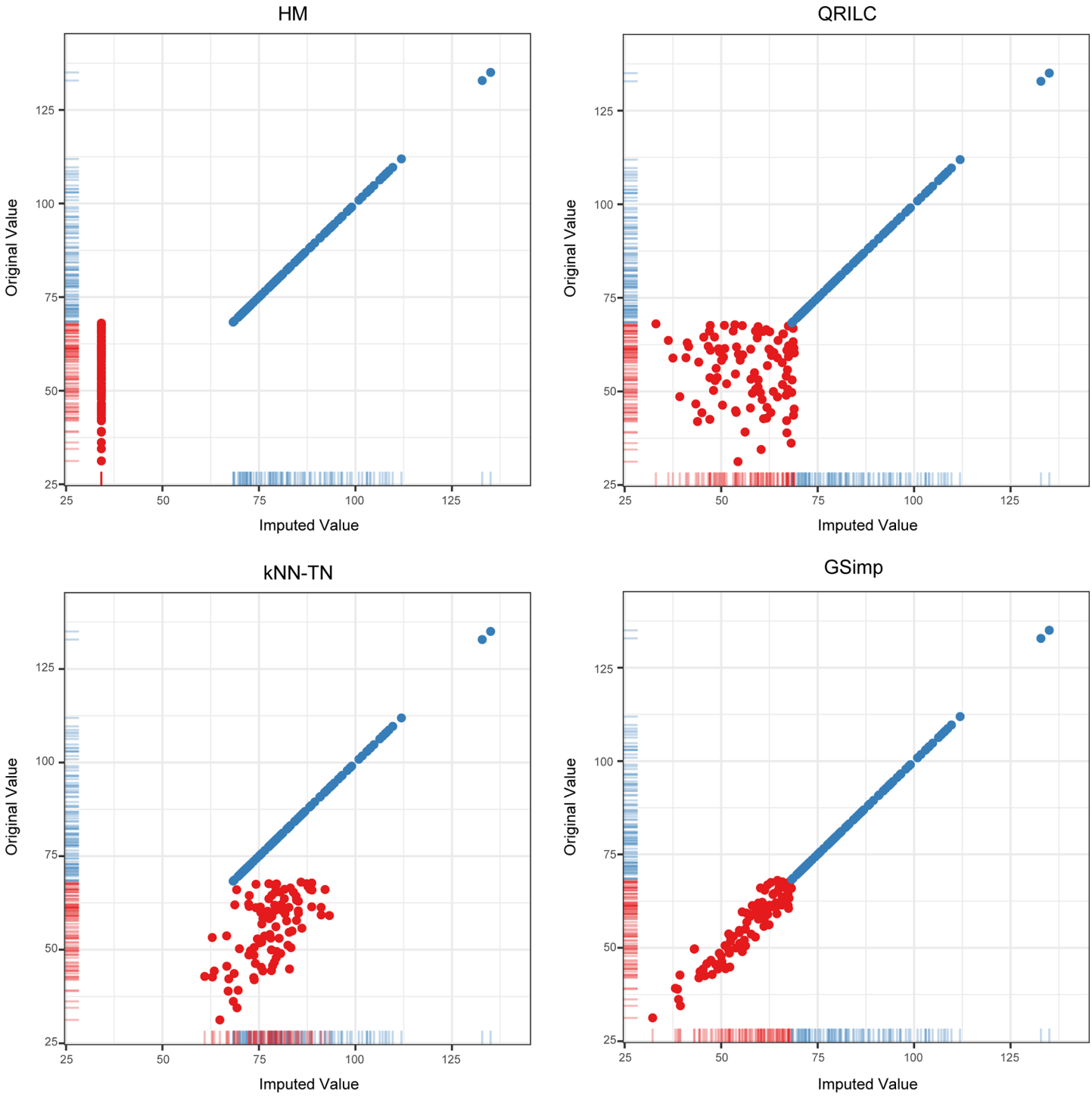

Fig 5. Comparisons of imputed values and original values on one variable. Scatter plots of imputed values (X-axis) and original values (Y-axis) on one example missing variable while non-missing elements represented as blue dots and missing elements as red dots based on four imputation methods: HM (upper left), QRILC (upper right), kNN-TN (lower left), and GSimp (lower right). Rug plots show the distributions of imputed values and original values.

https://doi.org/10.1371/journal.pcbi.1005973.g005

comparison, QRILC kept the overall data distribution and variances (upper right panel of Fig 5). However, stochastic values would be generated by this approach since QRILC imputes each

missing variable independently without utilizing the predictive information from other variables. Statistical learning based method, kNN-TN, applied a correlation based kNN algorithm with parameters of missing variables estimated with truncated normal distributions. This method utilized the information of highly correlated variables of targeted missing variable, thus kept a linear trend between original values and imputed values. However, since no constraint was applied for the imputation, a right shift of missing part occurs, causing imputed values to exceed the truncation point (lower left panel of Fig 5). In contrast, GSimp utilized the predictive information of other variables by employing a prediction model and held a truncated normal distribution for each missing element simultaneously, which ensured a favorable linear trend between imputed and original values as well as a reasonable bound for the imputed values (lower right panel of Fig 5).

In this study, we comprehensively evaluated our algorithm on targeted metabolomics datasets for the MNAR situation. We additionally tested a non-targeted GC/MS profiling metabolomics dataset and found that most of missing values are manually retrievable due to the miss-identification of peaks. These retrievable missing elements were randomly distributed across the dataset and irrelevant to their true abundances (S1 Appendix). Based on this, we assumed the majority of missing values are MCAR/MAR situation for non-targeted GC/MS data before manually missing retrieval. For the rest un-retrievable missing elements, we found much lower signal to noise ratios which could be assigned to the situation of left-censored MNAR. Thus, for non-targeted profiling datasets, missing retrieval from raw spectral data will be most recommended. Since we applied the minimum observed value of missing variable as an informative upper truncation point and $-\infty$ as a non-informative lower truncation point for left-censored missing, GSimp with this default settings might be applicable for the imputation of post-missing retrieval non-targeted data.

GSimp is more than that, other truncation values could also be applied in real-world analyses, such as known LOQ/LOD of metabolites or quantile of observed values (e.g., 10%) can be set as upper truncation points for different conditions. Additionally, when signal intensity of certain compound is larger than the upper limit of quantification range or saturation during instrument analysis, an informative lower truncation point could be correspondingly applied for the right-censored missing value. What's more, when non-informative bounds for both upper and lower limits (e.g., $+\infty$, $-\infty$) were applied, GSimp could be extended to the situation of MCAR/MAR. With the flexible usage of upper and lower limits, our approach may provide a versatile and powerful imputation technique for different missing types. For other–omics datasets with missing values (especially MNAR) (e.g., single cell RNA-sequencing data), we could also apply this method with few modifications of default settings. Thus, it is worthy to evaluate our approach, GSimp, in other complex scenarios in the future.

Since GSimp employed an iterative Gibbs sampler method, a large number of iterations (*iters_all* = 20, *iters_each* = 100) are preferable for the convergence of parameters in Markov chain Monte Carlo (MCMC) method. However, as we tested on the simulation dataset with different number of iterations, a smaller number of iterations (*iters_all* = 10, *iters_each* = 50) won't severely affect the imputation accuracy (S2 Fig). Among iterations for the whole data matrix, we applied a sequential imputation procedure for missing variables from the least number of missing values to the most. To improve the computational efficiency of GSimp on large scale datasets, we additionally implemented a parallel version which can run Gibbs sampler on multiple missing variables simultaneously, then update all imputed values of missing elements. Increasing the number of cores will significantly decrease the running time of GSimp as we tested on a random generated dataset (S1 Table).

In conclusion, we developed a new imputation approach GSimp that outperformed traditional determined value substitution method (HM) and other approaches (QRILC, and

kNN-TN) for MNAR situations. GSimp utilized predictive information of variables and held a truncated normal distribution for each missing element simultaneously via embedding a prediction model into the Gibbs sampler framework. With proper modifications on the parameter settings (e.g., truncation points, pre-processing, etc.) GSimp may be applicable to handle different types of missing values and in different -omics studies, thus deserved to be further explored in the future.

Materials and methods

Diabetes datasets

We employed datasets from a study of comparing serum metabolites between obese subjects with diabetes mellitus (N = 70) and healthy controls (N = 130) where N represents the number of observations. Dataset 1: a total of 42 free fatty acids (FFAs) were identified and quantified in those participants in order to evaluate their FFA profiles [21]. Dataset 2: a total of 34 bile acids (BAs) were identified and quantified in a similar way using different analytical protocol [22].

Simulation dataset

For the simulation dataset, we first calculated the covariance matrix *Cov* based on the whole diabetes dataset (P = 76) where P represents the number of variables. Then we generated two separated data matrices with the same number of 80 observations from multivariate normal distributions, representing two different biological groups. For each data matrix, the sample mean of each variable was drawn from a normal distribution $N(0, 0.5^2)$ and *Cov* was kept using SVD. Then, two data matrices were horizontally (column-wise) stacked together as a complete data matrix (N×P = 160×76) so that group differences were simulated and covariance was kept.

MNAR generation

For two real-world targeted metabolomics datasets, we generated a series of MNAR datasets by using the missing proportion (number of missing variables/number of total variables) from 0.1 to 0.6 in a step of 0.05 with quantile cut-off for each missing variable drawn from a uniform distribution $U(0.1, 0.5)$. The elements lower than the corresponding cut-off were removed and replaced with NA. For the simulation dataset, we generated a series of MNAR datasets by using the missing proportion from 0.1 to 0.8 step by 0.1 with MNAR cut-off drawn from $U(0.3, 0.6)$ for a more rigorous testing.

Prediction model

A prediction model was employed for the prediction of missing values by setting a targeted missing variable as outcome and other variables as predictors. Different prediction models (e.g., linear regression, elastic net [23], regression trees [24] and random forest [25], etc.) could be embedded in our imputation framework. Elastic net was applied in our approach as an ideal prediction model considering its stability, accuracy, and efficiency. This model is a regularized regression with the combination of L1 and L2 penalties of the LASSO [26] and ridge [27] methods. The estimates of regression coefficients in elastic net are defined as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda[(1-\alpha)/2\|\beta\|_2^2 + \alpha\|\beta\|_1]) \quad (1)$$

The L2 penalty $(1-\alpha)/2\|\beta\|_2^2$ improves the model's robustness by controlling the multicollinearities among variables which are widely existed in high-dimensional–omics data. And the

L1 penalty $\alpha\|\beta\|_1$ controls the number of predictors by assigning zero coefficients to the "unnecessary" predictors. From a Bayesian point of view, the regularization is a mixture of Gaussian and Laplacian prior distributions of coefficients which can pull the full model of maximum likelihood estimates $\underset{\beta}{\operatorname{argmin}} \|y - X\beta\|^2$ towards the null model of prior coefficients distribution, thus controls the risk of overfitting and increase the model robustness. R package *glmnet* was used for the elastic net. We set hyperparameters λ as 0.01 (default setting for high-dimensional data) and α as 0.5 (an equally mixture of LASSO and ridge penalties) [28].

Gibbs sampler

Gibbs sampler is a MCMC technique that sequentially updates parameters while others are fixed. It can be used to generate posterior samples. For each missing variable in the dataset, we applied a Gibbs sampler to impute the missing values by sampling from a truncated normal distribution with prediction model fitted value as mean and root mean square deviation (RMSD) of missing part as standard deviation while truncated by specified cut-points. Assuming we have a $n \times p$ data matrix $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \ldots, \boldsymbol{X}_p)$ with only one variable $\boldsymbol{X}_j$ containing left-censored missing values. We denote $\boldsymbol{X}_j$ as $\boldsymbol{y}$ and the missing part as $\boldsymbol{y}_m$ with length m and non-missing part as $\boldsymbol{y}_f$ with length f, and the rest of matrix $\boldsymbol{X}_{-j}$ as $\boldsymbol{X}$'. We can then set the lower truncation point lo as $-\infty$ (centralized data) or 0 (original data) and upper hi as the minimum/quantile value of $\boldsymbol{y}_f$ or a given LOQ. The truncation bounds ensure imputation results are constrained within $[lo, hi]$. Then, the Gibbs sampler approach can be described as following steps:

Step-1 (initialization): we initialize missing values (QRILC in our case), and get $\boldsymbol{y}$';

Step-2 (prediction): we then build a prediction model (elastic net in our case): $\boldsymbol{y}' \sim \boldsymbol{X}'$;

Step-3 (estimation): based on the prediction model, we get the predicted value $\hat{\boldsymbol{y}}$ and the root mean square deviation (RMSD) of missing part $\sigma = \sqrt{\frac{\sum_{i=1}^{m} (\hat{y}_{m_i} - y'_{m_i})^2}{m}}$ where y'_{m_i} and $\hat{y}_{m_i}$ are ith initialized/imputed value and fitted value respectively;

Step-4 (sampling): we draw sample $\tilde{y}_{m_i}$ from a truncated normal distribution $N(\hat{y}_{m_i}, \sigma^2 | [lo, hi])$ for ith missing element and update $\boldsymbol{y}$'.

We iteratively repeat step-2 to step-4 and update $\boldsymbol{X}_j$.

GSimp framework

A whole data matrix $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \ldots, \boldsymbol{X}_p)$ contains a number of k ($k \leq p$) left-censored missing variables. We present our imputation framework as following algorithm.

Algorithm: Gibbs sampler based left-censored missing value imputation approach

Require: $\boldsymbol{X}$ an $n \times p$ data matrix, *iters_all* the number of iterations for imputing the whole matrix $\boldsymbol{X}$, *iters_each* the number of iterations for imputing each missing variable, a vector of upper limits $\boldsymbol{U}$ ($+\infty$ for non-missing variables) with length p and a vector of lower limits $\boldsymbol{L}$ ($-\infty$ for non-missing variables) with length p.

```
1. X^imp ← initialize the missing values for X;
2. K ← vector of indices of missing variables in X with increasing
amount of missing values;
3. for 1:iters_all do
4.     for j in K do
5.         y' ← X_j^imp, y' can be divided into two parts: y'_m is a vector
of the imputed part (original missing part) with length m and y'_f is a
vector of the non-missing part with length f while n = m + f;
```

```
6.        X' ← X_{-j}^{imp}, represents the matrix X with jth column removed;
7.        lo ← L_j and hi← U_j;
8.        for 1:iters_each do
9.           Gibbs sampler step 2 to 4;
10.        end for
11.        Update X_j^{imp};
12.       end for
13. end for
14.return X^{imp}
```

Other imputation approaches

Other three left-censored missing imputation/substitution methods were conducted in our study for performance comparison:

- kNN-TN (Truncation *k*-nearest neighbors imputation) [19]: this method applied 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.
- QRILC (Quantile Regression Imputation of Left-Censored data) [18,29]: this method imputes missing elements randomly drawing from a truncated distribution estimated by a quantile regression. R package *imputeLCMD* was applied for this imputation approach.
- HM (Half of the Minimum): This method replaces missing elements with half of the minimum of non-missing elements in the corresponding variable.

Assessments of performance

Normalized Root Mean Squared Error (NRMSE) [30] has been commonly used to evaluated the differences between true values and imputed values. Considering the skewed distribution of missing values in MNAR, NRMSE based sum of ranks (SOR) was derived, a robust non-parametric measurement, to compare different imputation methods. The formula is as follows [31]:

$$SOR = \sum_{i=1}^{k} Rank_i(NRMSE) \quad (2)$$

where $Rank_i(NRMSE)$ represent the NRMSE ranks of different imputation methods in i_{th} missing variable.

Procrustes analysis, a statistical shape analysis, could be used to evaluate the similarity of two ordinations by calculating the sum of squared errors [32]. We applied principal component analysis (PCA) as the unsupervised (un-labeled) ordination measurement and Procrustes analysis to measure the alteration of the original sample distribution and the imputed sample distribution in the space of top PCs. R package *vegan* was applied for Procrustes analysis [33].

Labeled measurements include correlation analysis for log-transformed *p*-values between true data and imputed data from Student's *t*-test, partial least square (PLS)- Procrustes analysis that measures the differences between original and imputed sample distributions on top PCs using supervised PLS for the dimensional reduction. R package *ropls* was applied for PLS analysis [34]. These measurements were done using our imputation evaluation pipeline from our previous study [31], which is also accessible through: https://github.com/WandeRum/MVI-evaluation.

Furthermore, we evaluated the impacts of different imputation methods on the statistical sensitivity of detecting biological variances. On the simulation dataset, we calculated *p*-values

from student's *t*-tests between two groups from original and imputed datasets. We marked a set *S* as real differential variables at a significant level of *p*-cutoff (e.g., 0.05) from original simulation data, and a set *S'* as detected differential variables at the same significant level from imputed simulation data. Then we calculated the true positive rate $TPR = \frac{\#of\ (S \cap S')}{\#of\ S}$ to evaluate the effects of different imputation methods in terms of detecting differential variables.

Supporting information

S1 Appendix. A step by step tutorial of GSimp.
(PDF)

S1 Fig. Evaluations of different imputation methods on simulation dataset. SOR (upper left), PCA-Procrustes sum of squared errors (upper right), Pearson's correlation between log-transformed *p*-values of student's t-tests (lower left), and PLS-Procrustes sum of squared errors (lower right) on simulation dataset along with different numbers of missing variables based on three imputation methods: QRILC (green triangle), GSimp (blue square), and kNN-TN (purple cross).
(TIF)

S2 Fig. Evaluations of different numbers of iterations using GSimp on simulation dataset. SOR on simulation dataset along with different numbers of missing variables based on four different numbers of iterations: *iters_each* = 50 and *iters_all* = 20 (red circle), *iters_each* = 100 and *iters_all* = 20 (green triangle), *iters_each* = 50 and *iters_all* = 10 (blue square), *iters_each* = 100 and *iters_all* = 10 (purple cross).
(TIF)

S1 Table. Table of computational efficiency of GSimp on a 200×200 random generated large dataset. # missing variables: number of missing variables; iters_each: number of iterations for imputing each missing variable; iters_all: number of iterations for imputing the whole matrix; n_cores: number of cores.
(XLSX)

Acknowledgments

R. Wei and J. Wang would like to thank their parents (B. Wei, X. He, K. Wang, and Q. Peng) for their endless love and support. They are also grateful to Mr. Link who is always curious about the unexplored land.

Author Contributions

Conceptualization: Runmin Wei, Jingye Wang.

Data curation: Runmin Wei, Jingye Wang, Erik Jia, Tianlu Chen.

Formal analysis: Runmin Wei, Jingye Wang.

Funding acquisition: Tianlu Chen, Wei Jia.

Methodology: Runmin Wei.

Resources: Wei Jia.

Software: Runmin Wei, Jingye Wang, Erik Jia.

Supervision: Jingye Wang, Yan Ni, Wei Jia.

Writing – original draft: Runmin Wei, Jingye Wang.

Writing – review & editing: Runmin Wei, Jingye Wang, Erik Jia, Tianlu Chen, Yan Ni, Wei Jia.

References

1. Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models [Internet]. Cambridge University Press. 2006. https://doi.org/10.2277/0521867061
2. Little RJ a, Rubin DB. Statistical Analysis with Missing Data. Statistical analysis with missing data Second edition. 2002. https://doi.org/10.2307/1533221
3. Hrydziuszko O, Viant MR. Missing values in mass spectrometry based metabolomics: An undervalued step in the data processing pipeline. Metabolomics. 2012; 8: 161–174. https://doi.org/10.1007/s11306-011-0366-4
4. Guo L, Milburn M V, Ryals JA, Lonergan SC, Mitchell MW, Wulff JE, et al. Plasma metabolomic profiles enhance precision medicine for volunteers of normal health. Proc Natl Acad Sci. 2015; 112: E4901–E4910. https://doi.org/10.1073/pnas.1508425112 PMID: 26283345
5. Liu J-J, Ghosh S, Kovalik J-P, Ching J, Choi HW, Tavintharan S, et al. Profiling of plasma metabolites suggests altered mitochondrial fuel usage and remodelling of sphingolipid metabolism in individuals with type 2 diabetes and kidney disease. Kidney Int Reports. 2016; 2: 470–480. https://doi.org/10.1016/j.ekir.2016.12.003 PMID: 29142974
6. Butte NF, Liu Y, Zakeri IF, Mohney RP, Mehta N, Voruganti VS, et al. Global metabolomic profiling targeting childhood obesity in the Hispanic population. Am J Clin Nutr. 2015; 102: 256–267. https://doi.org/10.3945/ajcn.115.111872 PMID: 26085512
7. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001; 17: 520–525. https://doi.org/10.1093/bioinformatics/17.6.520 PMID: 11395428
8. Hastie T, Tibshirani R, Sherlock G. Imputing missing data for gene expression arrays. Tech Report, Div Biostat Stanford Univ. 1999; 1–9.
9. Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. pcaMethods—A bioconductor package providing PCA methods for incomplete data. Bioinformatics. 2007; 23: 1164–1167. https://doi.org/10.1093/bioinformatics/btm069 PMID: 17344241
10. Stekhoven DJ, Bühlmann P. Missforest-Non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012; 28: 112–118. https://doi.org/10.1093/bioinformatics/btr597 PMID: 22039212
11. Mak TD, Laiakis EC, Goudarzi M, Fornace AJ. MetaboLyzer: A novel statistical workflow for analyzing postprocessed LC-MS metabolomics data. Anal Chem. 2014; 86: 506–513. https://doi.org/10.1021/ac402477z PMID: 24266674
12. Katajamaa M, Miettinen J, Oresic M. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. Bioinformatics. 2006; 22: 634–6. https://doi.org/10.1093/bioinformatics/btk039 PMID: 16403790
13. Kessler N, Neuweger H, Bonte A, Langenkämper G, Niehaus K, Nattkemper TW, et al. MeltDB 2.0-advances of the metabolomics software system. Bioinformatics. 2013; 29: 2452–2459. https://doi.org/10.1093/bioinformatics/btt414 PMID: 23918246
14. Luedemann A, Von Malotky L, Erban A, Kopka J. TagFinder: Preprocessing software for the fingerprinting and the profiling of gas chromatography-mass spectrometry based metabolome analyses. Methods Mol Biol. 2012; 860: 255–286. https://doi.org/10.1007/978-1-61779-594-7_16 PMID: 22351182
15. Xia J, Sinelnikov I V., Han B, Wishart DS. MetaboAnalyst 3.0-making metabolomics more meaningful. Nucleic Acids Res. 2015; 43: W251–W257. https://doi.org/10.1093/nar/gkv380 PMID: 25897128
16. Xia J, Psychogios N, Young N, Wishart DS. MetaboAnalyst: A web server for metabolomic data analysis and interpretation. Nucleic Acids Res. 2009;37. https://doi.org/10.1093/nar/gkp356 PMID: 19429898
17. Xia J, Mandal R, Sinelnikov I V., Broadhurst D, Wishart DS. MetaboAnalyst 2.0-a comprehensive server for metabolomic data analysis. Nucleic Acids Res. 2012;40. https://doi.org/10.1093/nar/gks374 PMID: 22553367
18. Lazar C, Gatto L, Ferro M, Bruley C, Burger T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. J Proteome Res. 2016; 15: 1116–1125. https://doi.org/10.1021/acs.jproteome.5b00981 PMID: 26906401
19. Shah JS, Rai SN, DeFilippis AP, Hill BG, Bhatnagar A, Brock GN. Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical

metabolomics studies. BMC Bioinformatics. 2017; 18: 114. https://doi.org/10.1186/s12859-017-1547-6 PMID: 28219348

20. Gaude E, Chignola F, Spiliotopoulos D, Spitaleri A, Ghitti M, M Garcia-Manteiga J, et al. muma, An R Package for Metabolomics Univariate and Multivariate Statistical Analysis. Curr Metabolomics. 2013; 1: 180–189. https://doi.org/10.2174/2213235X11301020005
21. Ni Y, Zhao L, Yu H, Ma X, Bao Y, Rajani C, et al. Circulating Unsaturated Fatty Acids Delineate the Metabolic Status of Obese Individuals. EBioMedicine. 2015; 2: 1513–1522. https://doi.org/10.1016/j.ebiom.2015.09.004 PMID: 26629547
22. Lei S, Huang F, Zhao A, Chen T, Chen W, Xie G, et al. The ratio of dihomo-γ-linolenic acid to deoxycholic acid species is a potential biomarker for the metabolic abnormalities in obesity. FASEB J. Federation of American Societies for Experimental Biology; 2017; fj.201700055R. https://doi.org/10.1096/fj.201700055R PMID: 28490483
23. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005; 67: 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
24. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. The Wadsworth statisticsprobability series. 1984. https://doi.org/10.1371/journal.pone.0015807
25. Breiman L. Random forests. Mach Learn. 2001; 45: 5–32. https://doi.org/10.1023/A:1010933404324
26. Tibshirani R. Regression Selection and Shrinkage via the Lasso [Internet]. Journal of the Royal Statistical Society B. 1996. pp. 267–288. https://doi.org/10.2307/2346178
27. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. 1970; 12: 55–67. https://doi.org/10.1080/00401706.1970.10488634
28. Friedman AJ, Hastie T, Simon N, Tibshirani R, Hastie MT. Lasso and Elastic-Net Regularized Generalized Linear Models [Internet]. 2015. Available: http://www.jstatsoft.org/v33/i01/.
29. Lazar C. Imputation of left-censored missing data using QRILC method [Internet]. 2015.
30. Oba S, Sato M -a., Takemasa I, Monden M, Matsubara K -i., Ishii S. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics. 2003; 19: 2088–2096. https://doi.org/10.1093/bioinformatics/btg287 PMID: 14594714
31. Wei R, Wang J, Su M, Jia E, Chen S, Chen T, et al. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. Sci Rep. 2018; 8: 663. https://doi.org/10.1038/s41598-017-19120-0 PMID: 29330539
32. Dryden IL, Mardia K V. Statistical Shape Analysis. J Hum Evol. 1998; 4: 376. https://doi.org/10.1006/jhev.1999.0391
33. Oksanen J. Multivariate Analysis of Ecological Communities in R: vegan tutorial [Internet]. 2015.
34. Thévenot EA, Roux A, Xu Y, Ezan E, Junot C. Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive Workflow for Univariate and OPLS Statistical Analyses. J Proteome Res. 2015; 14: 3322–3335. https://doi.org/10.1021/acs.jproteome.5b00354 PMID: 26088811