PhD Thesis
Zhan-long Mei

An Expert Analysis System for Metabolomics
Data Analysis

Supervisor: Karsten Kristiansen & Siqi Liu
Submitted on 7th April 2020
Name of department: Department of Biology

Author(s): Zhan-long Mei

Title and subtitle: An expert analysis system for metabolomics data analysis

Topic description: Reviewing the challenge of untargeted metabolomics data analysis; evaluating multiple approaches at each stage of data pre-processing and pre-treatment; novel missing value imputation strategy; developing an integrate pipeline for untargeted metabolomics data analysis.

Supervisor: Karsten Kristiansen & Siqi Liu

Submitted on: 7th April 2020

Grade: [X]

Number of study units:

☐ 2  □ 3
PREFACE

This Ph.D. project started in 2017 as part of the collaboration between the Department of Biology, University of Copenhagen, and BGI Shenzhen. The work presented here was performed at both institutions under the supervision of Prof. Karsten Kristiansen and Prof. Siqi Liu.
ACKNOWLEDGEMENTS

I want to thank my supervisors, Prof. Karsten Kristiansen and Prof. Siqi Liu, for guiding me to explore the fantastic research area of metabolomics data analysis, for providing me absolute support during my Ph.D. study. You always keep your enthusiastic in research, and I do learn a lot from your research spirit. I sincerely thank both of you for all the help and suggestions.

I also express gratitude to my colleges in BGI-Shenzhen, especially JinZi, for giving me suggestions and encouragements during my study, Minling Liang and Haiyu Weng, for helping me to test my new ideas.

Finally, I would like to thank my family and friends, for giving me motivation and support to finish my research, especially my mother for taking good care of the whole family.
ABSTRACT

With the advancement of the MS techniques, the throughput and coverage of the untargeted metabolomics have been greatly improved, making it a powerful tool for screening altered metabolites associated with phenotypes or simulations. Tens of thousands of metabolites could be detected in one run and quantitively measured as the peak areas of the MS features. Due to the nature of the quantitation, the precision of the quantitation is affected by a variety of factors, such as the batch effects, the ionization efficiency, and the dilution effects. To combat these unwanted variations, several data processing and pretreatment steps are needed, and multiple algorithms have been developed. However, until now, there was no widely accepted workflow for the untargeted metabolomics data analysis. On the one hand, limited options of data analysis algorithms were implemented in the integrated pipelines without systematic evaluation of their performance. On the other hand, the users were encouraged to try different data processing algorithms to choose the best one.

In this study, I developed an expert analysis system, MetaboPro, for untargeted metabolomics data, in which multiple approaches were implemented, and systematic evaluations were provided. I showed that there is no single solution to the missing value imputation, batch effects removal, sample normalization, transformation, and scaling, because the performances of different approaches for these analyses varied a lot, and no approaches always outstand the other on different datasets. I reviewed and compared the existing evaluation criteria for each step of analysis and implemented the commonly used data analysis approaches and evaluation criteria into MetaboPro. Besides, I also developed a stepwise imputation strategy by classifying the missing values into three classes and imputing according to their origin, which greatly improved the imputation accuracy.

This expert analysis system may serve the community in at least three ways: firstly, MetaboPro provides guidance of the necessity of each data processing and how to evaluate the processing outcome step by step, which benefits new users to better understand the data analysis. Secondly, this integrated tool greatly improves the robustness of the statistical outcome, leading to precise interpreting of the phenotypes. Finally, this study will advance the development of untargeted metabolomics data analysis and speed up the formation of a widely accepted workflow.
Table of contents

ABBREVIATIONS ................................................................................................. 8

1.0 INTRODUCTION .............................................................................................. 9

1.1 Challenge of pre-processing in metabolomics analysis ........................................ 13
  1.1.1 Missing value .............................................................................................. 13
  1.1.2 Batch effects .............................................................................................. 15
  1.1.3 Sample normalization ................................................................................. 18

1.2 Challenge of data pre-treatment ....................................................................... 19

1.3 Challenge of statistical analysis and differential feature selection ........................ 23
  1.3.1 Univariate analysis approaches ................................................................. 24
  1.3.2 Multivariate analysis approaches ............................................................... 25

1.4 Objectives ........................................................................................................ 26

2. LIST OF ARTICLES RELATED TO THE TOPIC OF THIS THESIS ................... 27

LIST OF ARTICLES NOT RELATED TO THE TOPIC OF THIS THESIS AND
COMPLETED DURING THE PHD CAREER ................................................................ 28

3. RESULTS ........................................................................................................... 29

Part one. The establishment of an integrated untargeted metabolomics pipeline ............ 29

Result summary of Article 1: .................................................................................. 29

Part two. Novel strategy for missing values imputation and evaluation to imputation results ............................................................... 30

Result summary of Article 2: .................................................................................. 30

Part three. Evaluation to batch effects removal ................................................................ 32

Result summary of Article 3: .................................................................................. 32
Part four. Expert analysis system for metabolomics datasets

Result summary of Article 4:

4. DISCUSSION AND PERSPECTIVE

Missing value imputation:

Batch effects removal:

Transformation and scaling:

Development of integrated pipelines:

CONCLUSIONS

REFERENCES

APPENDIX
ABBREVIATIONS

NMR: nuclear magnetic resonance
LC-MS: liquid chromatography–mass spectrometry
GC-MS: gas chromatography–mass spectrometry
TM: truly missing
LIM: low intensity induced missing
UM: unclear missing
MCAR: missing completely at random
MAR: missing at random
MNAR: missing not at random
kNN: k-nearest neighbors
MF: missforest
RMSE: root mean square errors
PCA: principal component analysis
PLS-DA: partial least squares discriminant analysis
PPCA: probabilistic PCA
SVD: singular value decomposition
cZero: values close to zero
IA: imputation accuracy
DAF: differential abundance features
QC: quality control
QCRSC: quality control-robust spline correction
QCSVR: quality control-support vector regression
PQN: probabilistic quotient normalization
VSN: variance stabilization normalization
SVA: Surrogate Variable Analysis
1.0 INTRODUCTION

Untargeted metabolomics is a very powerful tool for screening and discovering biological meaningful metabolites responding to experimental factors \(^1\). With the advances of technical platforms, databases and statistical approaches, untargeted metabolomics has been widely applied in various domains, including mammals, plants, microbes, and environmental systems, especially in the studies of human diseases, biomarker screening, and association between the environment and genotypes \(^2\). Combined with other Omics, such as genomics and proteomics, metabolomics can help us better understand biological processes.

There are three main platforms for untargeted metabolomics studies: nuclear magnetic resonance (NMR), liquid chromatography–mass spectrometry (LC-MS), and gas chromatography–mass spectrometry (GC-MS). In NMR, a metabolite is represented by the chemical shifts and the signal intensities, while in LCMS and GCMS, a metabolite is represented by three descriptors, the retention time, the mass to charge ratio, and the signal intensities. In these platforms, the quantitative measurement of the metabolites’ intensities is achieved by calculating the peak area of the corresponding MS features, which is supposed to be positively correlated to the metabolite concentration. The main advantage of the metabolomics profiling is high throughput and coverage. Tens of thousands of metabolites can be quantitatively measured in one LC-MS run, which greatly improves the power and applicability of untargeted metabolomics profiling. However, the increase in data size and the confounding noise and variations also present a significant challenge to the statistical analysis\(^3\).

Since the quantitative measurement of the MS feature is solely represented by the peak area in the untargeted metabolomics studies, the MS signals are inevitably subjected to a variety of noise and variances, such as severe missing values, the decrease of responding rates due to long acquisition time, and the ionization difference among metabolites. According to De Livera et al. \(^4\), the observed MS signals in untargeted metabolomics datasets are a combination of the biological variations relating to the study design, the systematic experimental variations, the unwanted biological variation and the random noises. To obtain the clean data, it is necessary to perform the data preprocessing to remove the variations not related to the true biological difference, such as the batch effects and the concentration differences among samples (Figure 1A). It is merit to note that the clean data cannot be
directly applied for downstream analysis because of the scale difference in signal intensities among the MS features and the biased data distribution, which lower the power of statistical analysis. A data pretreatment step is needed to bring all the MS features to similar weights and adjust the left-censored distribution to the normal distribution, which fits the underline hypothesis of the most statistical model (Figure 1B). Lastly, the statistical analysis is applied to identify the differential abundant MS features (DAF) or to develop a model to precisely predict the phenotype of the new samples. For each statistical approach, there are some assumptions of the datasets, such as the normality and the homogeneity. If these assumptions are not met, the power of the statistical analysis decreases and the risk of false discovery increases. Thus, it is important to check the characteristic of the datasets and apply the appropriate statistical approaches to achieve reliable conclusions.

![Diagram](image.jpg)

**Figure 1.** the unwanted variations and noises in metabolomics datasets. (A) the subtypes of unwanted variations. (B) the scale difference and biased distribution of MS features.

To address these unwanted variations and precisely identify the metabolites responding to the study design, four main steps, including missing value imputation, data processing, data pretreatment and statistical analysis, are necessary for untargeted metabolomics data analysis. As shown in Figure 2 the four main steps can be further divided into several sub-steps, and in each sub-step, there are many approaches available and several factors have to be considered. In most cases, the selection of the approaches could have a significant influence on the statistical outcome, and the inappropriate implementation of data processing, pretreatment, and statistical analysis may lead to biased conclusions. For example, Oh et al. claimed that the screening of the DAFs is sensitive to the selection of the missing value imputation methods. Gromski, et al. concluded that both the missing values imputation strategies and the scaling approaches have a significant impact on the classification accuracy of the supervised multivariate models. Due to the difference of various datasets and the diversity of the underline theories behind the algorithms of statistical approaches, there is no single
solution to each step of the data analysis. For instance, Webb-Robertson et al. 9 evaluated ten imputation approaches on three LCMS datasets and concluded that there is no single solution to the missing values imputation. Li et al. 10 claimed that the sample normalization performance depended heavily on the nature of metabolomics datasets and summarized five criteria for the normalization performance evaluation. Vinaixa 11 reviewed the existing univariate analysis approaches and proposed a framework listing the essential considerations for selection of the univariate analysis appropriate. In summary, in each step of the untargeted metabolomics data analysis, there is no widely accepted processing approach. The optimal handling should be determined according to the intrinsic data characteristics, the assumptions of statistical approaches and systematic evaluations are needed to ensure the robustness of each step of the analysis.

Multiple steps and systematic evaluations are required in untargeted metabolomics data analysis to ensure robust conclusions. It is impractical for common users to review all the approaches and assess the performances all by themselves, and the integrated pipeline is needed to facilitate untargeted metabolomics data analysis. Currently, several integrated pipelines have been developed for metabolomics analysis, and most of the pipelines have implemented the data processing, pretreatment and the statistical analysis steps (Table 1). However, there are two limitations of these pipelines. Firstly, in most pipelines, limited number of data preprocessing, pretreatment and statistical analysis approaches were included. For example, in MetFlow 12, only Mean, Median and SUM 13 are available.
for sample normalization, while the other commonly used approaches are not included. Secondly, no or limited evaluations were provided to guide users to select the best approach. For example, in MetaboAnalyst, there are seven normalization approaches, four scaling, and two transformations available. It is very difficult for common users to determine which combination is the optimal one. What’s more, after implementing these normalization and scaling approaches, there is no effective way for the users to validate that the unwanted biological difference among the samples and the scale difference among the MS features have been correctly and properly removed.

Table 1. Qualitative assessment of existing integrated metabolomics pipelines

<table>
<thead>
<tr>
<th>Integrated pipelines</th>
<th>Metabo Analyst</th>
<th>MSPrep</th>
<th>Galax-M</th>
<th>Metabo Nexus</th>
<th>MetMS Line</th>
<th>Metabox</th>
<th>Stat Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing value imputation</td>
<td>Min</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>kNN</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>MF</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>BPCA</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>PPCA</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Mindet</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Batch effects removal</td>
<td>Combat</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>QCRSC</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>QCSVR</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Limma</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>SVA</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Batchratio</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>EigenMS</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Ber</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Sample normalization</td>
<td>PQN</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>SUM</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>VSN</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Quantile</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Transformation and scaling</td>
<td>Log</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Cubic</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Auto</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Pareto</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Level</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Vast</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>
With the existence of the noise signals and the unwanted variances, it is vital to apply appropriate data processing approaches to the complex metabolomics data to extract meaningful results. Unfortunately, there are no widely accepted standard pipelines for analyzing these complex and high-dimensional data. Thus, in the following part, I review the challenges of untargeted metabolomics data analysis, the available data processing, pre-treatment, and statistical analysis approaches. Moreover, I discuss the considerations of the selection of the analysis approaches and offer guidance on how to make decisions in each step of the analysis.

1.1 Challenge of pre-processing in metabolomics analysis

1.1.1 Missing value

1.1.1.1 The origin and impact of missing values

Missing values are non-detects of MS signals in the samples, often marked as “NA”. It is observed in almost all kinds of datasets, as well as in omics datasets. It is estimated that the metabolomics datasets containing more missing values than other omics datasets, occupying as much as 20% in total and affecting 50-80% of MS features. Generally, there are three main sources of missing values: (1) the concentrations of some metabolites are relatively low, resulting in MS signals lower than the detection limits. (2) Some metabolites may only exist in one biological group due to the different stimulations. For example, the drug compounds may only be found in the treated group of mice, not in the healthy control group. (3) Some random errors, such as iron suppression, or failures in peak picking may also introduce missing values. According to Robbins, the pattern of missing values could be missing not at random (MNAR) if the missing values are caused by low intensity while the random errors caused missing could be missing at random (MAR) or missing completely at random (MCAR). The existence of missing values hinders the application of most statistical approaches, lower the power of statistical outcome and may even lead to biased conclusions if not properly processed.

1.1.1.2 The possible ways to deal with missing values

Currently, a variety of approaches have been proposed for missing value imputation. These imputation strategies can generally be grouped into five classes: (1) simple imputation with a constant
value, such as the mean, median, or minimum value imputation; these imputations are criticized for biasing the true means and standard deviations \(^{17}\). (2) Local structure-based imputation, including k-nearest neighbors (kNN), local least square (LLS) \(^{18}\), which estimates the missing values through linear regression based on a subset of the most correlated MS features. (3) Global structure-based imputation, such as Probabilistic principal component analysis (PPCA) \(^{19}\) and Bayesian principal component analysis (BPCA) \(^{19}\), which applies dimension reduction analysis to the incomplete dataset with missing values and the missing values are iteratively reconstructed through the expectation maximization algorithm. (4) Machine learning-based imputation, such as MissForest (MF) \(^{20}\). In MF, each MS feature in the dataset is set as the outcome and a random forest is fitted to it. The missing values in the MS feature are then predicted using the random forest iteratively. These methods are often computationally intensive, which hinder their application to large datasets. (5) Left-censored imputation methods, including Mindet \(^{21}\) and QRLIC \(^{22}\). These methods assume that the missing values in the metabolomic dataset are caused by low intensity and thus should be imputed with values lower than the detection limit. Since the underline theories behind each imputation strategies vary significantly, and different metabolomic datasets have their specific missing pattern, different imputation approaches may generate inconsistent imputation results. It is important to distinguish the best imputation algorithm for a specific metabolomic dataset.

1.1.1.3 Limitations of the current missing value imputation approaches
As I have discussed above, the five classes of imputation approaches have diverse assumptions of the missing patterns and varied imputation strategies. In untargeted metabolomics datasets, the missing pattern is complex, with at least three sources of missing values. An imputation without the consideration of the origin of the missing values may lead to a poor result. For instance, the kNN is reported to work well on the MCAR/MAR missing patterns, but poorly on the MNAR \(^{25}\). However, most of those imputations, except the Mindet and the QRLIC, are not designed for the metabolomics datasets, which do not take the missing patterns of metabolomics datasets into consideration. The Mindet and QRLIC assume that the missing values are mainly caused by the low intensities and impute the missing values with small values, which may be oversimplified as there are also random
error induced missing. Hence, imputation the missing values according to their origins may be a promising area and better fit the untargeted metabolomics datasets.

1.1.1.4 The evaluation of missing value processing
Evaluation of the imputation performance has always been a challenge since the true values behind the missing values are often unknown. An alternative way is to compare the performance of missing value imputation on simulated datasets, in which the missing values are artificially introduced into a complete dataset, and the imputation performances are determined by comparing the imputed results to the complete dataset. Root mean square error (RMSE) is a commonly used measurement of the similarity between the imputed and original datasets. One concern about the simulation is that the patterns of the artificially introduced missing values cannot always well represent the patterns in the real-life datasets, limiting the applicability of the conclusions drawn from RMSE. The classification accuracy, which measures how imputation improves the ability of the supervised model to predict the biological groups of the new samples, was proposed to assess the imputation performance on real datasets. This evaluation is based on the hypothesis that if the missing values are well imputed, it could reduce the noise signals and better recover the real biological difference, resulting in better classification accuracy. However, this measurement is limited by the low sensitivity as the classification accuracy is an indirect measurement to the imputation performance and is influenced by a variety of factors, such as the true biological difference among the biological groups. Gromski et al. calculated the classification accuracies of six imputed results, which is indistinguishable among different imputation methods. Hence, new evaluation criterion is needed to aid the selection of best imputation strategy on real datasets.

1.1.2 Batch effects
1.1.2.1 The origin and classification of batch effects
In a large-scale metabolomics study, the samples are often collected and analyzed on several different batches, resulting in variations in MS signals relating to the experimental operators, reagent batches, and injection orders. The batch effects are largely composed of two parts: the between batch effect and the within batch effect or injection order variation. The between batch effects are systematic
differences caused by the independent acquisition of samples in various batches, while the within batch effects are often observed as the changes in the MS signals response over injection order in the same batch. The statistical analysis without proper management of batch effect may lead to biased conclusion, one critical step of the metabolomic data analysis is to remove these batch effects without distributing the real biological variance.

1.1.2.2 The approaches for batch effects correction

To address these batch effects, several approaches have been developed. One of the commonly used approach is to include internal standards, in which the batch effects could be estimated and removed by calibrating to the signals of these internal standards 29. However, it is difficult to select a group of internal standards that can fit all metabolites. The sample wise normalization methods, such as the SUM, PQN 30, are also suggested for batch correction, and the batch effects are removed by bringing the total intensities of all the samples to a similar level. Quality control (QC) samples-based approaches are also widely adopted, the QC-RLSC 2, and QC-SVR 31 are two representative approaches. These methods model the variations of the MS features in the QC samples as batch effects and extract them from the test samples. There are also some correction methods initially used in genomics data work well on metabolomics data, such as Combat 32, Limma 33, Surrogate Variable Analysis (SVA) 34 and Ber 35. Most of them use two-way ANOVA, while Combat applies an empirical Bayes method to calibrate MS signals into a linear model that contains both the biological and batch covariates. Due to the variance of batch effects in different datasets and the theories behind the correction strategies varied to each other, the performance of the batch effects removal approaches varied a lot on different datasets. It is important to select the optimal batch correction approach to ensure the robustness of the statistical analysis.

1.1.2.3. The evaluation to batch effects correction

Currently, several evaluation strategies have been proposed to assess the batch effects. The first strategy is to use the QC samples. Since the QC samples are technical replicates acquired at a consistent interval during the MS analysis, the variation in the QC samples could be used to model both the between and within batch effects and the evaluation is based on the reduction that the better the batch effects are removed, the higher the consistency among the QC samples would be. Wang et
al. proposed three measurements based on QC samples: 1) the reduction of the relative standard deviation of the MS features in QC samples, 2) the correlation coefficients among the QC samples and 3) the size of the cluster of the QC samples on the PCA plot. To better assess the removal of the within batch effects, Wang et al. and Sanchez-Illana et al. proposed the measurement of the correlation between the injection orders and the projection of the QC samples on PC1. The measurement is successful on some datasets but may fail if the batch effects are not mainly explained on PC1. The second strategy for batch effects evaluation is to use real samples, and the hypothesis is that samples from the same biological group should not be significantly different since these samples are randomly assigned into different batches. Wehrens et al. developed two measurements based on real samples: 1) the overlap of the samples from different batches on PCA score plot and 2) the reducing of the within-group variance after batch effects correction. The fundamental question is that whether the QC sample-based and the real sample-based evaluations are independent? Do both of them are sufficient for the batch effects assessment, or a combination of the two strategies is necessary?

1.1.2.4. Limitations of the batch effects evaluation and available tools

Although a variety of measurements for checking the existence of batch effects have been proposed, it is unclear if these evaluations end up with a similar conclusion. Besides, most of the proposed measurements focus on the between batch effects, while not the within batch effects removal. The reported measurement for within batch effects, which estimates the correlation between the injection order and PC1 scores, suffers from a high risk of failure if the within batch effects are not represented by the PC1. Hence, it is necessary to establish systematic evaluation criteria by comparing the existing between batch effects measurement and develop robust measurements for the within batch assessment.

In most widely used integrated pipelines, the option for the batch correction in these pipelines is often limited without systematic performance evaluation. To overcome this problem, Manimaran developed BatchQC, in which the Combat and SVA are implemented, while the commonly used methods such as QCRSC, Limma, and Ber are not included, and the performance is solely evaluated by the between batch effects removal. Above all, comprehensive evaluation criteria and an easy to use tool are needed to aid the selection of batch effects removal approaches.
1.1.3 Sample normalization

1.1.3.1 The necessity of sample normalization

In metabolomics studies, one of the main goals of statistical analysis is to compare the samples from different groups and find out the differential MS features. An important assumption for metabolomic data analysis is that all samples are comparable with each other. In other words, the total amount of the metabolites should be almost the same. To ensure this premise, equal sample volume or weight is used. Unfortunately, the sample concentration varied a lot among samples, making the total amount of metabolites incomparable despite the insurances of the same volume/weight. For example, the concentration difference between two urine samples could be as large as 14 folds \(^{13}\), leading to diverse total metabolites in the same volume of urine. These biological variations, which are not associated with the study design, may confound the real biological variances and lower the statistical power to identify the important metabolites related to the experimental factors. For instance, if the intensity of a metabolite in sample A is two times higher than that in sample B, it is hard to tell whether this difference is caused by the study design or by the concentration difference between the two samples. Thus, a process to equalize the total metabolites in the samples is needed, termed as sample normalization \(^{40}\).

1.1.3.2 The strategies for sample normalization

To improve the comparability among samples, two strategies, namely experimental-driven and data-driven correction, have been proposed \(^{41}\). The experimental-driven approaches try to adjust the concentration of samples according to specific measurements, such as the absorbency of the samples or the creatinine in the urine samples \(^{13}\). However, none of these measurements is the direct measurement of the total metabolites, and a data-driven approach is suggested to be applied after the experimental-driven approach. The commonly used data-driven approaches include the SUM, Mean, Median \(^{36}\), Probabilistic Quotient Normalization (PQN) \(^{30}\), Quantile \(^{42}\) and variance stabilization normalization (VSN) \(^{43}\) normalization. The SUM, Mean and Median normalization assume the total/mean/median signals are the same among all study samples and a dilution coefficient is determined correspondingly. While the PQN calculates the dilution coefficient based on all the MS features to infer the most proper dilution coefficient, which is regarded as more stable and robust.
The Quantile normalization was initially used for oligonucleotide array data, and also found application in NMR and LCMS data to remove the unwanted biological variance. It aims at achieving a similar distribution of MS signals across all biological samples. Lastly, the VSN normalization is a non-linear approach aiming at keeping a constant variance of the whole dataset. VSN applies the logarithm to the large values to remove heteroscedasticity and linearly transform the small values to make variance unchanged. Since the underline algorithms of these sample normalization approaches significantly varied, their performance on different datasets varied a lot from each other and dependent heavily on the nature of the datasets.

1.1.3.3 Evaluation to sample normalization
The sample normalization has received wide attention, and a series of evaluation criteria and tools have been reported, such as the NOREVA, NormalizeMets, MetabogroupS. At least five criteria have been proposed for sample normalization, including A) the approach’s capability of reducing the within-group variance measured by relative standard variation (RSD). The underline hypothesis is that if the samples are more comparable to each other, less fluctuation would be measured within the same biological group. B) the influence of the sample normalization methods on DAFs, the more DAFs, the better the sample normalization approach. This criterion is not direct and strong evidence of the sample normalization performance because the true number of the DAFs is unknown, and the number of DAFs not only depends on the biological difference but also the statistical approaches. C) the consistency of identifying DAFs among the different subsets of the whole datasets; D) the approach’s impact on the classification accuracy of the supervised models. E) the similarity between the normalized and the referent data, which is only useful when the internal standard metabolites are available, and the internal standards corrected dataset is used as the golden standard to evaluate different approaches. Although a variety of the sample normalization approaches and evaluation criteria have been proposed, these approaches and criteria were not commonly included in the integrated pipelines, limiting their applicability to complex metabolomics datasets.

1.2 Challenge of data pre-treatment
After the missing value processing, batch effects removal, and the sample normalization, a clean data without obvious unwanted variance is obtained. However, this clean data could not be used directly for statistical analysis because of the scale difference in MS signal among MS features and the biased distribution. The scale difference is often observed in metabolomics datasets as the difference in orders of magnitude among the MS intensities, and this difference may significantly reduce the statistical power of the multivariate analysis. For instance, in an untargeted metabolomics dataset, if the MS signals of a few MS features are two to three folds higher than the rest thousands of MS features, the first and second principal components of the PCA model could be dominated by these few MS features. In other words, the contribution of the rest thousands of MS features is much less than the few high-intensity MS features. There are two main sources of scale difference: firstly, the concentration of the metabolites varied a lot in samples; secondly, the ionization and responding rates varied among MS features, resulting in significant intensity difference. The scale difference should be corrected because it is not a reflection of the concentration difference, but the difference in ionization and responding rates. Besides, the low-intensity MS feature may also have a significant biological influence. Hence, it is necessary to bring the MS features to a similar level of weights during multivariate analysis through scaling and transformation.

Another factor hinders the direct use of clean data for statistical analysis is that the intensities of the MS features are mostly left-censored distributed. While most of the statistical analyses, such as the Student’s t-Test and the principal component analysis (PCA), require the input data to be normally distributed. The left-censored distribution could reduce the power of the statistical analysis, lower the robustness of the outcome and increase the rates of false positives/negatives. As shown in Figure 3, the application of the transformation could bring the left-censored distribution to near normal distribution and reduce the scale differences. The centering followed the transformation could bring all MS features to the same means and further eliminate the scale difference. Hence, it is necessary to adjust the data scale and distribution to ensure robust statistical outcomes.
Transformation and scaling are two effective strategies to deal with the scale difference and biased data distribution. A list of the commonly used transformation and scaling methods are shown in Table 2. Transformation, including the log transformation and the power transformation, can transfer the skewed distribution to the symmetric distribution and reduce the scale difference. The log transformation is reported to be better at these two aspects, while the power transformation can handle the zero values. Scaling eliminates the scale difference among MS features by dividing the MS signals with a scale factor, which is calculated based on the intensities of each MS feature. After scaling, the variance was converted to the difference around the scaling factors, thus eliminating the scale difference. However, this may also enlarge the noise signals in the MS features with low MS signals. According to the ways to determine the scale factors, there are several kinds of scaling approaches, such as the auto, pareto, level and vast scaling. And the selection of the scaling and transformation approaches could significantly influence the statistical outcome. For instance,
Gromski et al. argued that the commonly used auto scaling is not always the most suitable option. The comparison of six scaling methods on four classification models showed that the selection of scaling methods significantly influences the classification accuracy, and the optimal option is context dependent.

Table 2. Overview of the transformation and scaling approaches. Adapted from “Centering, scaling, and transformations: improving the biological information content of metabolomics data”, Robert A van den Berg et al., BMC Genomics, 2006, 7:142.

<table>
<thead>
<tr>
<th>Class</th>
<th>Method</th>
<th>Formula</th>
<th>Goal</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Centering</td>
<td>$x_{ij} = x_{ij} - \bar{x}_i$</td>
<td>Focus on the differences and not the similarities in the data</td>
<td>Remove the offset from the data</td>
<td>When data is heteroscedastic, the effect of this pretreatment method is not always sufficient</td>
</tr>
<tr>
<td></td>
<td>Auto scaling</td>
<td>$x_{ij} = \frac{x_{ij} - \bar{x}_i}{S_i}$</td>
<td>Compare metabolites based on correlations</td>
<td>All metabolites become equally important</td>
<td>Inflation of the measurement errors</td>
</tr>
<tr>
<td></td>
<td>Range scaling</td>
<td>$x_{ij} = \frac{x_{ij} - \bar{x}<em>i}{(x</em>{max} - x_{min})}$</td>
<td>Compare metabolites relative to the biological response range</td>
<td>All metabolites become equally important. Scaling is related to biology</td>
<td>Inflation of the measurement errors and sensitive to outliers</td>
</tr>
<tr>
<td>II</td>
<td>Pareto scaling</td>
<td>$x_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{S_i}}$</td>
<td>Reduce the relative importance of large values but keep data structure partially intact</td>
<td>Stays closer to the original measurement than autoscaling</td>
<td>Sensitive to large fold changes</td>
</tr>
<tr>
<td>Vast scaling</td>
<td>$x_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$</td>
<td>Focus on the metabolites that show small fluctuations</td>
<td>Aims for robustness can use prior group knowledge</td>
<td>Not suited for large induced variation without group structure</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Level scaling</td>
<td>$x_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$</td>
<td>Focus on relative response</td>
<td>Suited for identification of e.g., biomarkers</td>
<td>Inflation of the measurement errors</td>
</tr>
<tr>
<td>III</td>
<td>Log transformation</td>
<td>$x_{ij} = 10 \log(x_{ij})$</td>
<td>Correct for heteroscedasticity, pseudo scaling. Make multiplicative models additive</td>
<td>Reduce heteroscedasticity, multiplicative effects become additive</td>
<td>Difficulties with values with large relative standard deviation and zeros</td>
</tr>
<tr>
<td></td>
<td>Power transformation</td>
<td>$x_{ij} = \sqrt[3]{x_{ij}}$</td>
<td>Correct for heteroscedasticity, pseudo scaling</td>
<td>Reduce heteroscedasticity, no problems with small values</td>
<td>The choice for square root is arbitrary.</td>
</tr>
</tbody>
</table>

There are two ways to adjust the scale difference and data distribution, 1) only applying the scaling or transformation approaches, and 2) applying the scaling after the transformation. It is impractical
to apply the transformation after the scaling because scaling could result in negative values, which could not be used for transformation. Since two transformation methods and six scaling approaches are available, there are multiple combinations of these two strategies. However, there have been minimal investigations of evaluation to scaling and transformation on untargeted metabolomics data and no tools have been proposed to aid the selection of the scaling and transformation strategies. In the integrated pipelines such as the MetaboAnalyst 14 and MetFlow 12, a variety of the scaling and transformation approaches have been implemented without systematic evaluation and the users are recommended to try different methods to select the best one. Since the main goal of the scaling and transformation is to adjust the biased data distribution and reduce the scale difference, the performance might be assessed in the following ways: 1) the ratios of the MS features in a dataset that follow the normal distribution, 2) the existence of the domination effects in the PCA model, which could be assessed by checking the correlations between the signal intensities and the contributions on the PCA loadings. Besides, the scaling and transformation have a significant influence on the outcome of the statistical analysis and may introduce noise into the dataset and result in outliers, and an outlier analysis on the PCA model is always necessary to monitor the side effects of the adjustment.

1.3 Challenge of statistical analysis and differential feature selection

The goal of the statistical analysis in the untargeted metabolomics study could be largely grouped into two categories: predicting the outcome of new samples and identifying the MS features associated with a specific phenotype. The statistical approaches are broadly classified into the univariate and multivariate analysis. The univariate analysis takes one metabolite into consideration at each time, while the multivariate holds a whole view on the datasets. The univariate analysis describes the extent of differential expression and the confidence of the difference, which is simple to implement and easy to interpret. However, the univariate analysis may fail to find the conditional significant DAFs due to ignoring the interactions among the MS features. The multivariate analysis is powerful at handling the intercorrelations and the conditional significant. While in metabolomic datasets, the number of the variables is much larger than that of observations, which may lead to
overfitting, so-called the ‘Curse of dimensionality’. Vinaixa et al.\textsuperscript{11} argued that when the number of the variables is much larger than the samples, which is often the case of untargeted metabolomics dataset, the performance of the model could decrease and has a high risk of overfitting. Since both the univariate and multivariate have pros and cons, they all need to be carefully applied to avoid misconducting.

1.3.1 Univariate analysis approaches

There are a variety of issues needed to be considered during the selection of univariate analysis strategies, such as the study design, the number of the MS features, the distribution of the signal intensity, and the false discovery rate (FDR) control. Vinaixa et al.\textsuperscript{11} proposed a workflow to aid the determination of the univariate analysis method (Figure 4). There are six steps in this workflow: 1) remove the MS features that do not contain biological information, which could be determined by QC samples. The MS features in the QC samples with a CV <0.2 are suggested to remain. In case the QC samples are not available, the MS features in the real samples with a CV >0.2 are kept. 2) check the experimental design, is the samples from the two biological groups paired or not, and the paired test should be applied if the study follows the paired design. 3) check the normality of the data distribution to decide the parameter/non-parameter test. The t-test requires the data to be normally distributed if most of the MS features are not normally distributed, then the non-parameter test such as the Wilcox test, should be used instead. 4) calculate the mean differences of the MS features from two groups using the selected parameter/non-parameter test. 5) apply the FDR correction. The number of the MS features and the selection of FDR correction method are two key issues affecting the FDR results. 6) apply the foldchange cutoff and rank the importance of MS features. Unfortunately, in most integrated pipelines, the univariate analysis did not follow this workflow, limiting the robustness of the univariate analysis.
Figure 4. The workflow of univariate analysis for untargeted metabolomics data. Adapted from “A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data”, Maria Vinaixa et al., Metabolites 2012, 2, 775-795.

1.3.2 multivariate analysis approaches

A variety of multiple variate approaches, which initially used in other Omics, could also be used for analyzing the untargeted metabolomics data. Several reviews have investigated the performance of different multivariate models. One of the most important issues in the supervised approaches is overfitting. To avoid this problem, the permutation test is always necessary to all supervised models.
However, Rubingh et al. 48 argued that when the number of variables is much larger than that of the sample numbers, the power of the permutation test could decrease, in other words, even the permutation test result indicates the model is not overfitting, there is still high risk of overfitting. Hence, it is necessary to control the number of MS features input into the supervised models and carefully check the overfitting problems.

1.4 Objectives

The aim of this Ph.D. study was to develop an expert analysis system for the untargeted metabolomics data analysis. The specific aims can be summarized as three major points listed here:

1. Development of a novel tool for missing value imputation in untargeted metabolomics datasets.
2. Development of the systematic batch effects evaluation criteria by comparatively evaluating the existence between batch effects measurements and proposing new within batch effects measurement.
3. Development of an integrated pipeline, in each step of data analysis rich options are available and systematic evaluations are implemented to aid the achievement of robust results.
2. List of articles related to the topic of this thesis


List of articles not related to the topic of this thesis and completed during the PhD career


3. RESULTS

Part one. The establishment of an integrated untargeted metabolomics pipeline

Result summary of Article 1:
As I have mentioned in the introduction section, in the reported integrated pipelines, many of the data processing, data pretreatment approaches were not included, which limited their applicability to complex datasets. Hence, I developed the metaX (Figure 5), which provides multiple choices to deal with the complex scenario of the untargeted metabolomics datasets, and the main results are listed as follows:

- The qualitative comparison of the functions included in metaX with other reported pipelines showed that metaX was one of the most comprehensive pipelines, in which six missing value imputation, three batch effects removal, four sample normalization, two transformations, six scaling approaches have been implemented in metaX. Besides, metaX also provided multiple options in functional analysis, such as the correlation network analysis and pathway analysis;
- Quality control is vital to ensure reliable results. MetaX provides easy and quick control by checking eight measurements relating to the repeatability of the data before and after the data processing.
- A demo dataset from a coronary heart disease (CHD) study was used to demonstrate the usability of metaX.

Figure 5. The workflow of metaX.
The main advantage of metaX is that many options were implemented, which facilitates the need for customizing the pipeline according to the research requirements. While the limitation of metaX derives from the multiple choices and a lack of guidance in each step of the analysis, and a more comprehensive pipeline is needed.

In this paper, I reviewed the existing strategies for data processing, tested metaX using two publicly available datasets, developed the user-interface of metaX, and revised the manuscript.

Part two. A novel strategy for missing values imputation and evaluation to imputation results

Result summary of Article 2:
In this article, we reported a stepwise imputation tool termed as SIM (Figure 6). In SIM, the missing values are firstly classified into truly missing (TM), low intensity induced missing (LIM), and unclear missing, which are imputed with corresponding optimal approaches. The main results are listed as follows:

- The definition of the TMs, LIMs, and UMs. The TMs are missing values missed completely in one biological group but observed in more than 50% of samples in other groups. Since the TMs are caused by the absent of metabolites, they should be replaced with values close to zero (cZero). The LIMs refer to the missing values origin from low intensity. If the largest observed value of a MS feature from a certain group of samples are lower than the 10% quantile of the whole dataset, the missing values in this MS feature are classified as LIMs. The Min imputation, which is designed for the low intensity induced missing, is applied for LIMs imputation. The remaining missing values without clear missing resources are classified as UMs, and kNN is recommended for UMs imputation.

- The existence of TMs, LIMs, and UMs in real datasets. Four publicly available untargeted metabolomics datasets were recruited in this study. According to the classification criterion of missing values, a few TMs were observed in dataset 3 and 4. The LIMs occupied 33%-56%, while the UMs accounted for 44%-64% in the four datasets.
Determination of the best imputations for TMs, LIMs, and UMs. Five imputation approaches were applied to the TMs imputation on dataset 4, except cZero, none of the imputed results of kNN, MF, PPCA, Mindet were close to zero. To assess the imputation to LIMs, a simulated dataset was built in two steps: 1) a complete dataset was obtained by removing the MS features containing missing values in dataset 4; 2) the values at the lower 1% or 2% quantile were removed. The introduced missing values were imputed using five approaches, and the performances were evaluated by comparing the imputed and the true values. The imputed results of the kNN, MF and PPCA were much larger than the true values, while that of Mindet and Min were comparable to the true values. The simulated datasets were also employed for UMs assessment. The UMs were mimicked by randomly removing 1% or 2% observed values in the complete dataset and imputed using four strategies respectively. To avoid the bias of the randomness, this procedure was repeated 100 times. The comparison between the imputed and true values showed that imputed values of MF and kNN were closest to the true values, while those of Mindet were much smaller than the true values. Since the kNN is much faster than MF, it is recommended for UMs imputation.

SIM outstands other imputation approaches concerning imputation accuracy. The performances of SIM, kNN, PPCA, MF and Mindet were compared on the four untargeted metabolomics datasets with a varied amount of missing values. The imputation accuracy (IA), which measures the similarity of the imputed and observed values in the samples from the same biological group, was used to assess the imputation performance. In spite of the varied missing values, the IAs of the SIM on all four datasets were the smallest compared to that of the rest imputation approaches. The performance of kNN and MF were dataset dependent, good on dataset 3 and poor in 1-2, while the PPCA and Mindet were globally unstratified.
In summary, in this study, I proposed a new imputation strategy to classify the missing values in untargeted metabolomics datasets into different types and impute accordingly, which achieves higher precision compared to the existing methods.

In this paper, I proposed the idea of stepwise imputation, contributed to the codes of SIM, and wrote the manuscript.

**Part three. Evaluation of batch effects removal**

**Result summary of Article 3:**

Although several batch-correction approaches have been proposed, there are no commonly acceptable evaluation criteria to select the optimal batch correction strategy. Here, I developed an online tool for batch effects removal, termed as BCM (Figure 7), and the main results are listed as follows:

- The batch effects correction and evaluation strategies included in this study. Seven commonly used batch effects correction approaches were evaluated in this study, including two QC-based approaches, QCRSC and QCSVR, three ANOVA-based approaches, Limma, Ber and SVA, two other methods Combat and EigenMS. Six measurements were used to evaluate the performance of batch effects correction. Three out of six measurements are QC-based
assessments, including a) the mean relative standard deviation of MS features in QC samples, b) the distribution of QC samples in the PCA score plot, and c) the mean correlation coefficients among QC samples. The other three are the real sample-based measurements, including d) the overlap of samples from different batches on the PCA score plot, e) the ratio of the MS features differentially expressed among batches, and f) the improvement of classification accuracy after batch effects correction.

The comparison of the six reported measurements for batch effects evaluation. The measurement f was the classification accuracy, which was all 1 in dataset 4 and 0.27-0.29 on dataset 2. The PCA plot of the seven corrected datasets showed that the samples from different groups were completely separated in dataset 2 while overlapped in dataset 4, indicating selection of the batch correction approaches did not have a significant influence on the separation of samples in some cases and the classification accuracy was unsuitable for the evaluating of the batch correction performance. The measurement c was the average correlation coefficients among QC samples, the score of the seven batch correction methods were 0.96-0.99 in dataset 4 and 0.98-0.99 in dataset 2, which was not sensitive enough to distinguish different results. The measurement a and b, which were QC-based measurements, showed that the QCRSC and QCSVR were the best on dataset 2 and 4. In contrast to measurements a and b, measurements d and e, which utilize the real samples to assess the batch effects, showed that the samples from different batches were not completely overlapped on QCRSC and QCSVR corrected dataset, and a few features in the QCRSC and QCSVR corrected datasets were still significantly differential among batches.

A measurement was proposed to evaluate the within batch effects by calculating the ratios of MS features that had a correlation coefficient larger than 0.6 between the injection order and signal intensities in the QC samples. The measurement on the four datasets showed that there were considerable within batch effects on all four datasets, and 29%, 46%, and 19% MS features in dataset 1, 2, and 3 showed strong within batch effects, respectively. After the QCRSC and QCSVR correction, less than 3% MS features in dataset 2 and 3, and less than 9% MS features in data set 1 showed strong within batch effects. While ratios of MS features showed strong within batch effects in the rest corrected results were similar to that of the uncorrected datasets.
The evaluation of seven batch effects removal approaches on four publicly available datasets using the three selected measurements showed that the Ber and Limma were the best for between batch effects removal on dataset 2 and 3, while the Combat was the best for between bath effects on dataset 1 and 4. The QCRSC and the QCSVR were the best for within batch effects removal on all four datasets.

An online tool termed as BCM was developed to aid the batch effects removal in untargeted metabolomics datasets, and a dataset from the serum profiling of the coronary heart disease patients was used to demonstrate the usability of BCM.

In summary, there is no single solution to the batch effects removal in the untargeted metabolomics datasets. BCM implemented rich functions and comprehensive evaluation criteria, which could greatly improve the robustness of the untargeted metabolomics study.

In this study, I reviewed the existing batch effects correction methods, performed the comparison of the existing batch effects correction methods, proposed the new within batch effects correction approach, wrote the codes of BCM and the manuscript.

Part four. Expert analysis system for metabolomics datasets

Result summary of Article 4:

An integrated pipeline, termed as MetaboPro, was proposed, in which multiple options and rich evaluations were implemented in each step (Figure 8). An untargeted metabolomic profiling of urine
samples from the coronary disease heart patients was used to demonstrate the reasonability and utility of MetaboPro. The main results were listed as the following:

- Missing value processing: the missing values occupied 5% of the whole dataset, and the missing matrix showed that the health group contained more missing values than the other. After the TMs and LIMs imputation, the remaining UMs were evenly distributed. Four approaches were applied for the remaining missing values imputation, and the imputation accuracy showed that the kNN was the best one.

- Batch effects removal: the boxplot of all the MS features in QC samples showed there was obvious batch effects, especially a strong within batch effects. Six batch effects removal approaches were applied and compared with respect to (1) the batch overlap, (2) the ratios of MS features differential among batches, and (3) the ratios of MS features have a correlation larger than 0.6 between the injection order and signal intensities in QC samples. The results showed that QCRSC was the best in all three measurements.

- Sample normalization: the boxplot of all the MS features in the real samples showed a high diversity of the total metabolite amounts among samples probably due to the dilution effects. Four sample normalization approaches were applied to remove the unwanted biological variation, and the normalization performances were evaluated using (1) the classification accuracy, (2) the reduction of the within-group variance, and (3) the PCA plot. The VSN normalized dataset had the highest classification accuracy and the lowest within-group variance. In consistence with the decreasing of within-group variance, the samples from the same biological group were tightly clustered on the PCA score plots.

- Transformation and scaling: After batch effects removal and sample normalization, a clean dataset was obtained, in which less than 10% of the MS features from the Disease and Health group followed the normal distribution. The application of three transformations greatly improved the ratio of normally distributed MS features. Among them, the Log transformed dataset had the highest ratio of near 70%. To check the necessity of scaling, we checked the existence of dominate effects by investigating the correlation of MS intensities and the contribution of PCA loadings. The PCA score plot of the auto, pareto, range scaled datasets were similar to that of none scaled data, while the samples could not be completely separated after vast and level scaling. Besides, there was no dominant effect on the transformed data.
- Statistical analysis: The log-transformed data was used for the statistical analysis. Since 70% of the MS features followed the normal distribution, the t-test was applied in univariate analysis. The PCA showed the samples from the two biological groups were completely separated from each other. The PLS-DA model was validated by 200 permutation test and the R²Y and Q² were significantly larger than the random model. Following the criteria of 1) foldchange >1.2 or <0.083 and 2) BH corrected p-value <0.05 and 3) VIP >1, 68 DAFs were identified.

Figure 8. The workflow of MetaboPro.

In summary, MetPro provides rich functions to deal with the unwanted variations and noise signals, and in each step of processing, comprehensive evaluations are provided to determine the optimal solution, thus ensuring the robustness of the statistical outcome.

In this study, I designed the workflow of MetaboPro, compared it to the existing pipeline, developed the codes of MetaboPro, and wrote the manuscript.
4. DISCUSSION AND PERSPECTIVE

Untargeted metabolomics studies have been widely applied in a series of research fields. The improvement of the technics allows the comprehensive profiling of large-scale samples, generating a huge amount of high dimensional data. Due to the nature of the data quantitation, long acquisition time, and the heteroscedasticity among samples, the untargeted metabolomics data is evitable subject to tremendous noise and unwanted variation. It is very important to remove these unwanted variations to reveal the real biological difference associated with the phenotype. Thus, a series of data processing, pre-treatment, and statistical approaches are developed. Unfortunately, in the integrated pipelines, only a few of these approaches are implemented. Besides, it is difficult for the users to determine which approach is optimal to their data and whether unwanted variations are completely removed at each step of data analysis. To address these issues, I extended current research and developed novel tools in the following aspects:

Missing value imputation:
It is reported that there are mainly three types of missing values in the untargeted metabolomics datasets: missing because of the experimental design, low intensity induced missing, and random error caused missing. In paper two, we show that the low intensity-induced missing and the unclear missing were two main sources of missing values, and truly missing was only observed in a few studies. Since the missing values in untargeted metabolomics datasets have multiple origins, it raises a question of whether a single imputation approach could properly handle all these missing values?

In paper two, we show that the optimal imputation for the TMs, LIMs, and UMs is not the same and stepwise imputation is necessary. Since the TMs are completely missed in one biological group but observed in the other, it is supposed to be caused by experimental design and should be imputed with values close to zero. However, none of the commonly used imputations replaced the TMs with values close to zero, but with values close to the observed values. In other words, the imputed values were not significantly different from the observed values, which may lead to false-negative findings as the TMs are caused by the absent of metabolites. According to Robin, the low intensity induced missing values belong to missing not at random (MNAR). Using simulated datasets, we show that the
commonly used approaches, such as kNN, PPCA failed to correctly recover the LIMs, while the imputed values of Mindet and Min, which impute the missing data with minimum values, were close to the true values. This result was consistent with our exception, and the reported results that the commonly used approaches could not handle the MNAR and LIMs need to be replaced with values from the lower end of the observed values. On the other hand, the Mindet, which is well suited for LIMs imputation did not work for the UMs because the assumption of Mindet did not fit the origin of UMs. Since various types of missing values exist and the optimal imputation for each type of imputation vary, it is necessary to classify the missing values and impute them step by step.

To evaluate the imputation performance on real datasets, we proposed a criterion, termed as imputation accuracy, which measures the similarity between the observed and imputed values in the samples from the same biological group. The reasonability of the imputation accuracy is that the imputed and the observed values for the samples undergone the same treatment should not be significantly different. On all the four real datasets with different sample sizes, the SIM always outperformed the other commonly used approaches with the highest imputation accuracy. The protentional explanation of this phenomenon is that SIM better imputed the TMs and LIMs than the kNN and MF and better imputed the UMs than the Mindet. For instance, on dataset 2 and 4, a considerable amount of the kNN and MF imputed values were higher than the observed values from the same group, while the majority of the Mindet imputed values were significantly lower than the observed values, which again proved that the stepwise imputation was optimal for untargeted metabolomics datasets.

The main advantage of SIM origin from the classification of the missing values. Although the low intensity is one of the major resources of missing values, the proportion of the LIMs and UMs varied a lot among different datasets. For example, the LIMs ranged from 36%-70% in the four datasets. In the case of the low proportion of LIMs, the improvement of SIM could be limited.

**Batch effects removal**

The batch effects are commonly observed and inevitable in untargeted metabolomics data, which needed to be carefully removed to ensure the robustness of the statistical outcome. Till now, multiple batch effects correction and various evaluation measurements have been proposed. However, there is still no widely accepted workflow for batch effects correction and evaluation, and comprehensive
evaluation is needed. In paper three, we developed systematic criteria for the between and within batch effects assessment and built an online tool to deal with batch effects.

There are at least six reported batch effects removal measurements available, the comparison of these measurements on the two datasets showed that two out of six measurements were unbiased. The classification accuracy is one of the measurements, which has been used for a variety of evaluations, such as the performance of missing value imputation \(^8\), the batch effects correction \(^50\), the sample normalization \(^51\) and the scaling effects \(^7\). The underline hypothesis is that if the noise signals were removed, the predictive power of the supervised model could improve. However, the classification accuracy is affected by a variety of factors, and not specific to any of these corrections, making the classification accuracy low in sensitivity and specificity. In paper three, we showed that the classification accuracies were one on all the seven batch effects corrected results of dataset 2.

Three out of the six measurements are QC-based assessments, which measure the mean correlation coefficients, the mean CV, and the cluster of QC samples. The robustness of these measurements depends on the hypothesis that the QC samples could perfectly represent the batch effects in all MS features, which is not always true. For example, in paper three, we show that in a few MS features, the batch differences represented by the QC samples were not the same as that represented by the biological replicates. In this case, even the QC samples showed that there were no batch effects, while significant batch effects could be observed on real samples. In dataset 1 and 3, the QC-based correction approaches outperformed the rest with respect to the QC-based assessments, while the PCA score plots of dataset 1 and 3 showed that the samples from different batches were still separated after batch effects correction using QCRSC and QCSVR. The other two real sample-based measurements, which estimates the batch overlap of the samples and the ratios of the MS features differentially expressed among batches, were unbiased, as they measured the batch effects in real samples.

The within batch effects removal: none of the above-mentioned batch effects measurements could reflect the removal of the within batch effects. For instance, the criteria d and e, which measure the batch overlap and ratios of the MS features differentially expressed among batches, showed that all the batch effects were removed after the Limma correction. However, the intensities of the QC samples were still highly correlated with the injection order, indicating the existence of strong within
batch effects, and the criteria d and e did not reflect the extent of within batch effects. The ratios of MS features had a correlation coefficient > 0.6 between the injection order, and intensities could be used as a measurement for the within batch effects.

In this study, we established systematic evaluations to both the between and within batch effects removal and built a tool to apply easily and evaluate commonly used batch effects correction approaches.

**Transformation and scaling:**

The selection of the data pretreatment strategies has a dramatic influence of the statistical outcome, and the transformation and scaling are two main classes of approaches. There are two main issues needed to be addressed: firstly, the order of the transformation and scaling, secondly, how to determine the optimal combination of these two data pretreatments. Currently, there is no evaluation of whether only transformation or scaling is enough, or a combination of these two is needed. Gromski et al.\(^7\) compared the selection of scaling on supervised models but did not take the transformation into consideration. In NOREVA, either transformation or scaling could be used to one dataset, and the combination of them is not available. In this study, we suggested to perform the transformation first and then check the necessity of scaling based on two considerations: theoretically, the transformation should adjust the data distribution while the scaling should not. The untargeted metabolomics datasets are mostly left-censored distributed. Hence, the transformation is always necessary. Besides, the transformation should also reduce the scale difference to some extent. Practically, the scaling could result in negative values, which should not be used for transformation.

To guide the selection of the transformation and scaling strategies, we proposed two measurements to assess the transformation and two estimations to assess the scaling approaches. The ratios of MS features follow the normal distribution, and the p-value distribution of the MS features was employed to assess the transformation strategies, while the dominating effect and the PCA plot were used to evaluate the scaling approaches. If there is no dominant effect in the multivariate model and the additional scaling does not improve the model, it is unnecessary to perform the scaling.
In summary, in this section, we claimed the right order of data pretreatment is to conduct transformation prior to scaling. Two criteria for evaluating the transformation approaches and other two measurements for the selection of scaling methods were proposed.

**Development of integrated pipelines**

In this study, we reviewed the challenges of the untargeted metabolomics studies, implemented the commonly used data processing and statistical approaches into an integrated pipeline, termed as MetaX. To help common researchers better analyze their data, we developed a series of evaluation measurements for each step of untargeted metabolomics data analysis, such as the missing value imputation, batch effects removal, sample normalization, transformation, and scaling, the improved integrated data analysis system is termed as MetaboPro. This analysis system should serve two roles for the research community: firstly, this analysis could act as a tutorial for the new users. During the data analysis, MetaboPro could indicate the necessity of certain data processing, such as sample normalization, how do different approaches perform on sample normalization, whether the samples are comparable after the application of certain sample normalization approach. The researchers could gain more insights and deepen their understanding of the data analysis through this step by step process. Secondly, MetaboPro could secure more reliable and robust statistical outcomes. For instance, in the original work ⁵² of the coronary heart disease (CHD) study, PCA analysis was applied to the urine samples from the disease and health group. The samples of the two groups could not be separated on the first components, but on the second components, and the PC1 and 2 only explained 24% of total variance, indicating the existing of noise signals and variances unrelated to the experimental design. While in MetaboPro, after careful execution of data processing and pretreatment, the PCA model could explain as much as 38% of the total variance, and the samples from the two groups were nicely separated on the first component. The improvement of the statistical outcome could be a result of better implementation of the data processing, which better interpreted the biological phenotype.
CONCLUSIONS

In this study, we developed a novel missing value imputation workflow, proposed evaluation criteria to assess the performance of missing value imputation, batch effects removal, transformation, and scaling. We developed an expert analysis system by integrating the commonly used approaches for untargeted metabolomics data analysis and implementing evaluation measurements at each step of the analysis. This pipeline outstands the other existing tools in several ways: firstly, multiple data processing and pretreatment approaches were implemented to deal with the complex variations and noises confounded in the datasets. Secondly, comprehensive evaluation measurements were provided in each step of the analysis to ensure the robustness of statistical outcome. The main challenge of untargeted metabolomics data analysis is to combat unwanted variations and extract meaningful biological outcomes. Our study could greatly satisfy these needs and further reliable statistical conclusions. Besides, the proposition/implementation of this system may improve widely accepted statistical workflow and boost the development of statistical analysis in metabolomics studies.
REFERENCES


20. Stekhoven, D. J.; Bühlmann, P., MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2011, 28 (1), 112-118.


APPENDIX

PAPER 1

metaX: a flexible and comprehensive software for processing metabolomics data

Bo Wen1,2, Zhanlong Mei1,2, Chunwei Zeng1,2 and Siqi Liu1,2*

Abstract

Background: Non-targeted metabolomics based on mass spectrometry enables high-throughput profiling of the metabolites in a biological sample. The large amount of data generated from mass spectrometry requires intensive computational processing for annotation of mass spectra and identification of metabolites. Computational analysis tools that are fully integrated with multiple functions and are easily operated by users who lack extensive knowledge in programming are needed in this research field.

Results: We herein developed an R package, metaX, that is capable of end-to-end metabolomics data analysis through a set of interchangeable modules. Specifically, metaX provides several functions, such as peak picking and annotation, data quality assessment, missing value imputation, data normalization, univariate and multivariate statistics, power analysis and sample size estimation, receiver operating characteristic analysis, biomarker selection, pathway annotation, correlation network analysis, and metabolite identification. In addition, metaX offers a web-based interface (http://metax.genomics.cn) for data quality assessment and normalization method evaluation, and it generates an HTML-based report with a visualized interface. The metaX utilities were demonstrated with a published metabolomics dataset on a large scale. The software is available for operation as either a web-based graphical user interface (GUI) or in the form of command line functions. The package and the example reports are available at http://metax.genomics.cn/

Conclusions: The pipeline of metaX is platform-independent and is easy to use for analysis of metabolomics data generated from mass spectrometry.

Keywords: Metabolomics, Pipeline, Workflow, Quality control, Normalization

Background

Biochemicals (metabolites) with low molecular masses are the ultimate products of biological metabolism, while a metabolome represents the total composite in a given biological system and reflects the interactions among an organism’s genome, gene expression status and the relevant micro-environment [1]. The most prevalent technology used in analysis of metabolomics is non-targeted mass spectrometry (MS) coupled with either liquid chromatography (LC-MS) or gas chromatography (GC-MS) [2, 3]. Generally, these techniques generate a set data of mass spectra with chromatography that includes retention time, peak intensity and chemical masses. Data analysis involves stepwise procedures including peak picking, quality control, data cleaning, preprocessing, univariate and multivariate statistical analysis and data visualization. A number of software packages are available for MS-based metabolomics data analysis as listed in Table 1, including propriety commercial, open-source, and online workflows. The MS manufacturers generally provide propriety software, like SIEVE (Thermo Scientific), MassHunter (Agilent Technologies) and Progenesis QI (Waters), which are often limited in scope and function. Open-source software, such as XCMS [4], CAMERA [5], MAIT [6], MetaboAnalyst [7] and Workflow4Metabolomics [8], usually cover limited processing steps. There is no such comprehensive pipeline that is used across the metabolomics community [9, 10]. Referring to the capabilities of the tools mainly used (as shown in Table 1), an automatic and comprehensive open source pipeline is urgent in
<table>
<thead>
<tr>
<th>Feature</th>
<th>metaX</th>
<th>MALI</th>
<th>Workflow4Metabolomics</th>
<th>MetLINe</th>
<th>metaMS</th>
<th>MetaboNews</th>
<th>MetaboAnalyst</th>
<th>XCMS_HDLine</th>
<th>MetIDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>R, Java</td>
<td>R, PERL, Python, Java</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R, Java</td>
<td>R, Java</td>
<td>R</td>
<td>perl, JavaScript and R</td>
</tr>
<tr>
<td>Platform</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Independents</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Open source</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Usable offline</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Power analysis</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Automatic outlier</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PCA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cluster analysis</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PLS-DA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>RDA analysis</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Normalization</td>
<td>Sum, PON, VSN, QC-FSC, Comet, SVR, quantiles</td>
<td>Linear or local polynomial regression fitting</td>
<td>QC-LSC</td>
<td>Internal standard or quantile normalization</td>
<td>Normalized by sum/median, Normalized by reference sample/feature, sample specific normalization and quantile normalization</td>
<td>Normalized by specific compound or feature</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biomarker analysis</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Correlation network</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Metabolite identification</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Functional analysis</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Quality assessment</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Peak picking</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>HTML-based report</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>(PDF)</td>
<td>-</td>
<td>(PDF)</td>
</tr>
</tbody>
</table>

51
<table>
<thead>
<tr>
<th>Feature</th>
<th>Minine</th>
<th>Metabo</th>
<th>agri-MS</th>
<th>R</th>
<th>EigenMS</th>
<th>Metab</th>
<th>Metabon</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td>JAVA</td>
<td>JAVA, R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Platform</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Open source</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Usable offline</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Power analysis</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automatic outlier samples finding</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
bioinformatics analysis of metabolomics. Basically, the pipeline aims for users to easily perform end-to-end metabolomics data analysis with a flexible combination of different methods to efficiently integrate new modules and to build customized pipelines in multiple ways. We herein developed a comprehensive workflow for analysis of metabolomics data, termed metaX. At the present time, R [11] is a popular statistical programming environment and provides a convenient environment for statistical analysis of metabolomic and other -omics data [12, 13]. We thus designed metaX as an R package that automates analysis of untargeted metabolomics data acquired from LC/MS or GC/MS and offers a user-friendly web-based interface for data quality assessment and normalization evaluation. This workflow, which is open source and rich in functions, encourages experienced programmers to improve the relevant functions or to build their own pipeline within the R framework. Overall, metaX aims to be a tool array that utilizes an end-to-end statistical analysis of metabolomics data.

Implementation

A stepwise overview of data processing using metaX is illustrated in Fig. 1.

Peak picking and inputs

In general, metaX can take mzXML files as input or a peak table file as input. If taking mzXML files as input, metaX will use the R package XCMS [4] to detect peaks, then use the CAMERA [5] package to perform peak annotation. If a peaks table file is an input, metaX transforms the table data from a peak detection software, such as Progenesis QI (exported comma separated value (csv) format file), into an R object compatible with the subsequent workflow.

Pre-processing of raw peak data metabolite

The raw peak intensity data was pre-processed in metaX. Firstly, if a metabolite feature is detected in < 50% of quality control (QC) samples or detected in < 20% of experimental samples, it is removed from data analysis [14]. Secondly, a missing value after the first filtering is retained and imputed. In metaX, four methods are implemented to perform missing value imputation: k-nearest neighbor (KNN), Bayesian principal component analysis replacement (BPCA), svdImpute and random forest imputation (missForest) [15].

Data scaling and transformation

Five different scaling approaches are offered in metaX: Pareto scaling, vast scaling, range scaling, autoscaling

---

**Fig. 1** Overview of metaX. This figure summarizes the main modules, functions and features of metaX. The input data and the functions are included in the figure.
and level scaling [16]. The formulas of these scaling approaches are described in detail elsewhere [16]. In addition, three transformation approaches are offered in metaX: log, generalized logarithm (glog) and cube root transformation.

**Removal of outliers**

metaX provides the ability to automatically remove the outlier samples in the pre-processed data based on expansion of the Hotelling’s $T^2$ distribution ellipse [17]. A sample within the first and second component principal component analysis (PCA) score plot beyond the expanded ellipse is removed, and then the PCA model is recalculated. In default mode, three rounds of outlier removal are performed.

**Normalization**

A metabolomics dataset usually contains unwanted variations introduced by signal drift/attenuation and multiplicative noise across the dynamic range. These effects can detrimentally impact the significant signal discovery and MS features that are required for rigorous quality assurance [14, 18]. In metaX, two types of normalization methods are provided: 1) Sample-based normalization is used to correct different concentrations of samples, such as normalization to total sum, probabilistic quotient normalization (PQN), variance stabilizing normalization (VSN) and quantile-based methods. 2) Peak-based normalization is implemented to correct data within batch experiment analytical variation and batch-to-batch variation in large-scale studies [19]. In this normalization, if a study contains QC samples, the QC-robust spline batch correction (QC-RSC) can be used to alleviate the effects of peak area attenuation [19]. During normalization, the degree of smoothingness is controlled by a parameter that sets the proportion of points for smoothing at each point, while in metaX, this parameter is automatically assigned by using leave-one-out cross validation. On the basis of QC samples, a metabolite feature with a coefficient of variation (CV) over the predetermined value is excluded after normalization. The CV threshold could be set by users; generally, CV values $\leq 30\%$ are recommended. Support vector regression (SVR) [20] and ComBat [21] normalization methods are also implemented in metaX. A user-friendly web-based interface (http://metax.genomics.cn) was offered for rapid evaluation of the data normalization methods for a specified dataset.

**Assessment of data quality**

Pre- and post-normalization, the data quality is visually assessed in several aspects, 1) the peak number distribution, 2) the number of missing value distribution, 3) the boxplot of peak intensity, 4) the total peak intensity distribution, 5) the correlation heatmap of QC samples if available, 6) the metabolite m/z (or mass) distribution, 7) the plot of m/z versus retention time, and 8) the PCA score or loading plot of all samples. There are two ways to perform data quality assessment in metaX, the command line mode and the user-friendly web-based interface at http://metax.genomics.cn/.

**Univariate and multivariate statistical analysis**

metaX offers both univariate and multivariate statistical analysis. For univariate statistical analysis, the parametric statistical test (Students t-test), non-parametric statistical test (Mann-Whitney U test), and classical univariate receiver operating characteristic (ROC) curve analysis are implemented. For multivariate statistical analysis, metaX offers functionalities for cluster analysis, multivariate modelling, including PCA, partial least squares-discriminant analysis (PLS-DA) and orthogonal partial least squares-discriminant analysis (OPLS-DA), with numerical and graphical results and diagnostics (optimal number of components estimated by cross-validation, $R^2$, $Q^2$, variable importance in projection (VIP), statistical significance of the model by permutation testing) [22]. In terms of the univariate test analysis, metaX also offers the false discovery rate (FDR)-corrected $p$-value by using the Benjamini-Hochberg FDR algorithm [23]. The PLS-DA was implemented based on the functions from the pls package [24], and the OPLS-DA was performed using the functions from the ropls package [25].

**Power and sample size analysis**

metaX offers an easy-to-use function to perform the power and sample size analysis. This function is based on the Bioconductor package SSPA [26] and outputs a figure to show the distribution curve of sample size versus the estimated power.

**Metabolite correlation network analysis**

metaX offers two types of network analysis. One is the correlation network analysis without regard for experimental groups information, and the other is differential correlation network analysis, which aims to identify metabolite correlation differences in a physiological state. The former was implemented using the core function from the stats package to calculate the correlation coefficient, and the latter was implemented using the function comp2.cc.fdr from the DiffCorr package [27] to calculate the significantly differential correlations. The igraph package [28] was used for network analysis and visualization. In addition, the network can be exported as a file in formats such as gml and pajek, which can be imported into Cytoscape [29] and Gephi [30] for network analysis and visualization. Both of the correlation network analyses aim to describe the correlation patterns among metabolites across samples, in which nodes represent metabolites and edges represent the correlation between different metabolites. The network analysis offers
Fig. 2 User interface of metaX for quality assessment and normalization evaluation
a complementary method to univariate and multivariate statistical analysis methods.

Metabolite identification
Currently, metaX provides a function for metabolite identification based on the Human Metabolome Database (HMDB) [31], KEGG [32, 33], MassBank [34], PubChem [35], LIPID MAPS [36], MetaCyc [37] and PlantCyc (www.plantcyc.org). Moreover, metaX can easily be extended to support the other databases. The metabolites having molecular weights within a specified tolerance to the query m/z or molecular weight value are retrieved from the databases as putative identifications. The information of adducts and isotopes is utilized to assist in metabolite identification if it is present. The default tolerance is 10 ppm.

Functional analysis
At present, metaX provides a function for metabolite pathway analysis based on IMPaLA [38].

Biomarker analysis
metaX uses functions from the R package “caret” to perform the biomarker selection, model creation and performance evaluation [39]. Currently, two methods, random forest [40] and support vector machine (SVM), are implemented to automatically select the metabolites which show the best performance. After the best set features are

---

**Fig. 3** QC charts generated by metaX. a The intensity of feature distribution before normalization. b The intensity of feature distribution after normalization. c The correlation plot of QC samples before normalization. d The correlation plot of QC samples after normalization. e The missing value distribution in experimental and QC samples. f The CV distribution of all features before and after normalization for each group.
selected, a randomForest model can be created and the ROC curve can be plotted.

**HTML-based report generation**

metaX outputs an HTML-based report by using the Nozzle package [41], which contains quality assessment plots and other analysis results.

**Results and discussion**

To illustrate the applications of metaX, a published nontargeted LC-MS metabolomics dataset from a coronary heart disease (CHD) study was used [42, 43]. The dataset consisted of two batches of 138 plasma samples (59 CHD patients, 43 healthy controls and 36 QC samples) acquired in positive ion mode on an LTQ Orbitrap Velos instrument (Thermo Fisher Scientific, MA, USA). LC-MS raw data files were converted to mzXML format using ProteoWizard (version 3.0.5941) [44] and then were processed by XCMS [4] and CAMERA [5] for peak picking and peak annotation, respectively. In total, 1,438 features were retained for downstream analysis. The mzXML files can be downloaded from the Dryad Digital Repository [43]. It merits to note that the study focus is mainly on the software application and its capabilities, not on the biological interpretation of the generated results.

**Quality assessment of metabolomics data using metaX**

In metabolomics studies, data quality checks are crucial prerequisites to achieve reliable results. metaX offers a quick and easy data quality check of metabolomics data. This can be done using the R function in metaX or a user-friendly web interface at the website http://metax.genomics.cn/ as shown in Fig. 2. The mainly QC charts generated by metaX for the CHD dataset are illustrated in Figs. 3 and 4. The number of features detected per sample over the analysis time (injection order) is illustrated in Fig. 4c, revealing that the peaks acquired from any group, disease, healthy and QC, are randomly distributed. The intensities of all features per samples before and after normalization over the analysis time (injection order) are illustrated in Fig. 3a and b, respectively. The missing value distribution is shown in Fig. 3e, which gives an overview of the percent of missing values of all features in both the QC and experiment samples. According to Crawade’s view, the total missing value plot and the total intensity plot derived from raw data and treated with/without normalization.

![Fig. 4 QC charts generated by metaX.](image)

- **a** The sum intensity of all features per sample before normalization over the analysis time (injection order).
- **b** The sum intensity of all features per sample after normalization over the analysis time (injection order).
- **c** The number of features per sample over the analysis time (injection order).
- **d** The score plot of PCA for the raw feature intensity data.
- **e** The score plot of PCA for the normalized data.
could be used to identify sample outliers [45]. Our analysis supported this. The correlation plots of QC samples before and after normalization by SVR are illustrated in Fig. 3c and d and indicate that the lowest correlation efficiency is enhanced from approximately 0.7 to 0.9. The CV distribution of all features before and after normalization for each group is displayed in Fig. 3f, implying that after normalization, the signal quality is obviously improved. The sum intensity of all features per sample before and after normalization over the analysis time (injection order) is illustrated in Fig. 4a and b, suggesting that normalization could narrow the signal variation. The score plots of PCA for the raw feature intensity data and the normalized data are shown in Fig. 4d and e, respectively, which give an overview of the dataset and showing trends, groupings and outliers before data normalization and after data normalization. The score plot of PCA (Fig. 4d) for the non-normalized data provided a simple and easily interpretable visual check of the presence of batch effects. In Fig. 4d, the two data batches appear as two separated groups upon PCA analysis without normalization, whereas in Fig. 4e, after normalization the batch effect was reduced and all of the QC samples were clustered tightly, which provides an initial evaluation of the data quality. Overall, these QC charts demonstrate the necessity of normalization for metabolomics data, while metaX enables overview of the data quality with different charts.

Evaluation of normalization methods using metaX
A systematic bias in high-throughput metabolomics data is often introduced by various steps of sample processing and data generation. Data normalization can reduce systematic biases. A question related to this issue is how to select a proper normalization method. metaX provides a user-friendly web-based Shiny application (http://metax.genomics.cn) for this purpose. To select the optimal normalization approach for the CHD dataset, seven methods are evaluated using metaX. Figure 5 shows the score plots of PCA using different normalization methods. They indicate that after normalization using QC-RSC, ComBat or SVR, all of the QC samples are clustered more tightly, and the batch effect is effectively reduced compared with other methods. Table 2 presents the quantitative comparison metrics acquired by the different methods. From the results it is clear that all normalization methods performed better than non-normalization used in most of the metrics. Specifically, SVR detects the largest number of features (1293) with CV ≤ 30% in QC samples, followed by QC-RSC (1191). For the average CV of features in QC samples, SVR achieved the best performance, followed by QC-RSC. This is similar to the findings in a previous study [20]. However, QC-RSC could detect the largest number of differentially expressed features (178), followed by SVR (170). Taken together, for this data set, SVR could be an

![Fig. 5 Comparison of different normalization methods from PCA.](image-url)

The different points in the figures refer to different samples, and the samples were color-coded according to their group information and shape-coded according to their batch information.
optimal normalization method, thus it was chosen as the default normalization method for the downstream analysis.

Univariate and multivariate statistical analysis

Data for the QC samples are removed from the dataset prior to univariate and multivariate analysis in metaX. For univariate analysis, Mann-Whitney U test and Student’s t-test are performed to compare disease and health groups, followed by false discovery correction using the Benjamini-Hochberg method using metaX. The results, along with the fold change of the disease group versus health group, are presented in Additional file 1: Table S1. In total, 171 features (13.22% of total features) are detected under the criteria of the corrected p-value (Mann-Whitney U test) ≤ 0.05, fold change ≥ 1.5 or ≤ 0.667 and VIP > 1, and 170 features (13.15% of total features) are detected under the criterion of the corrected p-value (Students t-test) ≤ 0.05, fold change ≥ 1.5 or ≤ 0.667 and VIP > 1. The result is comparable with that of the previous study [42].

For multivariate analysis, PCA, PLS-DA and OPLS-DA are performed by metaX. In PCA analysis, the normalized peak intensity matrix is log transformed, followed by Pareto scaling and centering, and then two components are selected. The PCA score and loading plots are shown in Fig. 6a and b, respectively. The score plot indicates that there is an apparent difference between the disease and health groups. For PLS-DA and OPLS-DA, the normalized peak intensity matrix is also log transformed, followed by Pareto scaling and centering. Two components are selected for PLS-DA and two components (one orthogonal and one predictive) for OPLS-DA. The score and loading plots for PLS-DA and OPLS-DA are shown in Fig. 7a and c, respectively. The R²Y and Q²Y values of the PLS-DA model, which are

<table>
<thead>
<tr>
<th>Methods</th>
<th>NO. of peaks</th>
<th>NO. of peaks (CV ≤ 30%)</th>
<th>DEF</th>
<th>Mean (CV) (mRDA)</th>
<th>Mean (CV) (health)</th>
<th>Mean (CV) (QC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrast</td>
<td>1438</td>
<td>930</td>
<td>127</td>
<td>0.4261</td>
<td>0.3816</td>
<td>0.1636</td>
</tr>
<tr>
<td>none</td>
<td>1438</td>
<td>527</td>
<td>65</td>
<td>0.4865</td>
<td>0.4759</td>
<td>0.2114</td>
</tr>
<tr>
<td>QC_RSC</td>
<td>1438</td>
<td>1191</td>
<td>178</td>
<td>0.5108</td>
<td>0.4664</td>
<td>0.1098</td>
</tr>
<tr>
<td>SVR</td>
<td>1438</td>
<td>1293</td>
<td>170</td>
<td>0.4853</td>
<td>0.4583</td>
<td>0.1081</td>
</tr>
<tr>
<td>PGN</td>
<td>1438</td>
<td>793</td>
<td>125</td>
<td>0.4945</td>
<td>0.4681</td>
<td>0.1777</td>
</tr>
<tr>
<td>Quantiles</td>
<td>1438</td>
<td>740</td>
<td>118</td>
<td>0.4911</td>
<td>0.4646</td>
<td>0.1895</td>
</tr>
<tr>
<td>sum</td>
<td>1438</td>
<td>761</td>
<td>119</td>
<td>0.5044</td>
<td>0.4733</td>
<td>0.1979</td>
</tr>
<tr>
<td>VSN</td>
<td>1438</td>
<td>772</td>
<td>120</td>
<td>0.5014</td>
<td>0.4761</td>
<td>0.1912</td>
</tr>
</tbody>
</table>

Note:

a: After normalization, the number of peaks with CV ≤ 30% in QC samples
b: DEF: differentially expressed features with q-value < 0.05, fold change ≥ 1.5 or fold change < = 0.667 and VIP > 1
Mean (CV) mRDA: The average CV of peaks in mRDA disease group
Mean (CV) health: The average CV of peaks in health group
Mean (CV) QC: The average CV of peaks in QC group

Fig. 6 The score and loading plots of PCA. a) Score plot of PCA and b) Loading plot of PCA. The different points in the figures refer to different samples, and the samples are color-coded according to their group information. The QC samples were removed before performing the PCA analysis.
0.908 and 0.854, respectively, indicate that the model has good goodness of fit and predictive ability. The R² and Q² values of the OPLS-DA model, which are 0.905 and 0.847, respectively, indicate that the model also has good goodness of fit and predictive ability. Overall, the two multivariate data analysis methods, PLS-DA and OPLS-DA, give similar results. To test the validity of the models of PLS-DA and OPLS-DA, a permutation test (n = 200) is performed. As shown in Fig. 7b and d, the test indicated that the two models are valid, and the good predictive ability of the model is not because of over-fitting with a p-value less than 0.05. Taken together, the results of PCA and PLS-DA (or OPLS-DA) show a distinct separation between the disease and health groups.

### Table 3: The biomarkers selected by metaX

<table>
<thead>
<tr>
<th>MZ</th>
<th>RT (min)</th>
<th>Mass</th>
<th>HMDB</th>
<th>Name</th>
<th>Delta (ppm)</th>
<th>Chemical formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>308.0498</td>
<td>10.46</td>
<td>285.0629</td>
<td>HMB814387</td>
<td>Cladribine</td>
<td>–8.18</td>
<td>C10H12ClN5O3</td>
</tr>
<tr>
<td>424.3412</td>
<td>11.94</td>
<td>423.3499</td>
<td>HMB806469</td>
<td>Linoyle carnitine</td>
<td>–2.31</td>
<td>C25H45NO4</td>
</tr>
<tr>
<td>155.0281</td>
<td>2.81</td>
<td>116.0664</td>
<td>HMB832411</td>
<td>2-Methyl-1-methylthio-2-butene</td>
<td>–8.77</td>
<td>C6H12S</td>
</tr>
<tr>
<td>130.0499</td>
<td>3.43</td>
<td>129.0426</td>
<td>HMB800267</td>
<td>Pyroglutamic acid</td>
<td>0.15</td>
<td>C3H7NOC3</td>
</tr>
<tr>
<td>174.9913</td>
<td>2.30</td>
<td>NULL</td>
<td>NULL</td>
<td>NULL</td>
<td>NULL</td>
<td>NULL</td>
</tr>
<tr>
<td>309.0533</td>
<td>10.47</td>
<td>270.0892</td>
<td>HMB833940</td>
<td>Vignafuran</td>
<td>3.44</td>
<td>C16H14O4</td>
</tr>
<tr>
<td>425.3446</td>
<td>11.94</td>
<td>424.3341</td>
<td>HMB806327</td>
<td>Alpha-Tocotrienol</td>
<td>7.62</td>
<td>C5OH44O2</td>
</tr>
<tr>
<td>324.0443</td>
<td>9.33</td>
<td>301.0563</td>
<td>HMB801062</td>
<td>N-Acetyl-D-Glucosamine 6-Phosphate</td>
<td>–3.86</td>
<td>C8H16NO9P</td>
</tr>
</tbody>
</table>
Biomarker analysis, metabolite identification and pathway analysis

To create the classification model between the disease and health groups, the functions implemented in metaX are used to conduct the biomarker selection, model creation and performance evaluation. A recursive feature elimination algorithm with the random forest model is used to select the best feature set. During the treatment, 5-fold cross-validation is used to optimize the model and reduce overfitting. As shown in Table 3, 8 features were selected. To further evaluate the performance of the 8 selected features, the 102 samples were randomly split into two sample sets. One sample set (Disease: 29, Health 29) was for model building and the other (Disease: 14, Health 30) was for testing. Based on the two data sets, the 8 features were used to build a random forest model, and a receiver operating characteristic (ROC) curve of this model was plotted and is shown in Fig. 8. The result indicated that the model based on the 8 features had a good result with an area under the ROC (AUROC) curve of 0.999. The 8 features were then identified based on the HMDB (version 3.6) database.

---

**Fig. 8** The ROC curve result of the six selected metabolites

**Fig. 9** The differential correction network. The top six largest numbers of nodes communities were color-coded. Detailed information about the samples and their communities are presented in Table S3.
through metaX. Seven out of the 8 features were identified with a mass accuracy of < 10 ppm (parts per million). The putative identified metabolites were then submitted to the IMPaLA website (version 9) through metaX to perform the pathway analysis, and the results are presented in Additional file 2: Table S2.

Correlation network analysis

Network-based correlation analysis is a complementary method to the traditional univariate and multivariate statistics that is taken in metabolomics analysis to identify metabolite changes in response to variable status of physiology. All of the features with the normalized intensity described above were used to perform the differential correction network analysis. This analysis can be used to detect the interconnection of metabolite pairs whose relationships are significantly altered due to the disease process. In this study, only the metabolite pairs that had significant differential correlations ($p$-value $c = 0.01$) between the disease and health populations were used to build the network. As shown in Fig. 9, of the network with 266 nodes and 444 edges, a giant component (198/266, 74.44%) was found and the community detection analysis using the fast greedy modularity optimization algorithm against this component resulted in seven communities, in which each one has equal to or greater than 10 nodes detected. In addition, metaX can estimate three centrality metrics (degree, closeness and betweenness) for each node, and they reflect the importance of the node in the entire network (Additional file 3: Table S3). Differentially correlation network analysis is expected to provide useful insights into the underlying biological processes of the clinical development of CHD.

Conclusions

metaX presents a complete data processing software that is easy to operate and capable of dealing with large-scale metabolomics datasets. A metaX user can customize the pipeline according to the research requirements. Compared to software for metabolomics datasets that requires high manual interaction, metaX requires much less manual interaction and can be used in a command line or web-based user-friendly interface. Based upon the fast process and the optimized workflow, therefore, metaX would greatly improve the interpretation of metabolomics data.

Additional files

Additional file 1: Table S1. The fold change and $p$-value for all of the features. (XLXS 146 kb)

Additional file 2: Table S2. The pathway analysis results for the 8 selected biomarkers. (XLXS 14 kb)

Additional file 3: Table S3. The centrality metrics for each node in the network. (XLXS 22 kb)

Abbreviations

FDR: False discovery rate; GUI: Graphical user interface; HMDB: Human metabolome database; MSMS: Tandem mass spectrometry; OPLS-DA: Orthogonal partial least squares discriminant analysis; PCA: Principal component analysis; PLS-DA: Partial least square discriminant analysis; QC: Quality control; ROC: Receiver operating characteristic

Acknowledgements

We thank Dr. David Broadhurst for the implementation of QC-RSC method.

Funding

This study was supported in part by the International Science & Technology Cooperation Program of China (2014DFB30020), Chinese National Basic Research Programs (2014CB040202-2, 2014CB040205) and the National High-Tech Research and Development Program of China (2012AA020202). The funding body was not involved in the design of the study and collection, analysis, and interpretation of data or in writing the manuscript.

Availability of data and materials

GPL-2 licensed and available in the Bioconductor framework.

Authors’ contributions

BW conceived of and designed the project. BW and CWZ wrote the R package. BW, MZL and CWZ developed the website of metaX. BW, MZL and CWZ performed the data analysis and tested the software. BW and SQL wrote the paper and all authors revised and approved.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Publisher’s Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 14 October 2016 Accepted: 3 March 2017
Published online: 21 March 2017

References

PAPER 2

SIM: an imputation pipeline improved by classification of missing values

Zhanlong Mei¹,²,³, Shenyan Zhang¹, Anna-Lisa Schaap-Johansen³, Huanming Yang ¹,²,⁴, Karsten Kristiansen⁵, Siqi Liu¹,²*

¹ BGI-Shenzhen, Shenzhen 518083, China
² China National GeneBank, Shenzhen, 518120, China
³ Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark
⁴ James D. Watson Institute of Genome Science, 310008 Hangzhou, China
⁵ Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark

* Correspondence: siqiliu@genomics.cn; Tel.: +86-13910021096 (S.L)

Zhanlong Mei: meizhanlong@genomics.cn
Shenyan Zhang: zhengshenyan@genomics.cn
Anna-Lisa Schaap-Johansen: alsj@dtu.dk
Huanming Yang: yanghuanming@genomics.cn
Karsten Kristiansen: kk@bio.ku.dk
Siqi Liu: siqiliu@genomics.cn
Abstract:

**Background:** Missing values are commonly observed in metabolomics studies, whereby proper imputation is necessary for metabolomics informatics. Although several imputation approaches have been used, a generally acceptable imputation is still missing.

**Results:** Herein, a new web tool for imputation is proposed, termed SIM, in which missing values are broadly categorized into different classes, and stepwise imputations are applied to treat the corresponding classified missing values. Moreover, SIM contains the functions to evaluate imputation results as well as reasonability of the differentially abundant features. SIM utilities and other imputations were carefully examined using several published metabolomics datasets.

**Conclusion:** SIM provides an easy tool in metabolomics imputation and improves the imputation results.

**Keywords:** missing value; metabolomics; stepwise imputation; imputation evaluation
Background

Metabolomics is a powerful tool for globally identifying and quantifying metabolites in a biological system. At present liquid chromatography combined with mass spectrometry (LC-MS) is a main technology in the field [1]. A metabolomics study based upon LC-MS generates a huge dataset because of the acquisition of MS signals from complex and structurally diverse metabolites and large number of samples. Missing values of MS signals for some metabolites in a dataset is a common phenomenon in a metabolomics project, which leads to distortion of data distribution and variances and hinders statistical interpretation of MS signals. As a matter of fact, missing values are commonly defined as a nonresponse of a variant observed for one or more items in science and technology. How to handle missing values is a tough challenge, while imputation, a process to replace missing values with substituted values, is a common approach to deal with such challenge. Although several imputation techniques are developed and applied in metabolomics, such as k-nearest neighbors (kNN) [2] and singular value decomposition (SVD) [3], there is still a lack of a widely accepted approach.

Missing values derived from complex analyses account for as much as 10% of the entire dataset in a metabolomics project and can be divided into three basic types according to the original causes: missing due to experimental design assigned as truly missing (TM), missing related to low abundance named as low intensity-induced missing (LIM), and missing with unclear reasons denoted as unclear missing (UM) [4, 5]. According to Robbins’s rule [6], TMs and LIMs belong to the missing not at random (MNAR) class because the missing values are either caused by the absence of the metabolites or low intensity, both of which are intensity dependent, whereas UMs are regarded as missing completely at random (MCAR) or missing at random (MAR). If an imputation is simply employed to treat missing values in metabolomics without consideration of the different types of missing values, it may result in biased interpretation of MS data. For example, the kNN is suitable for MCAR and MAR, but may have a poor performance on MNAR [7]. On the other hand, the Deterministic Minimum Imputation (Mindet) [8] is designed to deal with left-censored data or MNAR, which may be oversimplified if the dataset contains both MNAR and MCAR. Webb-Robertson et al. [9] evaluated ten imputation methods on the same metabolomic datasets and concluded that no single imputation could provide a satisfactory solution to all missing values. According to our view, global estimation towards the different types of missing values is a fundamental step to achieve a satisfactory imputation in a metabolomics dataset.
Another challenge in missing value processing is how to evaluate an imputed result. Classification accuracy [10, 11] is often used in metabolomics studies, which estimates how imputation improves classification among several groups. This evaluation method, however, appears to result in a lower sensitivity. For instance, Gromski et al. [10] assessed the classification accuracy using five imputations in a dataset of cell metabolome profiling, and reported that four out of five imputations shared relatively similar classification accuracies ranging from 92 to 98%. Normalized root mean square errors (NRMSE) [2, 12] is another evaluation method, in which the proximity of the imputed result to the reference values is statistically weighted. Nevertheless, in many metabolomics datasets, there is a lack of reference data. To define a reference dataset, a simulated dataset [11] is taken for evaluation, in which missing values of different types are artificially introduced. For example, Wei et al. [13] employed a simulated approach to four metabolomic datasets. A complete matrix generated by removing all the incomplete MS features was set as reference data. Missing values were introduced by randomly removing proportional data from the reference, then were imputed with a specific approach. Finally, NRMSE was used to evaluate the closeness between the imputed and reference data. An alternative method for evaluation of imputed result is to take replication data into consideration, in which a feature in replicates, regardless of whether it is observed or imputed, is assumed to be comparable [14]. Proximity of the imputed to the observed values in replicates, termed as imputation accuracy, is likely to be a criterion to evaluate the imputation performance. Generally, there seems lack of a global evaluation tool for imputations of metabolomics data, and therefore, a combination of the advantages of the evaluation methods described above is necessary to assess metabolomics data generated under different conditions.

In this study, we propose an imputation pipeline for metabolomics informatics, Stepwise Imputation for Metabolomics, simplified as SIM. It is generated on Shiny, a platform for developing an interactive program in R [15]. In SIM, missing values in a metabolomics dataset are categorized into three types according to their missing patterns; then the classified missing values are treated with the corresponding imputations. Importantly, the imputed rationality is surveyed by multiple evaluation approaches in SIM. SIM is user friendly because the graphics interface assists communication between software and users, while the source code for SIM placed onto GitHub is expected to undergo further improvements by interested programmers.
**Implementation**

**Workflow of SIM**

The infrastructure of SIM depicted in Figure 1A is basically divided into three steps, classification of missing values, imputation in a stepwise operation, and evaluation of the imputed results. The missing values are mainly classified into three types, TM, LIM, and UM. The imputation process is implemented in the order, TM first, then LIM, and finally UM. SIM is written in R, and its source code is provided in Supplemental File 1 and is also loaded onto GitHub at https://github.com/zhanlongmei/SIM.git. An online interface composed of five panels (Figure 1B) is offered at https://stepwise-imputation.shinyapps.io/git_sim/. The two files are basically required for testing the pipeline utilities, a file of the sample-by-feature matrix in csv format and another file for sample information in txt format. After data uploading, the missing values are imputed using SIM and other imputation approaches, respectively, and their imputation results are evaluated using imputation accuracy. Finally, the influence of imputation on differential analyses is assessed.

**Imputation of TM values**

Due to different treatments of biological groups, some metabolites or chemicals are only present in certain groups but are completely absent in other groups. For example, metabolites derived from drugs should only be present in the treated group and absent in healthy controls. If a MS feature is completely absent in one group but detected in the other group with at least 50% detection rate, the feature is classified as TM. Since TMs are caused by the true absence of a given metabolite, they should be filled with zero. However, zero may hinder some data transformations, such as log transformation, hence, the low values close to zero (cZero) are used instead. The cZeros are generated by random sampling from 0 to 0.01. The recognition and imputation of TMs in SIM are implemented by a custom-made R program.

**Imputation of LIM values**

LIMs are missing values caused by low intensity signals. If the maximum abundance of a MS feature from a certain group of samples is lower than the 10% quantile of the entire dataset, the missing values in this MS feature are termed as LIM. The LIMs are imputed using Min [9], which is specially designed for missing values due to low intensity in a metabolomics dataset. Briefly, missing values in a MS feature from a specific group of samples are imputed using the
minimum observed value from this MS feature. The program to recognize and impute LIMs is written in-house with R. The cut-off to recognize LIMs is adjustable according to the users’ demands.

*Imputation of UM values*

UMs are not generated by a clearly causal factor, and kNN is recommended for UM imputation. Thus, in SIM the local structure-based imputation kNN is available. In addition, several popular imputation strategies are also available in SIM, including 1) a global structure-based imputation, such as probabilistic PCA (PPCA), and 2) machine learning-based imputation, such as missForest (MF). Several R packages are utilized to achieve the UM imputations, *impute [2]* for kNN, *pcaMethods [3]* for PPCA, and *missForest [16]* for MF.

*Evaluation of imputed results*

Two parameters, NRMSE and imputation accuracy, are used for imputation evaluation. NRMSE measures the numerical differences between the imputed and reference values and is calculated using the ‘missForest’ R package [16]. If the reference dataset is simulated multiple times, the NRMSE evaluation is conducted as many times as the iteration number of simulations. Imputation accuracy for a specific MS feature is defined as the mean of the imputed values divided by the mean of the observed values in replicates. The mean of imputation accuracies of all MS features represents the overall performance of an imputation.

*Comparison of differential abundant features (DAFs)*

One core issue in metabolomics data analysis is to identify DAFs, where imputation approaches may exert a significant influence on DAF. The DAFs are determined by a criteria of fold changes of > 2 or < 0.5 with a p value of < 0.05. The overlap of commonly shared DAFs may indicate the influence of imputation approaches on statistical outcome. For the uncommonly shared DAFs, the intensity distribution of the imputed and observed values is provided to determine the imputation reliability. The programs for DAF analysis and comparison are made in-house with R.

*Benchmark datasets tested by SIM*
To test the utility of SIM, four datasets representing different sample types and sizes were selected (Table 1). Datasets 1 and 2 are metabolomic profiles from serum and urine samples from 59 Coronary heart disease (CHD) patients and 43 healthy controls, and the data were acquired using LC MS in the positive mode with a LTQ Orbitrap instrument (Thermo Fisher Scientific, Waltham, MA) with peak picking conducted using XCMS [17]. Dataset 3 presents metabolomic profiles of mouse spleen tissues, 10 from acutely infected mice, 10 from chronically infected and 10 from healthy control mice. Data were acquired in the positive mode with a SYNA PT G2 XS QTOF (Waters, UK)[18]. Dataset 4 represents metabolomic profiles of mice serum samples, in total 18 mice with 6 acutely infected (AI), 6 chronically infected (CI) and 6 healthy controls (Con). MS data were generated in negative mode with a SYNA PT G2 XS QTOF (Waters, UK)[19]. The peak tables and sample information of datasets 1-4 are available in Supplemental file 2.

<table>
<thead>
<tr>
<th>Experimental Dataset</th>
<th>Dataset1</th>
<th>Dataset2</th>
<th>Dataset3</th>
<th>Dataset4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>102</td>
<td>102</td>
<td>30</td>
<td>18</td>
</tr>
<tr>
<td>Species</td>
<td>human</td>
<td>human</td>
<td>mouse</td>
<td>mouse</td>
</tr>
<tr>
<td>Sample type</td>
<td>serum</td>
<td>urine</td>
<td>spleen tissue</td>
<td>serum</td>
</tr>
<tr>
<td>Missing ratio</td>
<td>1.23%</td>
<td>1.60%</td>
<td>10.09%</td>
<td>3.14%</td>
</tr>
<tr>
<td>Ratio of TMs</td>
<td>0%</td>
<td>0%</td>
<td>1%</td>
<td>3%</td>
</tr>
<tr>
<td>Ratio of LIMs</td>
<td>56%</td>
<td>50%</td>
<td>35%</td>
<td>33%</td>
</tr>
<tr>
<td>Ratio of UM s</td>
<td>44%</td>
<td>50%</td>
<td>64%</td>
<td>64%</td>
</tr>
</tbody>
</table>

**Results and Discussion**

**Assessment of the missing values**

As illustrated in Table 1, the typical metabolomics datasets were collected from two species and three sample types with varied sample sizes. Importantly, each individual dataset contains a specific distribution of missing values. No TM is observed in dataset 1 and 2, while dataset 3 and 4 have comparable levels of TMs. This probably reflects that samples in dataset 3 and 4 represented different experimental conditions. For example, dataset 4 comprises profiling of serum metabolites from mice with/without infection. Metabolites associated with inflammation may only be found in the
infected group, but not in the healthy controls. The LIM percentages in the four datasets range from 33% to 56%, while UM signals range from 44% to 64%, reflecting that LIMs and UMs are generally present in most metabolomics datasets.

Stepwise imputation in SIM

According to the pipeline design for SIM presented in Figure 1A, the missing values are treated in a stepwise manner, following the order TMs, LIMs and UMs. After the imputation process, the imputed datasets are evaluated and further used for differential abundant MS feature (DAF) analysis. To describe how SIM works for imputation, dataset 4 is taken as a typical case that requires imputation.

As mentioned, there are three groups in dataset 4 that were generated from different infectious treatments. Six MS features are completely absent in the AI group, two in the CI group and three in the Con group, likely to reflect the different infection conditions. With randomly picking up three TMs, 453.0495, 385.063 and 552.0492, these TMs are imputed by kNN, MF, PPCA, Mindet and cZero and the imputed results are shown in Figure 2A. Except for the imputed values of cZero, the results obtained from the other imputations are far from zero, demonstrating that cZero is an ideal imputation to treat TMs, whereas, the other imputation tools are unlikely to deal appropriately with TMs.

How is an imputation approach usable for LIMs? A simulated dataset could answer this question, at least partially. The simulated dataset is built in three steps: firstly, a complete dataset is obtained by removing all the MS features with missing values in dataset 4. Secondly, the LIMs are introduced by removing the values lower than the 1% or 2% quantiles in all the MS signals. Thirdly, the simulated missing values are imputed using different approaches. The rationality of the imputation is determined by NRMSE, which measures how close the imputed values are to the real ones. As shown in Figure 2B, the NRMSEs of the imputed results generated from Min and Mindet are much smaller compared with those obtained using kNN, MF and PPCA on dataset 4. Although the LIMs are defined at two levels of quantiles, 1% or 2%, the NRMSEs from Min remain identical with the smallest values in all imputation results. Moreover, the NRMSEs determined from kNN, MF and PPCA are not supersized because these approaches are theoretically workable on MCAR or MAR, but not on MNAR [7]. Taken together, Min stands out as the best imputation for LIMs.
Following a similar procedure to construct a simulation dataset as described above, the simulated UM datasets based on dataset 4 are generated by first removing all the MS features that are not completely observed in the datasets, then randomly introducing 1% or 2% missing values. To overcome the bias during random sampling of UMs, the random sampling is repeated 100 times. The distribution of the NRMSEs derived from different imputations are shown in Figure 2C. In all of the imputed results, MF and kNN exhibit the smallest NRMSEs, whereas Min and Mindet show the largest RMSEs, indicating that the Min and Mindet do not work for UMs, while both kNN and MF indeed perform an acceptable performance for UM imputation. In addition, the kNN running is much faster than MF, hence, it may be efficient in UM imputation.

Comparison of SIM with other imputation algorithms on real datasets

To further assess the advantages of SIM on real datasets, we compare the imputation accuracies of imputed results generated by using SIM and other approaches published. The missing values in the four real datasets with varied missing percentages are imputed using Mindet, kNN, MF, PPCA, and SIM imputation in parallel. The imputation accuracies with log 2 transformation are estimated in all the four datasets and are shown in Figure 3A-D. Ideally, the ratios between the imputed and observed values should be close to one, thus, an imputation accuracy with log 2 transformation close to zero is expected to produce a better imputation, whereas the value far from zero is indicative of a poor approximation to the real values. As compared with other imputations, the SIM imputation possesses the smallest variance of the imputation accuracies in all the four datasets, revealing that the imputed values based on SIM are quite close to the real values. Regarding the other imputations, the imputation performance of kNN and MF is dataset-dependent, good imputation for dataset 3 and poor for dataset 2, and the imputed results generated from PPCA and Mindet are generally not satisfactory. Taken together, SIM is capable to recover the missing values and is advantageous for analysis of metabolomic information.

SIM treatment towards a typical metabolomics dataset

As a typical metabolomic dataset, dataset 4 is treated by SIM to test all the functions in the pipeline, in which a total of 1889 missing values are found. According to the classification criteria for missing values, 66 TM values are identified in the three groups of AI, CI, and Con (Figure 4A). The 618 missing values are recognized as LIMs in all groups (Figure 4B). Except for TMs and LIMs, the remaining 1205 missing values are regarded as UMs. The missing values distribution
of all missing values and the UMs are shown in Figures 4C and D. The missing values in Figure 4C display a biased
distribution, while after treatment of TMs and LIMs, the UMs appear with an even distribution (Figure 4D), indicating a
pattern closer to the MCAR pattern, which might be suitable for an imputation method such as kNN. The complete dataset,
therefore, is obtained by stepwise imputation and is further used for differential analysis.

Identification of DAFs is a main task in a metabolomic study. Since various imputation approaches might generate
differently imputed data in the same dataset, DAFs are likely to be influenced by imputation style. Another useful function
in SIM is to check the influence of imputation on the differential analysis, and to evaluate the validity of DAFs only found
in one imputed result but not the other. Dataset 4 is used to evaluate the DAF utility through comparison of the kNN and
SIM imputed results. Under the criteria of fold change $> 2$ or $< 0.5$ with $p < 0.05$, 267 and 301 DAFs between the AI and
CI group are found based on the kNN and SIM imputation, respectively. As shown in Figure 5A, 265 DAFs are shared
between the kNN and SIM imputed results, while 2 are uniquely identified in kNN and 36 in SIM. Whether the unique
DAF is acceptable? Since all the unique DAFs contain missing values, looking for the imputation accuracy of these unique
DAFs possibly offers an answer. Of the 36 DAFs unique in SIM, TM, LIM and UM generally exist in the unique DAFs.
TMs are found in 6 DAFs, and the imputation results elicited form kNN and SIM to these DAFs are shown in Figure 5B.
The imputed values from kNN are almost equal to the measured values, whereas, these generated from SIM are close to
zero, indicating that the inference to TMs made by kNN is inappropriate and may lead to missing DAFs. Besides, LIMs
and UMs are found in the other 30 DAFs, the distribution of imputed and observed values for the ten typical DAFs are
shown in Figure 5C. Averagely for the 30 DAFs, the imputation accuracies for these DAFs are $0.72 \pm 0.37$ in SIM and
$4.21 \pm 3.8$ in kNN, suggesting that the imputed values derived from kNN are poorly comparable with the observed ones
in the replicates. Taking all the comparison together, inappropriate imputation is a plausible reason why some DAFs are
missed in the dataset impute by kNN, while SIM provides better imputation and results in reliable DAFs discovery.

Conclusions

In summary, the SIM workflow aims at automatic and correct imputation in metabolomics, in particular concerning
classification of missing values and evaluation of imputed results. Based on the simulated datasets, we provide evidence
that classification of the missing values and independent imputation are necessary because different types of missing
values require specific treatments. The rationality of SIM imputed data is supported by its application to four different metabolomics datasets. We envision that SIM is a useful tool for global imputation of missing values in metabolomics informatics.
Availability and requirements

Project name: SIM software

Project home page: https://stepwise-imputation.shinyapps.io/git_sim/

Operating system(s): Platform independent

Programming language: R

Other requirements: None

Any restrictions to use by non-academics: None
List of abbreviations

SIM: stepwise imputation for metabolomics datasets
TM: truly missing
LIM: low intensity induced missing
UM: unclear missing
MCAR: missing completely at random
MAR: missing at random
MNAR: missing not at random
kNN: k-nearest neighbors
MF: missForest
PPCA: probabilistic PCA
SVD: singular value decomposition
Mindet: Deterministic Minimum Imputation
cZero: values close to zero
DAF: differential abundance features
NRMSE: normalized root mean square error
Declarations

• Ethics approval and consent to participate
Not applicable.

• Consent for publication
Not applicable.

• Availability of data and materials
The datasets 1–4 supporting the conclusions of this article are included in Supplemental File 2.

• Competing interests
The authors declare no competing financial interests.

• Funding
This work was supported by the National Key Research and Development Program of China, grant number 2017YFC0908400 and by the National Natural Science Foundation of China, grant number 31500670 and 3170040728.

• Authors’ contributions
Z.M designed SIM, wrote the code and the first version of manuscript; S.Z collected the datasets and tested SIM; A.S, K. K, H. Y and S. L discussed the idea and results and revised the manuscript, with KK performing an extensive final revision of the manuscript.

• Acknowledgement
Not applicable.
Reference

1. Dettmer K, Aronov PA, Hammock BD: Mass spectrometry-based metabolomics. *Mass spectrometry reviews* 2007, 26(1):51-78.


15. R: A language and environment for statistical computing [https://www.R-project.org/]
Figure legends

Figure 1. The workflow (A) and online interface (B) of SIM. The missing values are categorized into three types, TM, LIM and UM, and are imputed sequentially followed by evaluation of the imputed results.

Figure 2. Comparison of the imputed results for a simulated dataset using different imputations. (A) The visualizing evaluation of the imputed results for TMs, (B) The NRMSE evaluation of the imputed results for LIMs, (C) The NRMSE evaluation of the imputed results for UMs.

Figure 3. The evaluation of the imputed results for four metabolomic datasets using different imputations. The imputation accuracies for the imputed results in dataset 1 (A), 2 (B), 3 (C) and 4 (D) using different methods for imputation.

Figure 4. Application of SIM to dataset 4. The distribution of different types of missing values and their imputations in dataset 4. (A) Distribution of the TMs. (B) Distribution of the LIMs. (C) Distribution of all missing values. (D) Distribution of the UMs.

Figure 5. The influence of imputation on DAFs in dataset 4. (A) The overlap of DAFs derived from kNN and SIM. The distribution of observed and imputed (kNN or SIM) values for the TMs found in 6 DAFs (B) and the LIMs and UMs found in the 10 typical DAFs (C).
Figures

A

Raw dataset

- Removal of the MS features with high frequent missing values

Classification of missing values

- TM
- LIM
- UM

Stepwise imputation of missing values

- TM with cZero
- LIM with Min
- UM with kNN

The influence of imputation on statistical outcome

Comparison of SIM to other strategies

Optimized dataset

B

BRI: Stepwise imputation for Heterogeneous Cells

- 1. Trend
- 2. Data subset
- 3. SIM imputation
- 4. Other imputation
- 5. Influence of imputation on differential analysis

Introduction:

TM

Due to the experimental design, some metabolites or chemicals are found only detectable in one group but completely undetectable in the other. A MET feature is completely measured in one group but is detected in the other group with 100% detection rate at least. It is classified into TM. Each missing value is filled with random values with uniform distribution ranging from 0 to 0.1.

LIM

LIMs are missing values missed by low signals. In the grouped dataset, if the maximum abundance of a MET feature from the target group is larger than the 75% quantile of the entire dataset, the missing values in this MET feature are called LIM. The LIMs are imputed with the maximum observed value from this MET feature.

UM

The remaining missing values without clear causes are called UM, and UM is recommended for UM imputation.

Recognition and imputation to TM

In the figure below, the obtained value is marked with the TM is marked red and the LIM/UM is marked blue.

A total of 50-MET group missing has been replaced with null values.
Figure 2
Figure 4
Mei, Z.; Yang, H.; Kristiansen, K.; Liu, S., Evaluation and online realization to batch effects removal in large-scale metabolomic studies. Metabolites 2020 (in preparation)
Evaluation and online realization to batch effects removal in large-scale metabolomic studies

Zhanlong Mei¹,²,³, Huanming Yang¹,²,⁴, Karsten Kristiansen¹,³, Siqi Liu¹,²,*

1. BGI-Shenzhen, Shenzhen 518083, China

2. China National GeneBank, Shenzhen, 518120, China

3. Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark

4. James D. Watson Institute of Genome Science, 310008 Hangzhou, China

*Correspondence to

Dr. Siqi Liu, siqiliu@genomics.cn, BGI-Shenzhen, Shenzhen 518083, China
Abstract

Batch effects are inevitable in metabolomics studies, manifested as signal drifts between or within analysis batches, lowering the power of the statistical analysis. Although several batch-correction approaches have been proposed, there are no commonly acceptable evaluation criteria to select the optimal batch correction strategy. Current evaluation tools emphasize between batch effects removal while failing to address the within batch effects, and the implemented correction approaches do not cover the commonly used strategies. Here, we present two unbiased and independent criteria for between batch effects evaluation by comparing the six reported evaluation approaches on two datasets and propose one measurement for the within batch effects assessment. The utility of the evaluation system was demonstrated by comparing seven commonly used batch correction methods on four datasets, and an online tool was developed for public use. The systematic evaluation criteria and the easy to use tool will contribute to lowering the impact of batch effects and improving the robustness of the statistical analysis.

Keywords

Between batch effects, Within batch effects
**Introduction**

Untargeted metabolomic studies aim at measuring as many metabolites as possible in a large number of samples to provide comprehensive profiling of small molecules in a biological system\(^1\). In a large-scale metabolomics study, the samples are often analyzed on several different batches, resulting in batch effects in MS signals related to the experimental batches and injection orders Goh et al.\(^2\). The batch effects are largely composed of two parts: the between batch effect and the within batch effect\(^3\) or injection order variation\(^4\). The between batch effects are systematic differences caused by the independent acquisition of samples in various running batches, while the within batch effects are often observed as the changes in the MS signal response over injection order in the same batch. Since the statistical analysis without proper management of batch effect may lead to biased conclusion, one critical step of the metabolomic data analysis is to remove these batch effects without distributing the real biological variance.

To check for the existence and assess the removal of batch effects, two main strategies have been proposed: the quality control (QC) sample-based approach and real sample-based measurement. During data acquisition, the real samples from the same biological group are randomly assigned to different batches, while the QC sample, the pooling of all the real samples, is analyzed at a constant interval of real samples. Since the QC samples are technical replicates, the MS signals are assumed to be consistent. Accordingly, three measurements are proposed for batch effects assessment based on QC samples, including (a) the method’s capability of reducing the relative standard deviation (RSD) for MS features in QC samples\(^5-7\); (b) the distribution of QC samples in the score plot of principal component analysis (PCA)\(^4-6\), and (c) improvement of the correlation coefficients among
the QC samples\textsuperscript{4, 5}. Since the real samples are randomly assigned into different batches, there should not be obvious differences among them, and three measurements were proposed based on this hypothesis, including (d) the distances of batch clusters in the PCA score plot\textsuperscript{7, 8}; (e) the improvement of the variance between biological groups and the reduction of variance within biological groups\textsuperscript{8, 9}; (f) the classification accuracy of the sample groups\textsuperscript{6}. A fundamental question is whether all these measurements provide unbiased and independent evaluation? It is necessary to select optimal criteria because calculating all these measurements is impractical, and these measurements may provide conflicting results. Furthermore, all these measurements emphasize the approaches’ ability to remove the between batch effects, while failing to assess the effects of within batch effects removal, which also affects the consistency and accuracy of the measurements\textsuperscript{4}. Above all, a set of evaluation criteria that fairly and systematically assess the removal of both the between and within batch effects is needed.

Currently, several approaches have been developed to remove the batch effects. One of the commonly used approaches is to include internal standards, in which the batch effects could be estimated by calibrating the signals of these internal standards\textsuperscript{10}. However, it is difficult to select a group of internal standards that can fit all metabolites. QC samples-based approaches are also widely adopted, the QC-RLSC\textsuperscript{11}, and QC-SVR\textsuperscript{12, 13} are two representative approaches. These methods basically model the variation of the MS features in QC sample as batch effects and extract them from the test samples. There are also some QC-independent approaches, such as Combat\textsuperscript{14}, Limma\textsuperscript{15}, Surrogate Variable Analysis (SVA)\textsuperscript{16} and Ber\textsuperscript{17}. Most of them adopt two-way ANOVA, while Combat applies an empirical Bayes method to calibrate MS signals in a linear model that contains both biological and batch covariates. Besides, Karpievitch et al.\textsuperscript{18} proposed a singular value decomposition-based method,
EigenMS, to deal with batch effects in metabolomics and proteomics datasets. Due to the variance of batch effects in different datasets and the theories behind the correction strategies vary, the performance of these batch effects correction approaches may vary a lot in different datasets, and there is no single solution to batch effects correction. It is important to select the optimal batch correction approach to ensure the robustness of the statistical analysis.

Although a series of batch effects correction approaches and evaluation measurements have been reported, the options for the batch effects correction in the commonly used integrated pipelines, such as MetaboAnalyst\textsuperscript{19}, metabox\textsuperscript{20} and metaX\textsuperscript{21}, are often limited without systematic evaluation (Table 1). Some evaluation tools, such as the BatchQC\textsuperscript{5} and MetaboDrift\textsuperscript{22}, have been reported to select the right batch correction strategy. In MetaboDrift, three QC-sample based approaches, including QUAD, CUBSPL and LOESS, are available, while in BatchQC, the Combat and SVA are implemented. Neither MetaboDrift nor BatchQC covered all the commonly used correction approaches. Besides, the evaluation criteria used in these tools did not account for the removal of the within-batch effects. Hence, an easy to use tool to implement and evaluate the commonly used batch effects correction strategies is needed.

In this study, we compared the six reported measurements to batch effects correction and selected two unbiased criteria for the between batch effects evaluation. A systematic evaluation model was built by combining the two selected criteria for between batch effects assessment and one measurement for within batch effects evaluation. The evaluation model was further applied to assess the performance of seven commonly used batch effects correction methods on four real datasets. The results showed that the no individual method always outstands other approaches in between batch
effects removal, while the QCRSC and SVR are always the best for the within batch effects removal on the four datasets. Finally, we propose an online tool, batch correction for metabolomics (BCM), for the public to perform batch effects evaluation in metabolomics datasets, and the tool was maintained at http://metax.genomics.cn/BCM.

Methods

Metabolomics data sets
In this study, four metabolomics datasets generated from different sample types with various number of batches were used to evaluate the batch effects correction approaches (Table 2). Dataset 1 was the metabolomic profiles of 102 human serum samples with/without Coronary heart disease. The samples were randomly assigned into two batches and were acquired on an LTQ Orbitrap instrument (Thermo Fisher Scientific, Waltham, MA) in the positive mode. Dataset 2 was the profiling of plasma metabolites in 258 samples, which were randomly divided into 6 batches. The samples coming from three biological groups were acquired on a YNAPT G2 XS QTOF (Waters, UK) instrument in the negative mode. Dataset 3 contains the lipidomic profiles of 90 human plasma, 45 psoriasis patients and 45 healthy control. The plasma samples were randomized and distributed into two batches: 40 in batch 1 and 50 in batch 2, and the data were produced in the positive mode with a SYNAPT G2 XS QTOF (Waters, UK). Dataset 4 was the metabolomic profiles of 134 heart tissues of sheep. The samples coming from the 2 biological groups were randomly divided into 8 batches and acquired on a hybrid 7-T Fourier transform ion cyclotron resonance mass spectrometer (LTQ FT Ultra, Thermo Fisher Scientific, Bremen, Germany). In all four datasets, the quality control (QC) samples were prepared by mixing all the study samples and injection at a constant interval.

Batch effects correction methods
Seven commonly used batch effects correction approaches were tested and evaluated, including two QC-based approaches, QCRSC and QCSVR, three ANOVA-based approaches, including Limma, Ber and SVA, two other methods Combat and EigenMS. Several R packages were implemented to perform the batch correction, including the removeBatcheffects function in limma \(^{27}\) R package, the sva \(^{28}\) for the SVA and Combat function, the Ber \(^{29}\) package for the Ber function. The QCRSC and QCSVR were achieved through in-house scripts; the span value in the LOESS regression of QCRSC was set to be 0.5.

**Criteria used for the evaluation of batch effects correction**

A total of six between and one within batch effects removal measurements were used in this study. Criterion a: the mean RSD of the MS features in the QC samples. The RSD of a MS feature is calculated by dividing the standard deviation of MS intensities to the mean MS intensity. The lower the mean RSD is, the better the batch effects are removed. Criteria b: the tightness of the QC samples in the PCA plot. The removal of the batch effects would improve the consistency among the QC samples, and thus, the QC samples would be tightly clustered. The quantitative measurement of the tightness is calculated by dividing the area of the QC samples on the PCA plot to that of all the samples. Criteria c: the improvement of the correlation coefficients among the QC samples. The higher the mean correlation coefficient is, the higher the QC samples are correlated, indicating that the batch effects are better removed. Criteria d: the overlap of samples from different batches on the PCA plot. Since the samples from the same biological group are randomly distributed on different batches, these samples from different batches should overlap well without the existence of batch effects. The extent of the overlap could be used to measure the batch effects. Criteria e: the difference of samples from different batches tested by ANOVA. The ANOVA tested whether the MS features in samples of the same biological group from different batches were significantly differential. The
ratio of the MS features that are identified as differential could be used to assess the extent of batch effects, with a high ratio indicating severe batch effects and a low ratio meaning minor batch effects. The within-batch effects are assessed using the QC samples, which are technical replicates and their intensities are expected to remain unchanged during the acquisition. If the MS signals of QC samples are highly correlated with the injection order, with a coefficient higher than 0.6 or lower than -0.6, this is taken to signify obvious within batch effects. The ratio of MS features highly correlated with the injection order could serve as the quantitative measurement of the within batch effects, with a lower ratio indicating low and a high ratio meaning high batch effects. All the evaluation criteria were achieved with in-house scripts.

**Implementation of the online tool**
The BCM was built based on Shiny, a R package for developing interactive interface application. The source code of BCM is uploaded into Github at http://metax.genomics.cn/BCM. BCM could be accessed through serval browsers, including Internet Explorer, Firefox, and Safari, without a login requirement. The input files required are a sample-MS feature matrix in csv format and a sample information table in txt format. In the sample information table, the sample names, batch information, group information, and the injection orders should be provided in the first column, and the column title should be named as “sample”, “batch”, “class” and “order” respectively. In the sample-feature matrix, the mass-to-ratio value and the retention time should be provided in the first two columns, which must be named as “mz” and “RT”, the third to the last column in the first row are sample names. Each row represents a MS feature, with the mass-to-ratio and retention time in the first column and intensities in the rest columns. Besides, an example dataset strictly following the above-mentioned requirements is provided in the Tutorial panel test purpose.
Results and Discussion:

1. Evaluation of the batch effects measurements

As mentioned in the “Introduction” section, a total of six measurements for assessing the performance of batch effects removal have been proposed. To explore whether these measurements provide unbiased and independent evaluation, we compared the consistency of the six measurements to the seven batch effects correction methods on datasets 2 and 4. As shown in Figure 1, the performances of the seven batch correction methods were distinguishable based on the criteria a, b, d and e but not c and f. Criterion c measured the mean correlation coefficients between the QC samples, which were 0.96-0.99 in dataset 4 and 0.98-0.99 in dataset 2. As shown in Supplemental Figure 1, the batch PCA plot of the dataset 2 showed that there were significant batch effects in the raw dataset and the Combat and Limma eliminated these batch effects, while the mean correlation coefficients of the QC samples in the raw and corrected datasets were all 0.98. Since the QC samples were technical replicates, they were highly correlated even in the raw datasets, making them sometimes indistinguishable. On the other hand, criterion f measured the classification accuracy of the sample groups, which were all 1 in dataset 4 and 0.27-0.44 in dataset 2. The class PCA plots showed that the two biological groups of datasets 4 were completely separated from each other (Supplemental Figure 2), while the three groups of datasets 2 overlapped extensively in both the raw and corrected datasets (Supplemental Figure 3). Since some other factors, such as the true biological difference, sometimes have a higher impact on the classification accuracy than the selection of batch correction approaches, and it is impractical to define to what extent should the ideal batch effects correction improve the classification accuracy, the classification accuracy is not an ideal batch correction performance measurements.
Another interesting phenomenon is that the criterion a and b, which measure the mean CV of the QC samples and the tightness of QC samples on a PCA plot were highly consistent and dependent, with a correlation coefficient of 0.95 on dataset 4 and 0.93 on dataset 2, which is not surprising because both criteria measure how the correction methods improve the consistency among the QC samples. According to these two measurements, the QCRSC and the QCSVR were the best approaches on both datasets as these two approaches resulted in the lowest CV and smallest QC clusters on the PCA plot. However, they were not the best in the other criteria, such as criterion d and e. For example, as shown in Supplemental Figure 1, the QCRSC did not completely remove the between batch effects on dataset 2 as the samples from the batch 1 and 2 did not completely overlap with samples from other batches, while all the batches completely overlapped in the Combat corrected dataset. The inconsistency between criteria a, b and criteria d, e was caused by the fact that the QC-based approaches remove the batch effects through eliminating the variation in QC samples, making these methods to outperform other methods, such as the Combat. However, there is no guarantee that QC-based approaches better remove the batch effects in both the QC and real samples. Criterion d measures the overlap of all samples from different batches on the PCA plots, which is a measurement at the whole dataset level, while criterion e calculates how correction eliminates the variation within biological groups from a single MS feature level. These two criteria were relatively unbiased and independent and were suggested for evaluation.

2. Evaluation of between batch effects removal
Based on the comparison above, the criteria d and e were unbiased and independent measurements for batch effects removal. We further evaluate the seven batch effects correction approaches on all the four real datasets. As shown in Figure 2, both criteria were distinguishable among different
methods on the four datasets and the two measurements were negatively correlated, which is consistent with the expectation that if the between batch variance was removed, the within replication variance could also decrease. According to criteria d and e, it seemed that the EigenMS was the best on all the four datasets for between batch effects removal as it had the highest ratio of batch overlap and lowest within replicates variation. However, the class PCA plots (Supplemental Figure 1) showed that the group variance might be increased after the correction as no within-group variance was observed and the between-group distance was much larger than the rest results. Nygaard et al. 30 warned that the methods remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Taken dataset 1 as an example, the EigenMS corrected dataset resulted in exaggerated p-value distribution, with more significantly differential MS features (Supplemental Figure 4A). The visual comparison of the intensities of five MS features in the uncorrected and EigenMS corrected datasets showed that there was no difference between the disease and health group in the uncorrected dataset, while clear separation was observed in the EigenMS corrected results (Supplemental Figure 4B). This may be caused by the fact that EigenMs corrects the batch effects by removing all the variations not related to the groups.

Except for EigenMs, there were no approaches that always outperformed the rest on all the four datasets (Figure 2). For dataset 2 and 3, it is clear that Ber and Limma were the best options as they had the highest batch overlap, and the lowest within replicates variance on the two datasets respectively. For dataset 1 and 4, the Combat, QCRSC and QCSVR had the highest batch overlap, while the Combat had the lowest within replications variance compared to that of QCRSC and QCSVR. Hence the Combat was chosen as the best one for dataset 1 and 4. Due to the variance in the nature of batch effects, different methods varied a lot in the performance of batch effects correction. For example, the Limma was best fitted for data set 3 but had a relatively poor performance on dataset
1. The selection of the batch effects correction methods without systematic evaluation or evaluate based on the biased criteria may lead to the inappropriate correction results.

It is worth noting that the QCRSC and QCSVR had the best performance on the whole dataset level in dataset 1 and 4, but not on the MS feature level. This may originate from the fact that the QC sample could well represent the batch effects on the whole, while not on a few MS features. To demonstrate this phenomenon, three MS features from dataset 1 were selected and their raw and corrected intensities generated from different approaches are shown in Supplemental Figure 5. The samples in dataset 1 were divided into two batches, injection 1-57 assigned as batch 1 and 58-138 as batch 2. Obvious between batch effects were observed in the QC group and the Disease/Healthy group, while the level of batch effects represented by QC was slightly different from that of the real samples. For example, in 200.0680023, the fold change of two batches was large in the Disease group while small in the QC group. Since the QC-based batch effects correction approaches remove the batch effects by bringing the QC samples to the same level, the differences between two batches in the real samples could not be completely removed if the batch effects in the real samples were not the same to that of QC samples. On the one hand, the QC-based approaches performed well in batch overlap in dataset 1 and 4, indicating most QC samples well represent the batch effects. On the other hand, the within-group variance indicated that there were in did a few MS features still contain batch effects after correction, and improvement may be needed to better treat these MS features.

3. Evaluation of the within-batch effects
The above-mentioned evaluation criteria are mostly designed for the measurement of the removal of the between batch effects, which may be not suitable for assessing the within batch effects removal. An intuitive way to measure the within batch effects or injection order effects is to use the technical
replicates, checking how the MS intensities of the technical replicates correlated with the injection order within batches. Taken dataset 1 as an example, a set of MS features from QC sample were positively or negatively correlated with the injection order in the raw dataset, indicating the existence of the within batch effects (Figure 3 A). The heatmap plots of the signal intensities of these MS features in the seven batch effects corrected datasets were shown in Figure 3 B-H. In the QCRSC and the QCSVR corrected datasets, none of MS features in the QC samples were correlated with the injection order, indicating the within batch effects were completely removed, while in the rest corrected results, only the SVA and the EigenMs corrected datasets partly removed the within batch effects. The ratio of the MS features in the QC samples that have a correlation coefficient larger/smaller than 0.6/-0.6 between the MS signals, and the injection order is used as the quantitative measurement of the within the batch correction. As shown in Figure 3 I, the QCRSC and QCSVR consistently outstand other approaches with the lowest ratio of MS features correlated with the injection order. The relatively poor performance of the non-QC-based approaches may be caused by the fact that the ANOVA-based approaches do not take the injection order effects into consideration, while the QC-based approaches remove the batch effects by fitting a regression line to the QC samples, which remove the between and within batch effects at the same time.

4. The online realization and case study of evaluation to batch effects removal
The evidence above showed that performance evaluation for batch effects removal is always needed and the criteria d and e were preferred measurements for between batch effects removal. Based on these conclusions, we developed an online tool (Figure 4), called BCM, which is maintained at http://metax.genomics.cn/BCM. In BCM, six batch correction approaches were implemented, including QCRSC, QCSVR, Combat, Limma, Ber and SVA. The EigenMs was excluded because of the high risk of introducing unwanted biological variance. After the data uploading and noise signal
filtering, the existence of batch effects is firstly checked using the extent of batch overlap, reducing in within-group variance, and the eliminating of injection order effects. If a considerable degree of batch effects was detected, as many as six batch effects correction approaches could be available and their performance could be evaluated using the above mentioned three criteria. All the datasets generated from different correction approaches could be downloaded and used for further analysis. Dataset 1 was used as an example to demonstrate the utility of BCM. As shown in Figure 5, the two batches were apparently separated from each other, with an overlap of 0.2. The evaluation of the six batch correction approaches showed that the Combat was best at between batch effects removal but poor in within batch effects removal (Figure 5). The QCRSC and the QCSVR were the best for within batch effects removal while less satisfied in the between batch effects removal. Although the QCRSC and QCSVR were not perfect in reducing the within-group variance due to the inconsistency between the QC and real samples on a few MS features, they still showed good performance on the overall between batch effects removal as the batch overlap was relatively high. Together with the best performance in within batch effects removal, the QCRSC and QCSVR were chosen as the optimized approaches.

**Conclusion**

In this study, we proposed a systematic model and developed an online tool for better removal of the batch effects. The systematic model was built based on the selection of two out of six between batch effects measurements and proposition of a within batch effects assessment. The application of the evaluation model on four real datasets showed that no commonly used batch effects correction approaches always outperformed the other on all the four datasets, and a solid evaluation is needed to determine the best result. With these evaluation criteria and user-friendly online tool, one can easily
apply different batch effects correct approaches and compare their performance in varies ways. By better removing the batch effects, this work may help users to achieve more robust statistical conclusions and thus more reliable biological findings.
ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (2017YFC0908400) and by the National Natural Science Foundation of China (31500670 and 3170040728).

AUTHOR INFORMATION

Corresponding Author

Siqi Liu

siqiliu@genomics.cn

Author Contributions

Zhanlong Mei performed the evaluation, developed the online tool and written the manuscript. Karsten Kristiansen, Huanming Yang and Siqi Liu revised the manuscript.

Competing Financial Interests Statement

The authors declare no competing financial interests.
**Figures Legends**

Figure 1. Comparison of the six measurements to the performance of between effects removal on dataset 4 (A) and dataset 2 (B).

Figure 2. Evaluation of the seven batch effects correction approaches using criteria d and e on dataset 1-4 (A-D).

Figure 3. Visual comparison of the within batch effects removal performance of None (A), QCRSC (B), QCSVR (C), Limma (D), Combat (D), SVA (E), EigenMS (F) and Ber(G) on dataset 1. The quantitative measurement of the within batch effects removal of seven methods on all four datasets (I).

Figure 4. The workflow of the BCM.

Figure 5. Performance evaluation of seven batch effects correction methods on dataset 1 using BCM.
Table 1. Qualitative assessment of the existing metabolomics tools for batch effects removal

<table>
<thead>
<tr>
<th>Tools</th>
<th>Batch correction strategies</th>
<th>Evaluations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Combat loess svr limma sva batchratio EigenMS Ber</td>
<td>a b c d e f g</td>
</tr>
<tr>
<td>BCM</td>
<td>Y Y Y Y Y Y Y Y Y Y Y Y Y</td>
<td></td>
</tr>
<tr>
<td>Metaboanalyst</td>
<td>Y N N N N N N N N N Y N</td>
<td></td>
</tr>
<tr>
<td>NOREVA</td>
<td>N Y N N N N Y N N N N Y</td>
<td></td>
</tr>
<tr>
<td>MSPrep</td>
<td>Y N N N Y N N N N N N N</td>
<td></td>
</tr>
<tr>
<td>metaX</td>
<td>Y Y Y N N N N N Y N Y</td>
<td></td>
</tr>
<tr>
<td>Galax-M</td>
<td>N N N N N N N N N N N</td>
<td></td>
</tr>
<tr>
<td>MetaboNexus</td>
<td>N N N N N N N N N N N</td>
<td></td>
</tr>
<tr>
<td>MetMSLine</td>
<td>N Y N N N N N N N N N</td>
<td></td>
</tr>
<tr>
<td>Metabox</td>
<td>N Y N N N Y N N N N N</td>
<td></td>
</tr>
<tr>
<td>statTarget</td>
<td>N Y N N N N N N N N N</td>
<td></td>
</tr>
<tr>
<td>MetaboCorr</td>
<td>N Y N N Y N N N Y Y N</td>
<td></td>
</tr>
<tr>
<td>BatchQC</td>
<td>Y N N N Y N N N Y N N</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. The information of the datasets used in this study

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sample Type</th>
<th>Platform</th>
<th>No. of Samples</th>
<th>No. of Batches</th>
<th>Biological Replicates</th>
<th>Batches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serum</td>
<td>LTQ Orbitrap</td>
<td>SYNAPT G2 XS QTOF</td>
<td>102</td>
<td>2</td>
<td>Disease: 59 Health: 43</td>
<td>Batch 1: 41 Batch 2: 61</td>
</tr>
<tr>
<td>Skin Tissue</td>
<td>SYNAPT G2 XS QTOF</td>
<td>LTQ FT Ultra</td>
<td>90</td>
<td>2</td>
<td>Disease: 45 Health: 45</td>
<td>Batch 1: 40 Batch 2: 50</td>
</tr>
</tbody>
</table>
Figures

Figure 1
Figure 2
Figure 3
Figure 4
Figure 5
Reference


Nygaard, V.; Rødland, E. A.; Hovig, E., Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses.

Evaluation and online realization to batch effects removal in large-scale metabolomic studies

Zhanlong Mei¹,²,³, Huanming Yang¹,²,⁴, Karsten Kristiansen¹,³, Siqi Liu¹,²,*

1. BGI-Shenzhen, Shenzhen 518083, China

2. China National GeneBank, Shenzhen, 518120, China

3. Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark

4. James D. Watson Institute of Genome Science, 310008 Hangzhou, China

*Correspondence to

Dr. Siqi Liu, siqiliu@genomics.cn, BGI-Shenzhen, Shenzhen 518083, China
Figure S1. The PCA plots colored in batches of the uncorrected and seven corrected results for dataset 2.

Figure S2. The PCA plots colored in biological groups of the uncorrected and seven corrected results for dataset 4.

Figure S3. The PCA plots colored in biological groups of the uncorrected and seven corrected results for dataset 2.

Figure S4. The p-value distribution of all MS features from uncorrected and seven corrected results from dataset 1 (A). The visual comparison of the intensities distribution of five MS features on the uncorrected and EigenMS corrected results from dataset 1 (B).

Figure S5. The visual comparison of the intensities distribution of three MS features on the uncorrected, Combat, QCRSC and QCSVR corrected results from dataset 1.
Figure S1. The PCA plots colored in batches of the uncorrected and seven corrected results for dataset 2.
Figure S2. The PCA plots colored in biological groups of the uncorrected and seven corrected results for dataset 4.
Figure S3. The PCA plots colored in biological groups of the uncorrected and seven corrected results for dataset 2.
Figure S4. The p-value distribution of all MS features from uncorrected and seven corrected results from dataset 1 (A). The visual comparison of the intensities distribution of five MS features on the uncorrected and EigenMS corrected results from dataset 1 (B).
Figure S5. The visual comparison of the intensities distribution of three MS features on the uncorrected, Combat, QCRSC and QCSVR corrected results from dataset 1.
MetaPro: an integrated pipeline for untargeted metabolomics analysis

Zhanlong Mei¹,²,³, Huanming Yang¹,²,⁴, Karsten Kristiansen¹,³, Siqi Liu¹,²,*

1. BGI-Shenzhen, Shenzhen 518083, China

2. China National GeneBank, Shenzhen, 518120, China

3. Laboratory of Genomics and Molecular Biomedicine, Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark

4. James D. Watson Institute of Genome Science, 310008 Hangzhou, China

*Correspondence to

Dr. Siqi Liu, siqiliu@genomics.cn, BGI-Shenzhen, Shenzhen 518083, China
ABSTRACT

With the advance in technology, metabolomics is now able to monitor thousands of MS features in as many as tens of thousands of samples. This presents new challenges in metabolomics data analysis due to the tremendous noise signals confounding with the real variation. A series of data pre-processing is needed to deal with the unwanted variation, and proper statistical approaches are required to identify the true biological difference. However, there are many options in each step of data processing and robust evaluations are needed to figure out the optimal solution, which is an unmet demand in the existing integrated pipelines. Here we presented an online tool, termed as MetaboPro, in which multiple evaluations are provided in each step of data processing to ensure the robustness of the statistical outcome. The MetaboPro is available at http://metax.genomics.cn/MetaboPro. The utility of MetaboPro was demonstrated using a urine metabolomics dataset. By reasonable determination of the optimal data processing, the MetaboPro resulted in a better interpretation of the phenotype. MetaboPro could improve the robustness of the statistical outcome, thus benefit to more reliable conclusions.

KEYWORDS

Metabolomics pipeline, data processing, evaluation measurements
INTRODUCTION

Metabolomics aims at revealing the biological response of an organism to different conditions by quantitative measurement of MS signals from thousands of metabolites [1, 2]. After the sample collection, metabolite extraction and data acquisition, the raw data files are converted into peak table through peak picking, and a series of data processing and statistical analysis are implemented to identify the different abundant MS features (DAF). However, the results of the quantitative analyses are often hampered by the unwanted variations, such as the experimental variation and the unwanted biological variation [3]. What’s more, the MS signals are often non-normally distributed and scale difference is commonly observed among MS features [4, 5]. As a result, the accuracy of the downstream analysis and the robustness of the differential abundant MS features heavily depends on the selection of the data pre-processing and statistical analysis. Till now, there is no widely accepted workflow for the metabolomics data analysis.

To remove these unwanted variations, several data pre-processing steps, including the missing values imputation, batch effects removal, sample normalization, scaling and transformation are needed before the statistical analysis. In each sub-steps of data pre-processing, there are several options and considerations, and the selection of the processing approaches significantly influences the statistical outcome. For example, Gromski et al. [4] argued the selection of scaling methods greatly influences the classification accuracy of the supervised multivariate models. Due to the diverse nature of different datasets, there is no solution to data pre-processing, it is important to try different approaches and identify the most appropriate result. For example, Scheel et al. [6] claimed that the DAFs are sensitive to the selection of the imputation approaches and Webb-Robertson et al. [7] evaluated 10
missing value imputation approaches and concluded that there are no approaches always outstanding other methods. However, it is difficult for common users to implement and evaluate by themselves.

Nowadays, several integrated pipelines have been developed for metabolomics analysis, such as the MeltDB [8], MetaboAnalyst [9], metaX [10], Workflow4Metabolomics [11] and MetFlow [12] (Table 1). However, there are two limitations of these pipelines. Firstly, in most pipelines, a limited number of data preprocessing approaches were included. For example, in the MetFlow, only Mean, Median and Toal intensity normalization approaches are implemented, while the commonly used PQN, VSN and Quantile are not included. Secondly, no or limited evaluations are provided to guide users to select the best approaches. For example, in MetaboAnalyst, there are six approaches for missing value imputation without guidance for method selection. It is very difficult for the common users to determine which combination is the best one and to validate the missing values are properly processed. Herein, we developed MetaboPro as a comprehensive and integrated online tool, in which multiple data processing and statistical options have been implemented and comprehensive evaluations are provided in each sub-steps of data processing to aid the robust statistical analysis.

IMPLEMENTATION

A workflow of the data process and statistical analysis using in MetaboPro is shown in Figure 1. After the data uploading, a series of data processing, including the missing value imputation, batch effects removal and sample normalization approaches are employed to obtain clean data without unwanted variations. The scaling and transformation are further applied to the clean data to better fit the requirements of the statistical analysis. The commonly used univariate and multivariate analysis, such as the Students T-test, Wilcox test, PCA, clustering, PLSDA, and OPLSDA are implemented to
obtain the DAFs. In each step of the data processing and statistical analysis, multiple options are implemented, and systematic evaluations are provided to determine the optimal result.

**Missing value processing**

Missing values are commonly observed in metabolomics datasets with several origins. Missing values are processed in two steps: firstly, the MS features with more than 50% missing values in the QC or real samples are discarded. Secondly, the remaining missing values are imputed by multiple approaches. Due to the complexity of the missing mechanism in metabolomics datasets, there is no single solution for missing values imputation. A total of 5 imputation approaches, including bPCA, pPCA, SVD [13], kNN and LLS [14], are implemented and their performances are assessed through imputation accuracy. The imputation accuracy measures the consistency of the imputed and observed values in the biological replications, which is calculated as the ratios between the means of the imputed and observed values from samples of the same biological group.

**Batch effects removal**

Batch effects are the variation origins from the independent acquisition of the samples in different batches and the experimental variation associated with the injection order. Seven batch effects removal approaches were implemented in MetaboPro, namely, BatchRatio, QCRSC [1], QCSVR [15], Combat [16], SVA [17], Limma [18] and EigenMS [19]. The consistency of the samples from the same biological group, which are randomly assigned into different batches, was used to assess the performance of between batch effects removal. Specifically, two measurements based on this hypothesis were proposed, (a) the ratio of MS features that are identified as differential between two batches in the replication samples; (b) the overlap of samples from different batches on the score plot.
of the principal component analysis [20]. As for the within batch effects estimation, the ratio of MS features whose intensities are highly correlated with the injection order, with a coefficient higher than 0.6 or lower than -0.6, is being used to quantitively measure the within batch effects.

**Sample normalization**

Biological samples, especially the urine, tissue and cell samples, often suffer from the concentration difference. In other words, the same volume and some weight of samples could not ensure the same total amount of metabolites. Five commonly used sample normalization approaches, including SUM, Median [21], VSN [22], Quantile [23] and PQN [24] normalization are implemented in MetaboPro. The optimal sample normalization approach is determined according to (1) the reducing of the within-group variance and (2) the improvement of the classification accuracy [25]. The assumption behind the evaluations is that if the samples are more comparable, the within-group variance could be lower and the classification accuracy of the supervised model could be higher. In MetaboPro, the within-group variance is measured as the mean of the relative standard deviation of MS features in samples of the same biological group, and the classification accuracy is estimated based on the PCA-LDA model.

**Data quality assessment**

High data quality is the base for the robust statistical outcome. In MetaboPro, a dataset could be classified as high quality and reproducibility if it could pass all the following three criteria [10], (1) more than 60% of the MS features in QC samples with a relative standard deviation (RSD) lower than 0.3, (2) the QC sample are tightly clustered in the PCA score plot without outliers and (3) the mean correlation coefficients among QC samples higher than 0.9.
Transformation and scaling

The distribution of the MS intensities is often the left-censored, which does not fit the assumption of several statistical analyses. A total of three transformations, namely, log, cubic and power transformation are implemented in MetaboPro. The Shapiro-Wilk's test is used to determine the ratio of MS features following normal distribution after transformation and the one with the highest ratio of normally distributed MS features is selected as the optimal transformation. Due to the difference in ionization and responding rates, the MS signals of different MS features varied a lot from each other, and a few MS features may be several times of scales larger than the rest MS features. This scale difference, which is not a reflection of the true biological concentration, causes the dominate effects of multivariate analysis and lowers the robustness of the statistical outcome [5]. Five scaling approaches, including, Auto, Pareto, Level, Range and Vast [5] are implemented in MetaboPro. The necessity of the scaling is determined upon the existence of the dominant effects on the principal component model. If the contribution of MS features on PCA loadings is strongly correlated with the MS intensities, the scaling is necessary.

Univariate analysis

The foldchange analysis and hypothesis test are included in the univariate analysis. The workflow of the hypothesis testing is adapted from Vinaixa’s suggestions [26] with slight modification. Firstly, the ratios of the MS features following the normal distribution is used to determine the parameter and non-parameter tests. If more than 50% of the MS features are normally distributed, the Students t-test is suggested, otherwise, the Mann-Whitney U test is adopted. Secondly, the paired and the unpaired test is determined according to the experimental design. The paired information table could be needed
if the paired test is selected. Lastly, false discovery is estimated. The volcano plot is provided to
demonstrate the result of the univariate analysis.

**Multivariate analysis**

Two unsupervised multivariate approaches are implemented, including the principal component
analysis (PCA) and the clustering analysis. The supervised multivariate approaches, including the
partial least square discriminate analysis (PLS-DA) and the orthogonal projections to latent structures
discriminant analysis (OPLSDA), are implemented in MetaboPro. To avoid overfitting, the
supervised models are validated through 200 times permutation test.

**RESULTS AND DISCUSSION**

**Datasets**

To illustrate the utility of the MetaboPro, an untargeted metabolomic profiling of urine samples from
59 CHD patients and 43 healthy controls [27] was used for demonstration purpose. The data was
acquired with an LTQ Orbitrap instrument (Thermo Fisher Scientific, Waltham, MA) in positive
mode. The raw mzXML files were available from the Dryad Digital Repository and the peak picking
was conducted using XCMS [28].

**Missing value processing**

The data integrate check showed that 5% of the values were missed and the healthy group contained
more missing values than the other groups (Figure 2A). After the processing of TMs and LIMs, 2%
of the missing values remained and were evenly distributed on the whole dataset (Figure 2B). Four
commonly used approaches were used to impute the UMs and the performances of these imputation approaches were estimated through imputation accuracy. As shown in Figure 2 C, the boxplot of the ratios between the imputed and observed values showed that most of the bPCA, pPCA and SVD imputed values were apparently larger than the observed values. The kNN and the LLS imputed values were comparable to the observed values, and the kNN was selected as the optimal one as the imputed values of kNN were closer to the observed values.

Removal of the unwanted variation

The unwanted variation often confounder the true biological variation and could be divided into two classes: the unwanted experimental and biological variation. The boxplot of all the MS features in the QC samples indicated the existence of obvious between and within batch effects as the total sample intensities distribution was different between two batches and a decreasing trend in both batches was observed (Figure 3 A). Six batch effects removal approaches were applied and compared with respect to the three criteria. As shown in Figure 3B, the samples from different batches in the QCRSC, Ber and Combat corrected results were highly overlapped, while in the QCRSC corrected result no MS features was differentially expressed among batches or correlated with the injection order. The results showed that the QCRSC performed the best in both the between and within batch effects removal.

After the batch effects removal, we further examined the comparability among samples. As shown in Figure 3 C, the total metabolite amounts varied a lot among different samples, properly due to the dilution effects. Four sample normalization approaches were applied to remove the unwanted biological variation and compared. The classification accuracy of the uncorrected dataset was around
0.6, while that of the VSN and the Quantile normalized datasets were higher than 0.7 (Figure 3 D). The CVs of the MS features in samples from the same group were the lowest on the VSN normalized dataset, followed by the Quantile dataset (Figure 3 E). In consistence with the decreasing of within-group variance, the samples from the same biological group were more tightly clustered on the PCA score plots than the uncorrected dataset (Figure 3 F). Based on the evaluation above, the VSN was the best option for improving the comparability among samples.

Data quality assessment

After the removal of the unwanted variation, the dataset was checked for quality and reproducibility. As shown in Figure 4 A, as much as 75% of the MS features in the QC samples had a relative standard deviation lower than 0.3. The QC samples were tightly clustered on the PCA score plot (Figure 4 B). The cluster analysis showed that the QC samples from the two batches were mixed with each other (Figure 4 C) and the mean correlation coefficient was 0.91. Based on these measurements, we concluded that the data showed high reproducibility.

Transformation and Scaling adjustment

A clean dataset was obtained after batch effects removal and sample normalization, and the Shapiro-Wilk test showed that less than 10% of the MS features from the Disease and Health group followed the normal distribution (Figure 5 B), which does not fit the assumptions of some statistical analysis, such as the T-test. The application of three transformations greatly improved the ratio of normally distributed MS features. More than 65% of the MS features in the Log transformed dataset followed the normal distribution, which was higher than that of the Cubic and Power transformed datasets. Thus, the Log transformation was set as the optimal option. Since the Log transformation could also
reduce the scale difference, we further investigated the necessity of scaling by checking the existence
of dominate effects. The PCA score plot of the Log transformed datasets and the correlation between
the MS intensities and contribution on loadings were shown in Figure 5 C and D. The contributions
of MS features on the loadings were not correlated with their MS intensities, indicating the non-
existence of the domination effects caused by the scale difference. The PCA score plot of the auto,
pareto, range scaled datasets were similar to that of none scaled data, while the samples could not be
completely separated after vast and level scaling (Figure 5 C and D). Since there was no scale
difference observed on the Log transformed dataset and the application of scaling approaches did not
improve the statistical outcome, we suggested using the Log transformed dataset for further statistical
analysis.

Univariate and multivariate statistical analysis

The log-transformed data was used for the statistical analysis and the QC samples were removed
before the univariate and multivariate analysis. For the univariate analysis, the raw MS intensities in
the clean data were used for the foldchange analysis while the Log transformed dataset was used for
Students’ t-test due to the fact that more than 65% of the MS features followed the normal distribution
after log transformation. The volcano plot was used to demonstrate the differential abundant MS
features of univariate analysis (Figure 6 A). The PCA showed the samples from the two biological
groups were completely separated from each other on PC1 and the within-group variance of the
Disease group was larger than the Health group (Figure 6 B). The score plot of the PLS-DA model
was shown in Figure 6 D and the R2Y and Q2 of the PLSDA were 0.96 and 0.88 respectively,
indicating good representation of the phenotype and predictive ability. To validate the PLSDA was
not overfitted, 200 times permutation tests were performed. The R2Y and Q2 of the random model
were significantly lower than the real model (Figure 6 E), indicating the PLS-DA model was not overfitted. Following the criteria of 1) foldchange >1.2 or <0.083 and 2) BH corrected p-value <0.05 and 3) VIP >1, 68 DAFs were identified (Figure 6 F).

For instance, in the original work of the coronary heart disease (CHD) study, the PCA analysis was applied to the urine samples from the disease and health group. The samples of the two groups could not be separated on the first components but on the second components and the PC1 and 2 only explained 24% of the total variance, indicating the existing of noise signals and variances unrelated to the experimental design. While in MetaboPro, after careful execution of data processing and pretreatment, the PCA model could explain as much as 38% of the total variance and the samples from the two groups were nicely separated on the first component. The improvement of the statistical outcome could be a result of better implementation of the data processing, which better interpreted the biological phenotype.

CONCLUSIONS
The main challenge of untargeted metabolomics data analysis is to remove the unwanted variations and noises to reveal the real biological difference. In this study, we proposed an integrated pipeline which guides the users go through the data analysis step-by-step and in each step, we provide the assessment to the necessity of certain data processing, the performance evaluation of the commonly used approaches and assessment to whether the unwanted variation is removed or not. The statistical outcome of the demo dataset showed that through careful evaluation in each step of processing, we could better recover the real biological difference. The MetaboPro is powerful in statistical analysis and easy to use with an interactive interface. Therefore, MetaboPro could greatly improve the interpretation and advance the development of untargeted metabolomics data analysis.
ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (2017YFC0908400) and by the National Natural Science Foundation of China (31500670 and 3170040728).

AUTHOR INFORMATION

Author Contributions

Zhanlong Mei designed MetaboPro, written the code and the manuscript; Karsten Kristiansen, Huanming Yang and Siqi Liu revised the manuscript

Competing Financial Interests Statement

The authors declare no competing financial interests.
Reference


**Figure legends**

Figure1. The workflow of MetaboPro.

Figure2. Evaluation of missing value imputation. The distribution of missing values before (A) and after (B) processing of TMs and LIMs. The comparison of the imputation accuracy of five imputation approaches to UMs (C).

Figure3. Evaluation of the removal of unwanted variation. The existence of the batch effects (A) and the comparison of six batch effects correction approaches (B). The assessment of the comparability among real samples (C). The comparison of four sample normalization approaches with respect to the classification accuracy (D) and reducing of the within-group variance (E). The influence of selection of sample normalization approaches on the PCA analysis.

Figure 4. Data quality assessment. The relative standard deviation of MS features from the QC sample (A). The distribution of the QC samples on the PCA score plot (B) and the correlation matrix of the QC samples (C).

Figure 5. Assessment of the scaling and transformation. The p-value distribution (A) and the ratio of normally distributed MS feature (B) of the clean datasets and three transformed datasets. The PCA score plot (C) and loading pot (D) of the Log transformed datasets and five scaled datasets.

Figure 6. The univariate and multivariate analysis. Volcano plot (A), PCA score plot (B) and clustering analysis to the differential abundant MS features (C). Score plot (D) and the permutation test (E) of the PLS-DA model. The overlap of the differential abundant MS features of the univariate and multivariate analysis (F).
### Table 1. Qualitative assessment of existing integrated metabolomics pipelines

<table>
<thead>
<tr>
<th>Integrated pipelines</th>
<th>Metabo Analyst</th>
<th>MSPrep</th>
<th>metaX</th>
<th>Galax-M</th>
<th>Metabo Nexus</th>
<th>MetMS Line</th>
<th>Metabox</th>
<th>Stat Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>kNN</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>MF</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>BPCA</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>PPCA</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Mindel</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Combat</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>QCRSC</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>QCSVR</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Limma</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>SVA</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Batchratio</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>EigenMS</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Ber</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>PQN</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>SUM</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>VSN</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Quantile</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Log</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Cubic</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Auto</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Pareto</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Level</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Vast</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Range</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>
Figure 1
Figure 2
Figure 3
Figure 4
Figure 6