Bioinformatics, 35(16), 2019, 2870–2872
doi: 10.1093/bioinformatics/bty1066
Advance Access Publication Date: 2 January 2019
Applications Note

Systems biology

MetFlow: an interactive and integrated workflow for metabolomics data cleaning and differential metabolite discovery

Xiaotao Shen[1,2] and Zheng-Jiang Zhu[1,*]

[1]Interdisciplinary Research Center on Biology and Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, China and [2]Univeristy of Chinese Academy of Sciences, Beijing 100049, China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on October 24, 2018; revised on December 13, 2018; editorial decision on December 21, 2018; accepted on December 27, 2018

Abstract

Summary: Mass spectrometry-based metabolomics aims to profile the metabolic changes in biological systems and identify differential metabolites related to physiological phenotypes and aberrant activities. However, many confounding factors during data acquisition complicate metabolomics data, which is characterized by high dimensionality, uncertain degrees of missing and zero values, nonlinearity, unwanted variations and non-normality. Therefore, prior to differential metabolite discovery analysis, various types of data cleaning such as batch alignment, missing value imputation, data normalization and scaling are essentially required for data post-processing. Here, we developed an interactive web server, namely, MetFlow, to provide an integrated and comprehensive workflow for metabolomics data cleaning and differential metabolite discovery.
Availability and implementation: The MetFlow is freely available on http://metflow.zhulab.cn/.
Contact: jiangzhu@sioc.ac.cn
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Metabolomics aims to discover differential metabolites to facilitate functional and mechanistic studies, as well as biomarker discovery (Johnson *et al.*, 2016). Mass spectrometry (MS) has been routinely applied to acquire metabolomics data. After data acquisition, software such as XCMS processes raw MS data to generate a feature table. Notably, during MS data acquisition, many unavoidable experimental factors (e.g. impurity accumulation and decreased sensitivity) cause the signal drift and complicate the metabolomics data through introducing uncertain degrees of missing values (MV), zero values and unwanted variations. The intensity distribution of metabolic features is usually of non-linearity and non-normality. All of these confounding factors in metabolomics data significantly influence the accuracy of subsequent statistical analysis, and present a challenge to discover the differential metabolites.

Therefore, various types of data cleaning are required for metabolomics data post-processing, such as MV imputation, data normalization, which effectively improve data quality and accuracy of statistical analysis. For example, Hrydziuszko *et al.* (2012) demonstrated that different MV imputation methods significantly influenced the biomarker selection. We also demonstrated that proper data normalization reduced unwanted variations and the selected metabolite biomarkers had better differential capabilities for disease diagnosis (Shen *et al.*, 2016).

However, we systematically reviewed various tools in the previous publications (References 1–12 in Supplementary Table S1), including the recent MetaboDiff (Mock *et al.*, 2018), and found several limitations: (i) no comprehensive and integrated workflow for data cleaning is available; most were separately developed as independent tools; (ii) lack of user-friendly, interactive and visualization

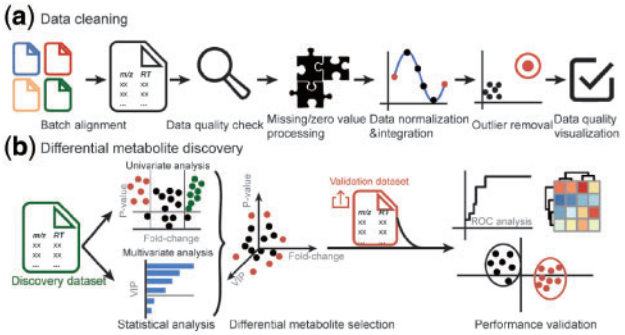

Fig. 1. The standardized workflow for MS-based metabolomics data post-processing using MetFlow: (**a**) data cleaning; and (**b**) differential metabolite discovery

interfaces for users with limited bioinformatics skills; and (iii) batch alignment and integration methods are not readily implemented for multi-batch-based metabolomics dataset. Herein, we developed MetFlow as a comprehensive and integrated web-based platform with an interactive and user-friendly interface. It enables non-bioinformaticians to perform their own post-processing data analysis using a flexible and standardized workflow.

2 Features and methods

The web server provides an integrated and standardized workflow with enough flexibilities and compatibilities to process metabolomics data from LC-MS, GC-MS and various data acquisition methods. Data analysis is performed in a step-by-step fashion, mainly including data cleaning, differential metabolite discovery and pathway enrichment analysis.

2.1 Data cleaning

Data cleaning is implemented as a step-wised and standardized workflow under 'Data Cleaning' tab, shown in Figure 1a.

2.1.1 Batch alignment

For large-scale metabolomics dataset with samples analyzed in multiple batches, batch alignment method was developed. The method contains a two-step alignment. Rough alignment is used to optimize the parameters for accurate alignment, while accurate alignment is used to match features across multi-batches and integrate those as one integral dataset (Supplementary Material Section S2). If there is only one batch, this step is skipped.

2.1.2 Data quality check and visualization

Insufficient data quality in metabolomics tends to generate high false discovery. Before and after data cleaning, data quality is assessed and visualized through different aspects. Specifically, distributions of missing/zero values, and data reproducibility including the distribution of peak variations, degrees of sample clustering in principle component analysis (PCA), auto-scaling boxplot and correlation matrix are employed to assess the data quality.

2.1.3 Missing/zero value processing

Missing/zero values are common in metabolomics data with an uncertain fraction. K-nearest neighbor (KNN) is recommended to impute MV, whereas other methods are also provided (Hrydziuszko *et al.*, 2012). We suggested that peaks with >50% of missing or zero values in all sample groups are considered as noisy peaks and discarded.

2.1.4 Data normalization and integration

Data normalization and integration are utilized to reduce unwanted variations in the dataset. Both of quality control (QC) sample-based normalization (Dunn *et al.*, 2011; Shen *et al.*, 2016) and sample-wised scalar normalization are included. Median/mean values of each metabolic feature from subject or QC samples are usually used as correction factors for data integration (Dunn *et al.*, 2011).

2.1.5 Outlier removal

Outlier samples are assessed and removed to avoid the biases. By default, we suggested that samples outside of Hotelling's T^2 95% CI in PCA score plot are labeled as outliers. Users can decide whether to remove outlier samples on their own consideration.

2.2 Differential metabolite discovery

Differential metabolite discovery is also implemented as a step-wised workflow under 'Differential Metabolite Discovery' tab (Fig. 1b).

2.2.1 Statistical analysis

Multiple common univariate and multivariate analyses are provided. For univariate analysis, volcano plot is displayed to illustrate the differential metabolites according to the *P*-value and fold-change. For multivariate analysis, transformation and scaling are first applied to reduce the contribution of the highly intense metabolic peaks (Guida *et al.*, 2016). Users can try different combinations, and inspect the clustering performance on PCA and partial least squares (PLS) analysis. The unsupervised PCA describes the metabolome-wide difference between samples, while supervised PLS effectively evaluates the contribution of individual metabolite through calculating VIP values. All data analyses are designed in an interactive fashion, and uses can efficiently explore and optimize the parameter combinations.

2.2.2 Differential metabolite selection

The analysis results from univariate and multivariate analyses are combined to select differential metabolites. The cutoffs for fold-change, *P*-value and VIP are set by users, and visualized in a 3D plot.

2.2.3 Performance validation

The performance of selected differential metabolites is finally validated and visualized using discovery and/or validation datasets. A variety of descriptive plots including PCA, PLS, hierarchical clustering analysis (HCA) and receiver operating characteristic are generated. The selected differential metabolites can be imported for pathway enrichment analysis (see Supplementary Material).

After data cleaning and differential metabolite discovery, all analysis results (.zip) and a report (.html) are generated and downloadable.

3 Case study

The usage and performance of MetFlow is showcased with three datasets. Dataset no. 1 is a LC-MS-based untargeted metabolomics study, which includes 434 samples in two batches. Dataset no. 2 is a GC-MS-based untargeted metabolomics study with 25 serum samples (from MetboLights, MTBLS321). Dataset no. 3 is a multiple reaction monitoring (MRM)-based targeted metabolomics study of 89 urine samples (see Supplementary Material).

4 Conclusion

MetFlow is an interactive and integrated web server for metabolomics data cleaning and differential metabolite discovery. The software enables common users with little knowledge in programing and statistics to perform metabolomics data analysis. MetFlow is also flexible to add new algorithms and methods.

Funding

The work is supported by National Natural Science Foundation of China [Grant No. 21575151] and Chinese Academy of Sciences Major Facility-based Open Research Program. Z.-J. Z. is supported by Thousand Youth Talents Program.

Conflict of Interest: none declared.

References

Dunn,W. *et al*. (2011) Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.*, **6**, 1060–1083.

Guida,R. *et al*. (2016) Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics*, **12**, 93.

Hrydziuszko,O. *et al*. (2012) Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. *Metabolomics*, **8**, 161–174.

Johnson,C. *et al*. (2016) Metabolomics: beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.*, **17**, 451–459.

Shen,X. *et al*. (2016) Normalization and integration of large-scale metabolomics data using support vector regression. *Metabolomics*, **12**, 89.

Mock,A. *et al*. (2018) MetaboDiff: an R package for differential metabolomic analysis. *Bioinformatics*, **34**, 3417–3418.