nature portfolio

https://doi.org/10.1038/s41596-021-00636-9

Supplementary information

Optimization of metabolomic data processing using NOREVA

In the format provided by the
authors and unedited

Supplementary Information for:

Optimization of Metabolomic Data Processing Using NOREVA

Jianbo Fu[1,†], Ying Zhang[1,†], Yunxia Wang[1], Hongning Zhang[1], Jin Liu[1], Jing Tang[1], Qingxia Yang[1], Huaicheng Sun[1,2], Wenqi Qiu[3], Yinghui Ma[4], Zhaorong Li[2], Mingyue Zheng[1,5] and Feng Zhu[1,2,*]

[1] College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China

[2] Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Hangzhou 330110, China

[3] Department of Surgery, HKU-SZH & Faculty of Medicine, The University of Hong Kong, Hong Kong, China.

[4] School of Economics and Management, Jiangsu University of Science and Technology, Zhenjiang 212100, China

[5] Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China

[*] E-mail: zhufeng@zju.edu.cn; Tel.: +86-189-8946-6518; Fax.: +86-0571-8820-8444; Lab home page: https://idrblab.org/Peoples-PI.php.

[†] These authors contributed equally: Jianbo Fu, Ying Zhang.

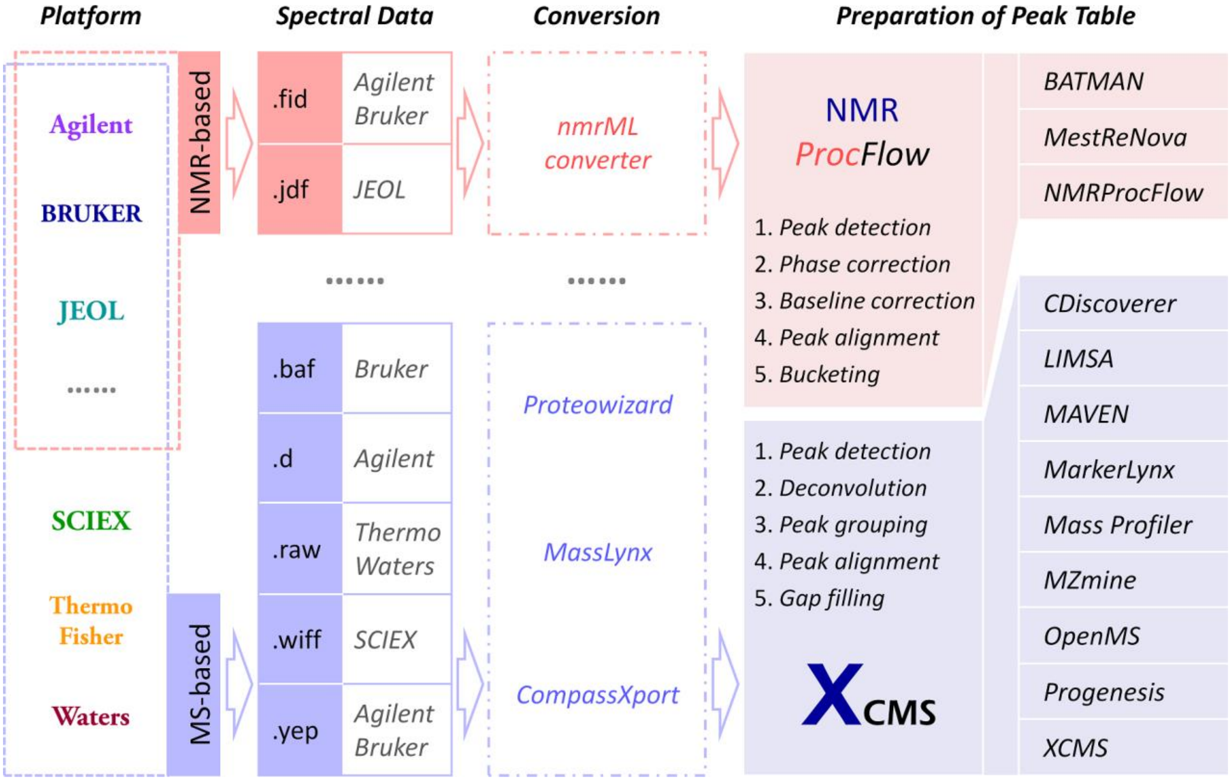


Figure S1. Pre-processing of the spectral data generated based on nuclear magnetic resonance (NMR) or mass spectrometry (MS). Spectral data were first generated using various platforms developed by different vendors, which were then *converted* to the open-source format. Finally, a peak table was *prepared* based on the resulting files of open-source format. The *preparation* process (peak detection, phase correction, baseline correction, peak alignment & bucketing) for NMR-based metabolomics is slightly different from that (peak detection, deconvolution, peak grouping, peak alignment & gap filling) for the MS-based ones. The resulting peak table gives a starting point for the protocol described in this study.

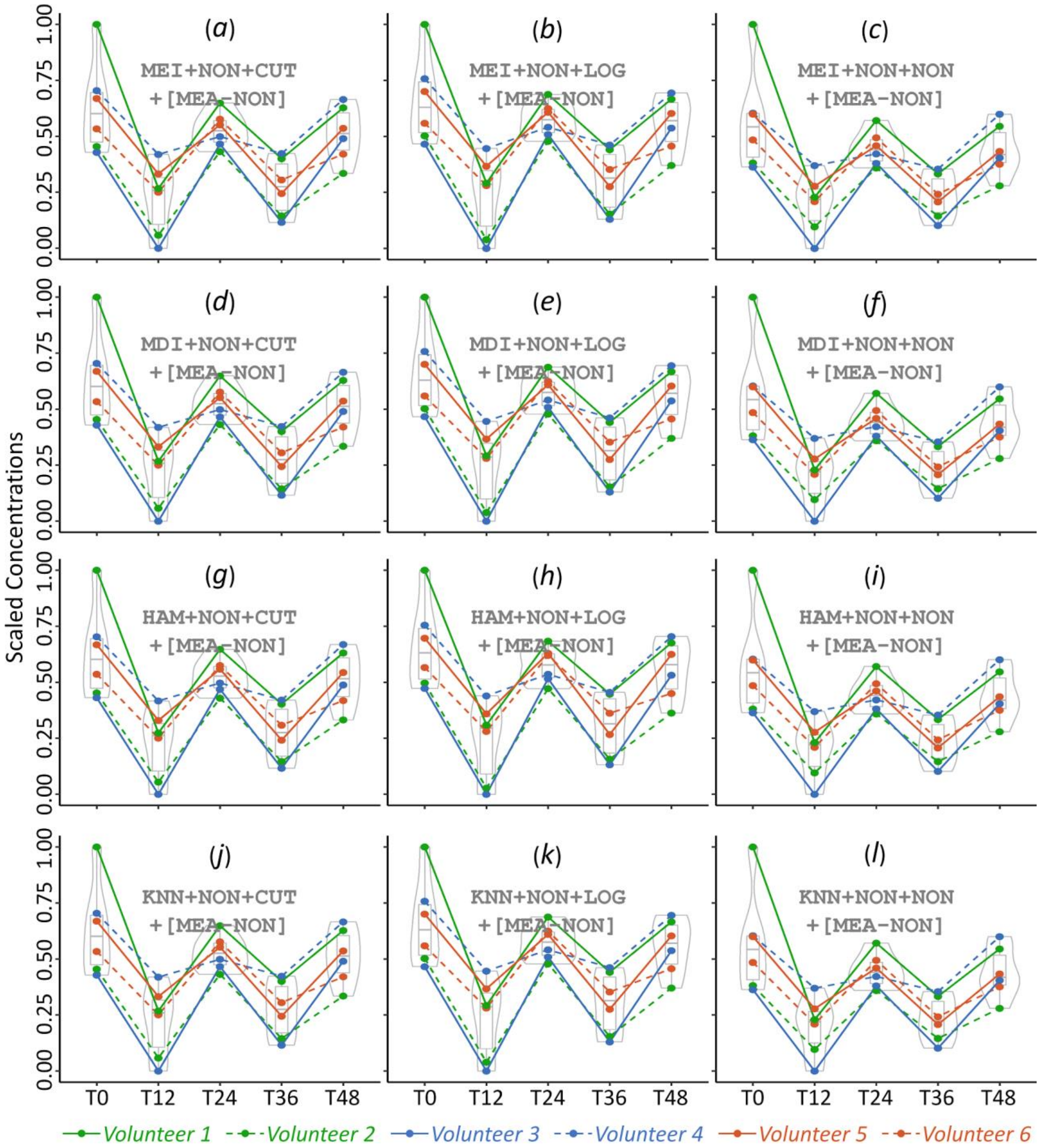


Figure S2. The performances of 12 possible workflows (that were used in *Skarke*'s pioneering study (1) to a time-course dataset PMID29215023) assessed using a well-established metabolic biomarker *Cortisol* (that elevates in the morning and declines in the evening). This dataset is a time-course consecutive sample collection of different time-points (T0: 0 hour in the morning; T12: 12 hours in the evening; T24: 24 hours in the morning; T36: 36 hours in the evening; T48: 48 hours in the morning). The time-dependent fluctuation pattern of metabolic marker *Cortisol* was successfully reproduced by all 12 possible workflows.

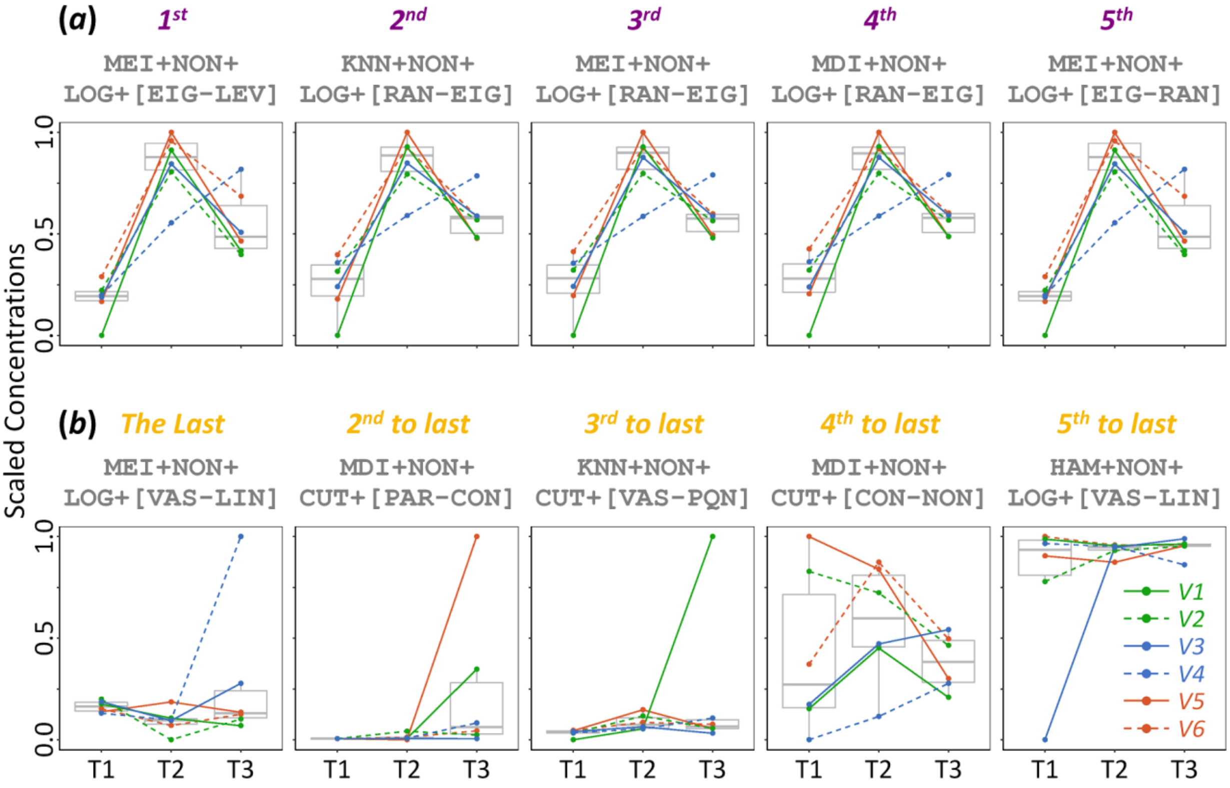


Figure S3. The processing outcomes of (***a***) five top-ranked and (***b***) five last-ranked workflows on the well-established metabolic marker '*kynurenine*' at three different time-points. **T1**: before malaria infection; **T2**: on the day of positive blood smear; **T3**: three weeks after anti-infectious treatment. (***a***) five top-ranked workflows can effectively preserve the 'true' biological variation of '*kynurenine*', which was reported (2,3) to elevate in patient plasma after infection (T1 to T2) and then decline after anti-malaria treatment (T2-T3); (***b***) the last-ranked workflows can hardly reproduce such 'true' biological variation (no statistically significant variation between any two time-points). V1, V2, V3, V4, V5, and V6 referred to six volunteers that were numbered from 1 to 6. Corresponding processing workflow was provided for each plot, and detailed descriptions on the processing methods in these workflows can be found in **Table S4** and **Table S5**.

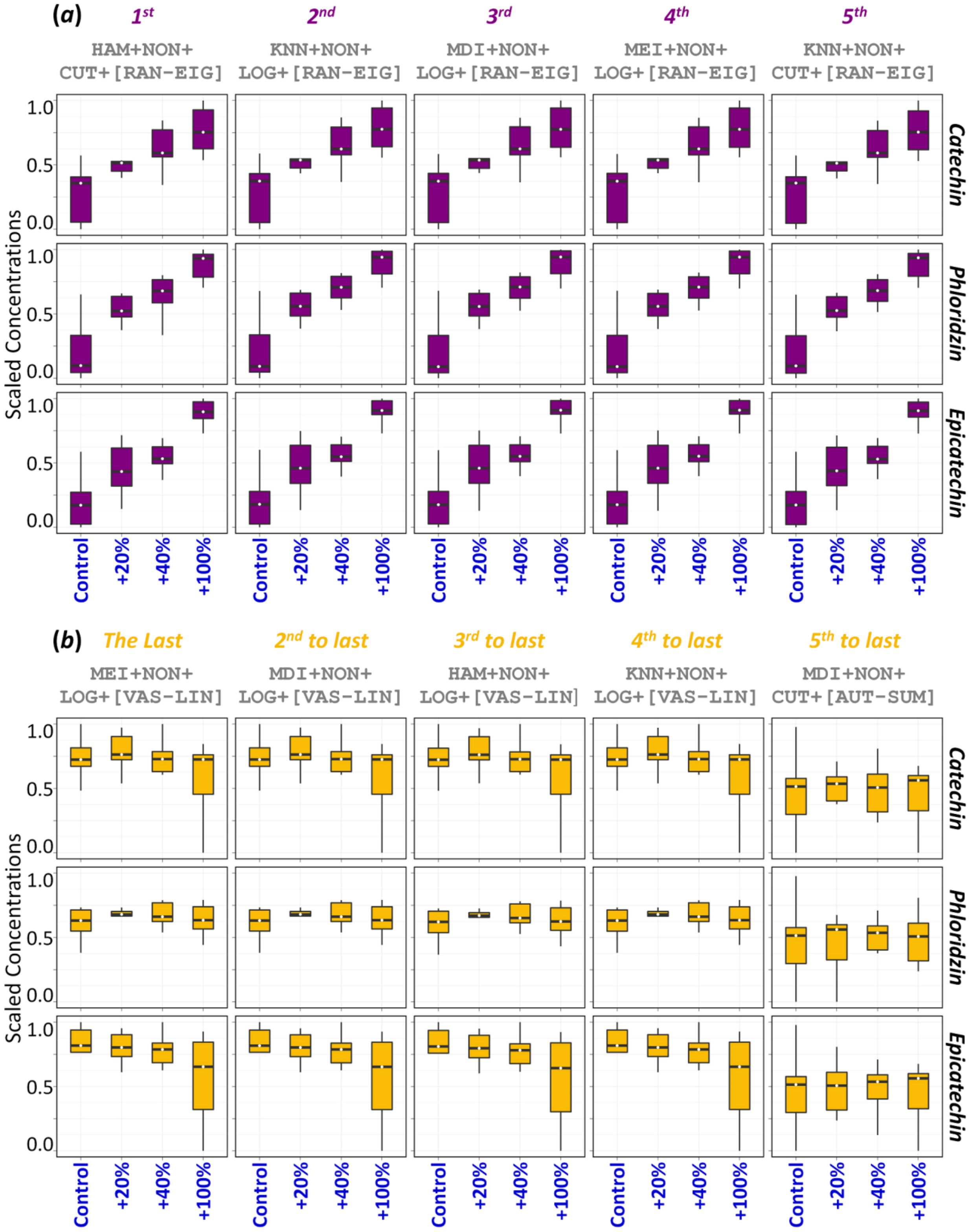


Figure S4. The processing outcomes of (***a***) five top-ranked and (***b***) five last-ranked workflows on three compounds (*catechin*, *phloridzin*, and *epicatechin*) spiked with the gradual increase of concentrations from the control to an increase of 20%, then 40%, and finally 100% (4). (***a***) five top-ranked workflows largely preserved the expected concentration variations of these spike-in compounds (from control to +20%, then to +40%, and finally to +100%); (***b***) five last-ranked workflows could not reproduce the spiked concentration variations for any of the compounds. The corresponding processing workflow was provided for each plot, and detailed descriptions on the processing methods in these workflows can be found in **Table S4** and **Table S5**.

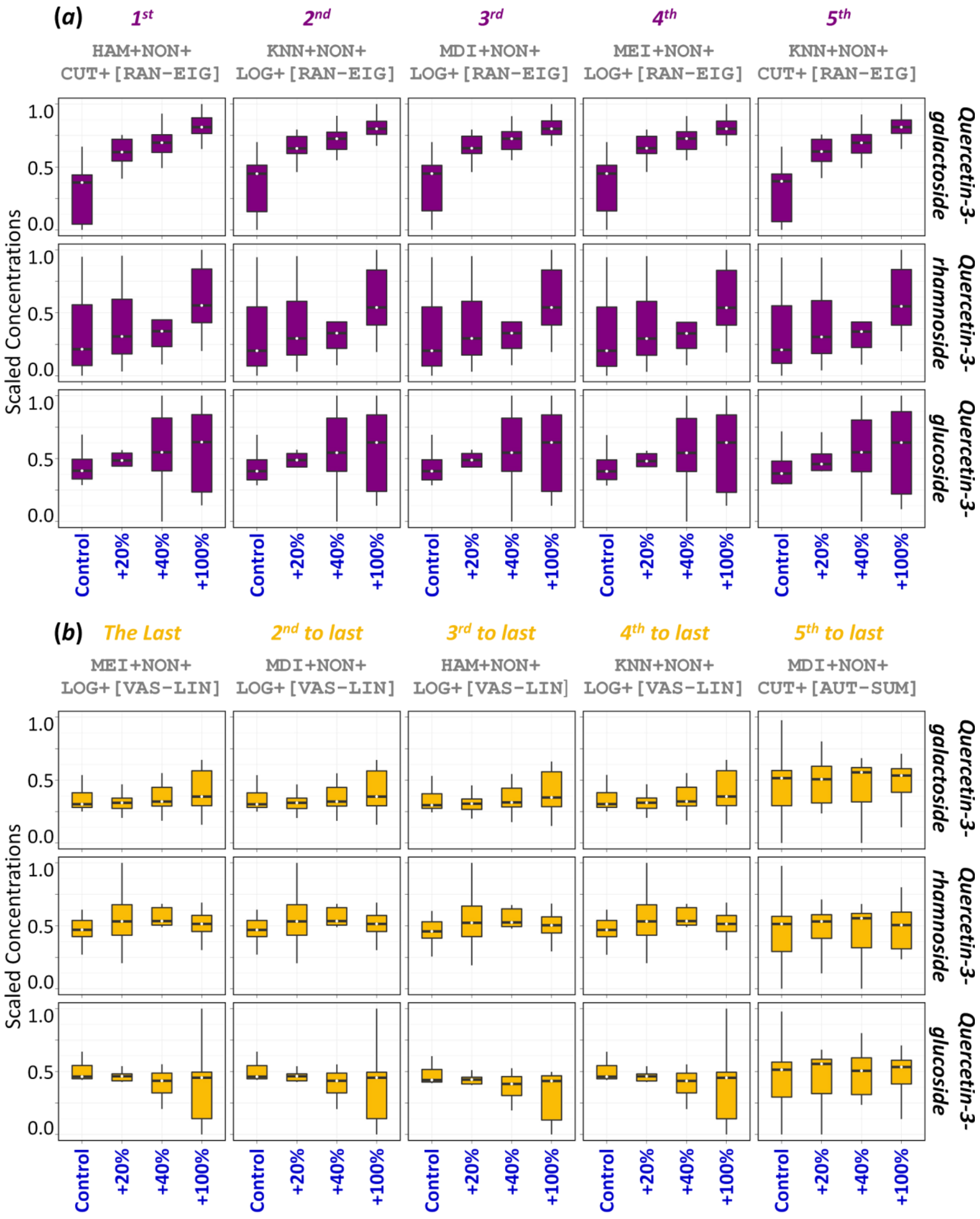


Figure S5. The processing outcomes of (***a***) five top-ranked and (***b***) five last-ranked workflows on three spike-in compounds (*quercetin-3-galactoside*, *quercetin-3-rhamnoside*, and *quercetin-3-glucoside*) spiked with the gradual increase of concentrations from the control to an increase of 20%, then 40%, and finally 100% (4). (***a***) five top-ranked workflows largely preserved those expected concentration variations of these spike-in compounds (from control to +20%, then to

+40%, and finally to +100%); (***b***) five last-ranked workflows could not reproduce those spiked concentration variations for any of those compounds. The corresponding processing workflows applied were indicated for all plots, and the detailed descriptions on the processing methods in these workflows can be found in **Table S4** and **Table S5**.

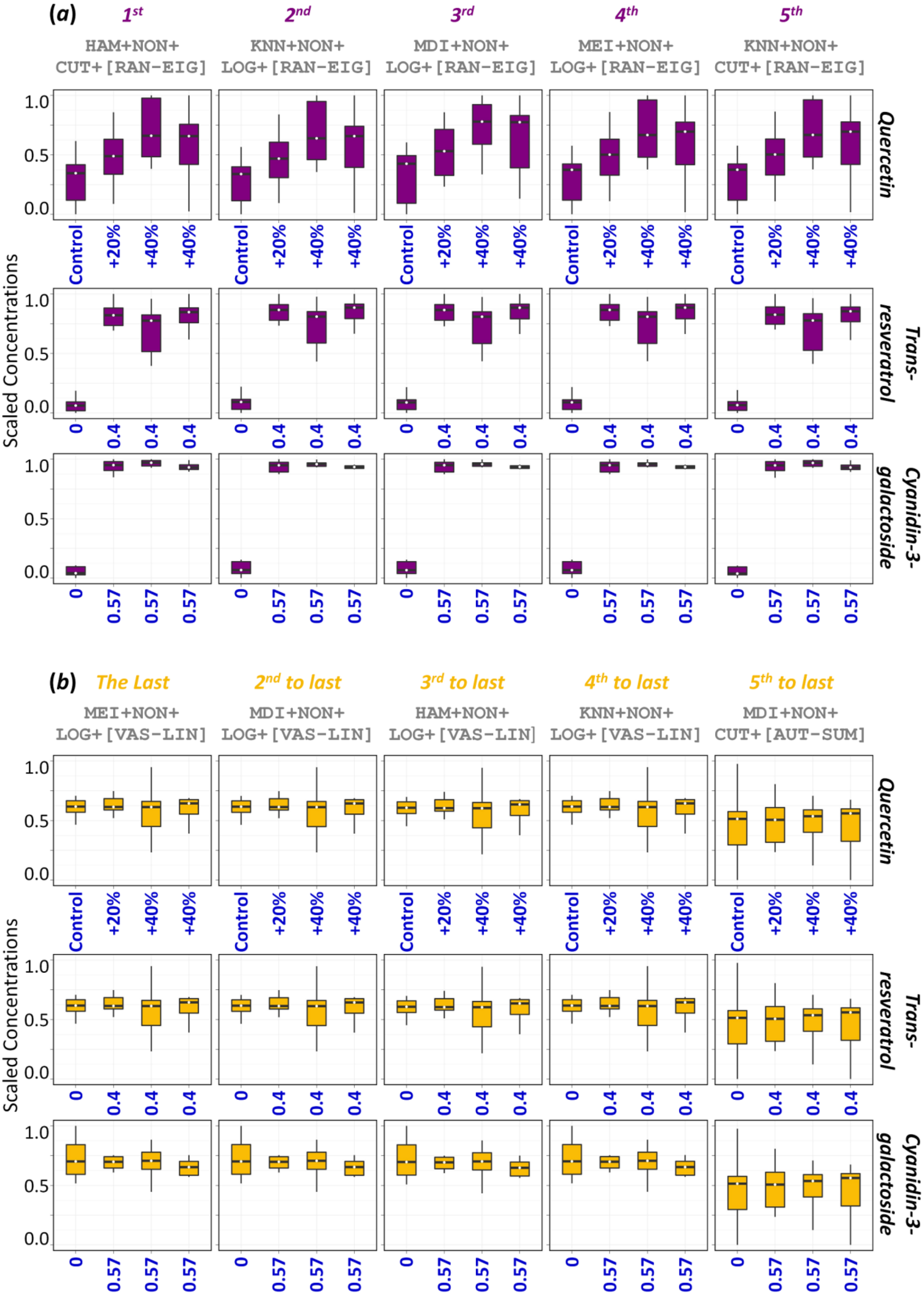


Figure S6. The processing outcomes of (***a***) five top-ranked and (***b***) five last-ranked workflows on three compounds (*quercetin*, *trans-resveratrol*, and *cyanidin-3-galactoside*). *Quercetin* was spiked with a concentration variation from control to an increase of 20%, then 40%, and finally 40% (4); *Trans-resveratrol*, and *Cyanidin-3-galactoside* were naturally unavailable in the first extract type, and spiked via constant concentration (0.4 and 0.57 mg/l, respectively) in another 3 extract types (4). (***a***) five top-ranked workflows largely preserved the expected concentration variations of these spike-in compounds; (***b***) five last-ranked workflows could not reproduce the spiked concentration changes for any of the compounds. Corresponding processing workflows applied were indicated for all plots, and the detailed descriptions on the processing methods in these workflows can be found in **Table S4** and **Table S5**.

Table S1. A variety of typical tools available for the pre-processing of spectral data acquired using various instruments of different vendors. NMR: nuclear magnetic resonance; MS: mass spectrometry. The detailed application of each tool in *spectral data pre-processing* procedure was provided in **Figure S1** (4 and 12 tools were applied to the sections of '*Conversion*' and '*Preparation of Peak Table*', respectively).

Data Pre-processing Tool	**Analytical Platform**	**Compatible Operating System**	**License (*vendor*)**	**Brief Description of Each Data Pre-processing Tool**
BATMAN	NMR	Linux MacOS Windows	General Public License	Preparing metabolomic peak table based on *Bayesian* model, which incorporates information on characteristic peak patterns of metabolites and is able to account for shifts in the position of peaks commonly seen in NMR spectra (5).
CompassXport	MS	Linux MacOS Windows	Apache License	Converting *Bruker* and some *Agilent* raw files to the universal mzXML format. These raw files include the following *Bruker* MS data file formats: analysis.baf, analysis.yep, *etc.*, and the *Agilent* MS data file format: analysis.yep (6)
Compound Discoverer	MS	Windows	Commercial (*Thermo Fisher*)	Preparing peak table for targeted/untargeted metabolomics, which comprises a workflow including peak picking, RT alignment, formula prediction, background annotation, and an automated library scanning for identification purposes (7).
LIMSA	MS	Linux Windows	General Public License	Preparing peak table for quantitative analysis of mass spectrometric metabolomic or lipidomic data, which sequentially carries out peak identification, integration, assignment, isotopic overlap correction, and quantification (8).
MarkerLynx	MS	Windows	Commercial (*Waters*)	Preparing the peak table for mass spectrometry-based metabolomics, which aims at conducting noise filtering, peak detection, raw data deconvolution, removal of isotope masses and alignment of the retention time (9).
MassHunter Profinder	MS	Windows	Commercial (*Agilent*)	Preparing the peak table from profiling and MSD data files, which is optimized to not only extract features from large datasets but also provides with an intuitive user interface to inspect and review each feature across the files (10).
MassLynx	MS	Windows	Apache License	Converting *Waters* and other raw files to mzXML format and controlling *Waters* mass spectrometers using integrated embedded PC technology, which acquires nominal mass, exact mass, MS/MS and exact mass MS/MS data (11).

MAVEN	MS	Linux MacOS Windows	General Public License	Preparing peak table for metabolomic quantitation from high-resolution full-scan mass spectrometry or multiple reaction monitoring datasets, which automatically detects and reports peak intensities for isotope-labeled metabolites (12).
MestreNova	NMR	Linux MacOS Windows	Commercial (*Mestrelab*)	Preparing peak table by aiming primarily at chemists dealing with 1D/2D spectra of small/midsized molecules, which covers all spectra pre-processing steps such as phase correction, baseline correction, peak alignment, and bucketing (13).
MZmine	MS	Linux MacOS Windows	General Public License	Preparing peak table for data processing of mass spectrometric metabolomic and proteomic data, which included noise reduction by filtering in chromatographic direction, cropping raw data range and removing scans by their width (14).
nmrML	NMR	MacOS Windows	MIT License	Converting the exchange syntax from the vendors' raw files into XSD-compliant nmrML by means of mappings from the *Bruker* 'acqus' or *Agilent* 'procpar' raw file to nmrML elements and controlled vocabulary terms (15).
NMRProcFlow	NMR	Linux MacOS Windows	General Public License	Preparing peak table for 1D spectra processing and metabolic fingerprinting of NMR metabolomic data, which covers all spectra pre-processing steps such as phase correction, baseline correction, peak alignment, and bucketing (16).
OpenMS	MS	Linux MacOS Windows	Berkeley Software Distribution	Preparing the peak table for addressing the most common tasks in quantitative metabolomics, which includes isotopic deconvolution, chromatographic peak-picking, RT alignment and feature-linking across multiple runs (17).
Progenesis QI	MS	Windows	Commercial (*Waters*)	Preparing the peak table for targeting the small molecule discovery analysis for metabolomics, which contains a series of pre-processing procedures including baseline correction, smoothing, deconvolution, and peak alignment (18).
Proteowizard	MS	Linux MacOS Windows	Apache License	Converting the obtained original MS data into mzXML format, which supports reading of mzML, mzXML and Thermo RAW files and provides a modular and extensible set of open-source, cross-platform tools and libraries (19).
XCMS	MS	Linux MacOS Windows	General Public License	Preparing peak table for targeted/untargeted LC-MS metabolomics by extracting metabolic features from raw MS data, which comprises chromatographic peak detection, sample alignment and peak correspondence (20).

Table S2. Representative tools available for the *statistical treatment & interpretation* of metabolomic data. ANOVA: analysis of variance; FC: fold change; HCA: hierarchical clustering analysis; *K*-means: *k*-means clustering; OPLS-DA: orthogonal partial least squares-discriminant analysis; PCA: principal component analysis; PLS-DA: partial least squares discriminant analysis; SAM: significance analysis for microarrays; SOM: self-organizing map; SVM-RFE: support vector machine-recursive feature elimination; WRS: wilcox rank sum test with permutation.

Tool	**Statistical Treatment**		**Interpretation**	
	No. of Methods for Sample Separation	**No. of Methods for Marker Identification**	**Availability of Pathway Analysis**	**Availability of Functional Enrichment**
KIMBLE (21)	≥2 (HCA, *etc.*)	≥1 (SVM-RFE)	NO	NO
MeltDB (22)	≥2 (PCA, *etc.*)	≥3 (ANOVA, *etc.*)	YES	YES
MetaboAnalyst (23)	≥4 (SOM, *etc.*)	≥11 (SAM, *etc.*)	YES	YES
Metabolomics Workbench (24)	≥2 (HCA, *etc.*)	≥5 (OPLS-DA, *etc.*)	YES	NO
metaP-server (25)	≥1 (PCA)	≥1 (Student's *t*-test)	YES	NO
metaX (26)	≥1 (PCA)	≥6 (PLS-DA, *etc.*)	YES	NO
MetDAT (27)	≥2 (HCA, *etc.*)	≥4 (FC, *etc.*)	YES	NO
MetFlow (28)	≥2 (HCA, *etc.*)	≥6 (WRS, *etc.*)	YES	NO
MMEASE (29)	≥4 (*K*-means, *etc.*)	≥13 (Relief, *etc.*)	YES	NO
muma (30)	≥1 (PCA)	≥4 (OPLS-DA, *etc.*)	NO	NO
W4M (31)	≥2 (HCA, *etc.*)	≥6 (FC, *etc.*)	NO	NO
WebSpecmine (32)	≥3 (*K*-means, *etc.*)	≥5 (SVM-RFE, *etc.*)	YES	NO

Table S3. Twenty representative studies that explicitly described the application of NOREVA in their metabolomic studies. These studies covered a very wide range of research fields, such as: *Microbiology*, *Molecular Biology*, *Pharmaceutical Science*, *Medical Science*, *Food & Environmental Science*, *Analytical Chemistry*, *Chromatography & Spectrometry*, and *Chemometrics* (for each field, 2~3 representative studies were described). CE-MS: capillary electrophoresis-mass spectrometry; GC-MS: gas chromatography-MS; LC-HR-MS/MS: liquid chromatography-high resolution-tandem MS; LC-HRMS: LC-high resolution MS; UPLC-HRMS: ultra-performance LC-high resolution MS; LC-MS/MS: LC-tandem MS.

Representative Publications	**The Application of NOREVA in Current Metabolomics as Described in Each Representative Publication**	**Research Field (*sub-field of research*)**	**Metabolomic Study Type**	**Platform**
(01) *Gut Microbes*. 11: 882, 2020	NOREVA was used to process the multi-class metabolomic data of 9 strains/combinations, which revealed the beneficial effects of microbiome regulation in the amelioration of non-alcoholic fatty liver disease (33).	Microbiology (*microbiome regulation*)	*Multi-class*	GC-MS
(02) *Front Microbiol*. 10: 1996, 2019	NOREVA was applied to process the metabolomic data that were generated based on the knockout of key biosynthetic gene, which revealed the mechanism of metabolite synthesis in the survival of a bacterial pathogen (34).	Microbiology (*metabolite biosynthesis*)	*Binary Classification*	UPLC-HRMS
(03) *Sci Rep*. 10: 17931, 2020	NOREVA was adopted to process time-course metabolomic data across 5 time-points, which helped the characterization of the bacterial and fungal diversity of three phyto-thermal baths performed in different months (35).	Microbiology (*microbiota dynamics*)	*Time-course*	GC-MS
(04) *Nucleic Acids Res*. 48: 385, 2020	NOREVA was employed as a data treatment module in the development of an interactive tool for single cell omics data interpretation, which facilitated the biological interpretation of single-cell multi-omics data by bench scientists (36).	Molecular Biology (*single-cell multi-omics*)	*Multi-class*	LC-MS GC-MS
(05) *Anal Chim Acta*. 1143: 124, 2021	NOREVA was recognized as one of the 'popular tools' for the processing of metabolomic data, which could transform complex raw data to a simplified data matrix, and reduced the effect from extreme outliers (37).	Molecular Biology (*single-cell metabolomics*)	*Multi-class* *Time-course*	LC-MS GC-MS

(06) *Aging Cell*. 19: e13213, 2020	NOREVA was utilized to remove systematic variations among samples in metabolomics study, which helped to discover the functional effect and metabolic profile of a drug combination for the treatment of cardiac aging (38).	Pharmaceutical Science (*drug synergistic efficacy*)	*Multi-class Time-course*	LC-MS/MS
(07) *Front Pharmacol*. 10: 127, 2019	NOREVA was used to process the metabolomic dataset for subsequent binary classification, which helped to assess the effectiveness of direct data merging strategy in long-term & large-scale pharmacometabonomics (39).	Pharmaceutical Science (*pharmaco-metabolomics*)	*Binary Classification*	LC-HRMS
(08) *J Proteome Res*. 19: 1913, 2020	NOREVA was adopted to process the metabolomic data and remove unwanted variations between samples, which assisted in determining whether urinary volatile terpenes levels could monitor breast cancer treatment efficacy (40).	Medical Science (*disease diagnosis*)	*Binary Classification*	GC-MS
(09) *Sci Rep*. 10: 16142, 2020	NOREVA was applied to process the metabolomic dataset for subsequent binary classification, which helped to discover the metabolic markers capable of predicting the development of a typical pregnancy complication (41).	Medical Science (*disease development*)	*Binary Classification*	LC-MS GC-MS
(10) *J Chromatogr B Analyt Technol Biomed Life Sci*. 1114: 119, 2019	NOREVA enabled the process of a lipidomic profile dataset for the subsequent binary classification, which facilitated the successful characterization of several novel lipidomic markers specific to the internet-gaming disorder (42).	Medical Science (*disease marker discovery*)	*Binary Classification*	LC-MS
(11) *Sci Total Environ*. 718: 137267, 2020	NOREVA facilitated the process of the metabolomic data in prior to binary classification, which helped to identify brain region-specific variations in metabolic pathway associated with the exposure to environmental ultrafine particles (43).	Food & Environmental Science (*environmental pollutant*)	*Binary Classification*	LC-MS GC-MS
(12) *LWT-Food Sci Technol*. 129: 109454, 2020	NOREVA realized the processing of a metabolomic dataset consisting of different sesame seeds from 6 countries, which quantified metabolic markers and enabled the evaluation of nutrition composition (44).	Food & Environmental Science (*nutrition composition*)	*Multi-class*	LC-MS GC-MS

(13) *Anal Chem*. 92: 203, 2020	NOREVA was regarded as one of the 'advanced tools' for the processing of metabolomic data, which facilitated the comparative evaluation of the performance of methods in processing the studied data matrix (45).	Analytical Chemistry (*instrument methodology*)	*Multi-class* *Time-course*	GC-MS CE-MS
(14) *Anal Chim Acta*. 1061: 60, 2019	NOREVA was used to check the effectiveness of a novel tool in improving the classification accuracy in metabolic marker discovery, which facilitated the removal of batch effects for large-scale untargeted metabolomics (46).	Analytical Chemistry (*quantitative analysis*)	*Binary Classification*	LC-MS GC-MS
(15) *Metabolomics*. 14: 54, 2018	NOREVA was employed as a data-processing module in the construction of a novel metabolomic tool, which allowed the visualization and comparative evaluation between different normalization algorithms (47).	Analytical Chemistry (*analytical software*)	*Multi-class* *Time-course*	LC-MS GC-MS
(16) *Anal Chem*. 91: 9836, 2019	NOREVA was considered to 'simplify' the investigation and selection of optimum processing workflow, which provided the platforms to perform and evaluate different normalization techniques on metabolomic dataset (48).	Chemometrics (*chemical calibration*)	*Multi-class* *Time-course*	LC-HR-MS/MS
(17) *Bioessays*. 40: e1700210, 2018	NOREVA was 'encouraged' to be applied for assessing the potential performance of different metabolomic processing methods on their empirical profiles, which could reflect the structure of empirical data in question (49).	Chemometrics (*cheminformatic pattern*)	*Multi-class* *Time-course*	GC-MS
(18) *Mass Spectrom Rev*. doi: mas.21672, 2020	NOREVA was considered as a 'user-friendly' metabolomic implementation with graphical interface for assessing the performances of both batch effect removal and biological information retention for mass spectrometry (MS) (50).	Chromatography & Spectrometry (*mass spectrometry*)	*Multi-class* *Time-course*	LC-MS GC-MS
(19) *Mass Spectrom Rev*. 40: 162, 2021	NOREVA was recognized as an MS-based batch processing service, which combined data-driven normalizations with internal standard/quality control-based methods and evaluated the performance for multiple testing (51).	Chromatography & Spectrometry (*mass spectrometry*)	*Multi-class* *Time-course*	LC-MS

(20) *Metabolites*. 9: 292, 2019	NOREVA was adopted to process the metabolomic data for subsequent binary classification, which helped to evaluate the performance of ammonium fluoride as additive salt in the hydrophilic interaction liquid chromatography (52).	Chromatography & Spectrometry (*liquid chromatography*)	*Binary Classification*	LC-HRMS

Table S4. A list of methods for data filtering, data imputation, quality control (QC) sample correction, and data transformation in the metabolomic *peak table processing*. A three-letter abbreviation (Abb.) code was used to represent each method of the different step (**Figure 1**) of metabolomics *peak table processing*. In a particular step of a processing workflow, if none of the above methods was applied, a three-letter code NON was used to indicate the non-application of any method in the corresponding step. Both method's introduction and research application(s) were provided.

Abb.	Method Name	Brief Introduction to Each Method and Its Application(s) in Metabolomic Studies
		Data Filtering
TPM	Tolerable Percent of Missing Values	**Method's Introduction**: This method calculates the percentage of missing values for each metabolite, and discards the one whose percent of missing values among all samples is higher than a threshold (53). **Research Application(s)**: It was used to filter the raw metabolomic data in hippocampal study for revealing neuroinflammatory factors in the transgenic hippocampus tissues of mice with *Alzheimer*'s disease (54).
TRS	Tolerance of Relative Standard Deviation	**Method's Introduction**: This method deletes the metabolite whose relative standard deviation (RSD) across all samples is higher than a threshold, since a lower RSD indicates a better reproducibility (55,56). **Research Application(s)**: It was integrated in a pipeline for the processing of both GC-MS and LC-MS open metabolomic profiling data based on the KNIME analytics platform (53).
		Data Imputation
HAM	Half of the Minimum Imputation	**Method's Introduction**: This method substitutes missing values with the half of the minimum value of non-missing values in the corresponding metabolites to reduce variation among experimental groups (57,58). **Research Application(s)**: It was adopted by the Q-TOF and HPLC-QqQ-MS metabolomics for identifying alterations of the exo-/endo-metabolite profiles in breast cancer cell lines (59).
KNN	*K*-nearest Neighbor Imputation	**Method's Introduction**: This method identifies *K* metabolites that are similar to the metabolite with missing value, and the missing values are imputed with the weighted average values of these neighboring ones (60). **Research Application(s)**: It was adopted in a metabolomic study, and facilitated the early identification of vincristine-induced peripheral neuropathy in pediatric leukemia patients (61).

MDI	Column Median Imputation	**Method's Introduction**: This method uses the median values, which are not easily affected by outliers, of non-missing values to impute those missing values in the corresponding metabolites (62). **Research Application(s)**: It was applied to discover the metabolic markers induced by metformin exposure and response, and to understand the metabolic mechanisms of metformin in ammonia detoxification (63).
MEI	Column Mean Imputation	**Method's Introduction**: This method replaces missing values with a median value of non-missing values in the corresponding metabolite, and tends to increase differences between diverse experimental groups (62). **Research Application(s)**: It was applied to a pharmacometabolomic assessment study for the discovery of novel drug response phenotypes of both atenolol and hydrochlorothiazide (64).
QC Sample Correction		
LLR	Local Linear Regression	**Method's Introduction**: This method corrects signals based on a local linear regression model which looks linear in small regions of input-space if the function has sufficient smoothness (65). **Research Application(s)**: It was utilized to process peak areas for quality control correction, and identify potential metabolic markers from a large-scale metabolomic dataset of hepatocellular carcinoma (66).
LPF	Local Polynomial Fits	**Method's Introduction**: This method is a nonparametric approach integrated in the QC-based LOESS signal correction (QC-RLSC) method for smoothing scatter plots and modeling functions (67). **Research Application(s)**: It was applied for metabolomic batch correction and the subsequent detection of significant metabolic variations in the athletes that were applied with growth hormone (68).
NWE	Nadaraya-Watson Estimator	**Method's Introduction**: This method provides a regression model which estimates the regression function by a weighted average of the raw data where the weights are a decreasing function of distance (69). **Research Application(s)**: It was integrated into some statistical *R* packages, which are the streamlined tools for signal drift correction and interpretations of quantitative MS-based metabolomic data (70).
Data Transformation		
CUT	Cube Root Transformation	**Method's Introduction**: This method increases the weight of metabolites of relatively lower concentrations and compresses the weight of metabolites of higher ones to an approximate normal distribution (71).

		Research Application(s): It was used to reveal metabolomic alterations in invasive ductal carcinoma of breast and help to identify diagnostic markers as well as potential therapeutic targets (72).
LOG	Log Transformation	**Method's Introduction**: This method tends to transform the distribution of metabolite abundance ratio to a more symmetrical (almost normal) distribution by minimizing the metabolites of extreme abundance (73). **Research Application(s)**: This method was utilized to enhance small signals in the metabolomics spectrum and facilitate the identification of metabolic markers for the early-stage diagnosis of oral cancer (74).

Table S5. A list of methods for data normalization in the metabolomic *peak table processing*. A three-letter abbreviation (Abb.) code was adopted to represent each method. If none of the methods was applied, a three-letter code NON was used to indicate the non-application of any method in normalization. For normalization method, there are three types of study assumption: (**SAα**) all metabolites should be equally important, which is the prerequisite for applying scaling methods (75-77); (**SAβ**) the level of metabolite intensity should be constant among all samples, which is the priori hypothesis for some normalizations, such as MED and SUM (78,79); and (**SAγ**) the intensity of the vast majority of the metabolites should be unchanged under the studied condition, which are demanded by some other normalization methods, such as PQN, VSN and QUA (78,80,81). It is essential to emphasize that due to the distinct assumptions, some methods are fundamentally inappropriate for certain dataset and thus cannot be assessed using NOREVA (81,82). Therefore, before any performance assessment, the nature of the studied dataset should be analyzed and whether the study assumption held for these data should be clarified.

Abb.	Method Name	Method's Introduction, Reported Applicable Domain(s), and Research Application(s)
		Sample-based Normalization
CON	Contrast	**Method's Introduction**: as a popular normalization, this method selects a baseline sample, to which other samples are normalized by fitting a nonlinear smooth curve (83,84). **Study Assumption *Gamma* (SAγ)**: the intensities of most metabolites are not changed under the studied conditions in the analyzed data (84). **Metabolomic Application**: it has been applied to improve data quality and remove unwanted variations in 1H NMR metabolite fingerprinting data in the case of unbalanced metabolite regulation (85).
CUB	Cubic Splines	**Method's Introduction**: this method aims to make the distribution of metabolite concentrations (geometric or arithmetic mean) among all samples comparable using the nonlinear baseline (84,86). **Study Assumption *Gamma* (SAγ)**: the intensities of most metabolites are not altered under the studied conditions in the analyzed data (87). **Metabolomic Application**: it has been used to eliminate unwanted biases and experimental variance for correctly classifying samples regardless of the dataset size (88).
EIG	EigenMS	**Method's Introduction**: this method is an adaptation of surrogate variable analysis, which identifies trends attributable to bias by utilizing singular value decomposition on model residuals (89,90).

		Study Assumption *Gamma* (SAγ): most of the metabolite intensities are not altered among samples (89), and this method reduces the sample-to-sample variations of unknown complexity (91,92). **Metabolomic Application**: it has been used to identify metabolomic biomarkers and dietary factors for characterizing the maternal metabolome with gestational diabetes (93).
LIN	Linear Baseline Scaling	**Method's Introduction**: this method maps linearly from each metabolite spectrum to a baseline through multiplying the metabolite intensities in all spectra using a particular scaling factor (84,94). **Study Assumption *Gamma* (SAγ)**: the intensities of most of the metabolites among samples are unchanged in the studied metabolomic dataset (95,96). **Metabolomic Application**: it facilitates the prediction of capecitabine-induced toxicity in patients with inoperable colorectal cancer based on pharmaco-metabonomic profiling (97).
LIW	Li-Wong	**Method's Introduction**: this method selects a baseline spectrum and normalizes other spectra by fitting a smooth curve on the level of metabolic feature intensities (83). **Study Assumption *Gamma* (SAγ)**: the intensities of the majority of the metabolites among samples are not changed in the studied data (95). **Metabolomic Application**: it has been utilized to eliminate unwanted sample-to-sample bias in 1H NMR metabolite fingerprinting datasets with unbalanced metabolite regulation (85).
LOE	Cyclic Loess	**Method's Introduction**: this method combines MA-plot and *Bland-Altman* plot by assuming the existence of non-linear bias (84), and it estimates a regression surface using multivariate smoothing procedure (98). **Study Assumption *Gamma* (SAγ)**: the majority of the intensities are unchanged in all samples (82,95), and the systematic bias nonlinearly depends on intensities (78). **Metabolomic Application**: it has been used in high throughput metabolomic studies to identify the underlying pathological mechanisms of the APP/PS1 model constructed for *Alzheimer*'s disease (99).
MEA	Mean Normalization	**Method's Introduction**: this method reduces variability among replicates by calculating the intensity of each metabolite in a given sample as the mean of intensities of all variables in samples (100,101). **Study Assumption *Beta* (SAβ)**: the mean level of intensities is consistent among all samples (78), and it ensures the metabolite intensity values in all samples comparable with each ones (102).

		Metabolomic Application: it has been utilized to normalize the pharmaco-metabolomics data and helped to differentiate L-carnitine outcomes in patients treated with septic shock (103).
MED	Median Normalization	**Method's Introduction**: this method removes unwanted variation among samples (101) by calculating the intensity of each metabolite in a given sample as the median of intensities of all variables in samples (78). **Study Assumption *Beta* (SAβ)**: the median level of intensities is consistent among all samples (78,102), and the metabolite intensity of each sample has the same median (78). **Metabolomic Application**: it has been applied to normalize metabolomic data generated by the quadrupole time-of-flight mass spectrometer for facilitating the analysis of human breath (104).
MST	MS Total Useful Signal	**Method's Introduction**: this method divides the intensity of each spectrum by the sum of intensities of all spectra, and makes metabolite intensities among all samples comparable (86,105). **Study Assumption *Beta* (SAβ)**: the level of metabolite intensity is constant among all samples by assuming that there is an equivalence between increased intensities and decreased intensities (86,106). **Metabolomic Application**: it has been used to correct the ionization efficiencies of the detected metabolite peaks and enhance accuracy for untargeted LC-MS based metabolomics data (107).
PQN	Probabilistic Quotient Normalization	**Method's Introduction**: this method integrally normalizes each spectrum and calculates a quotient between test and reference spectra, then all variables of the test spectrum are divided by the median quotient (108). **Study Assumption *Gamma* (SAγ)**: most metabolite intensities are unchanged among all samples (80), and it ensures the metabolite intensity values in all samples comparable with each ones (102). **Metabolomic Application**: it has been utilized in the 1H NMR-based metabolomics and identified as a robust method for complex biological mixtures attributing to various dilution concentration levels (108).
QUA	Quantile Normalization	**Method's Introduction**: this method replaces each point in the samples with the mean of the corresponding quantile and the distribution of the sample is made consistent on the basis of the sample quantile (94). **Study Assumption *Gamma* (SAγ)**: most metabolite intensities is unchanged among all samples (96), and it ensures the metabolite intensities in all samples comparable with each ones (102). **Metabolomic Application**: it was found as a well-performing normalization in processing 1D 1H urinary metabolomic data (109) and assisted the discovery of antihypertensive medication (110).

SUM	Total Sum Normalization	**Method's Introduction**: with the aim of reducing sample-to-sample variations, this method normalizes the metabolite intensities by assigning an appropriate weight to each sample (111). **Study Assumption *Beta* (SAβ)**: the average level of intensities is constant among all samples (112), and it ensures the metabolite intensity value in all samples comparable with each other (102). **Metabolomic Application**: it facilitates the identification of metabolites that associate with the different responses to gemcitabine-carboplatin chemotherapy in patients with metastatic breast cancer (113).
		Metabolite-based Normalization
AUT	Auto Scaling	**Method's Introduction**: this method is one of the simplest methods to adjust the metabolomic variances, which scales metabolite intensities based on the standard deviation of the metabolomic data (84,114). **Study Assumption *Alpha* (SAα)**: all metabolites are equally important (75), and it changes the emphasis from metabolites of high concentrations to those of moderate/small intensities (115,116). **Metabolomic Application**: it has been used in LC/MS-based metabolomics to facilitate the identification of urinary nucleosides as potential urogenital cancer markers (117).
LEV	Level Scaling	**Method's Introduction**: this method transforms the metabolic signal variations to that related to the mean metabolic signal, which are changed to values in percentages relative to the mean concentration (75). **Study Assumption *Alpha* (SAα)**: all metabolites are equally important, that is to say, metabolites with high intensities are not necessarily more important than those with low intensities (75). **Metabolomic Application**: it was applied to remove technical and biological variations in the UPLC-MS based untargeted metabolomic dataset, and to facilitate the investigation of differences in the liver metabolic profiles between distinct animal groups in the toxicology studies (118).
PAR	Pareto Scaling	**Method's Introduction**: this method uses the square root of the standard deviation of the data as the scaling factor, which can reduce the weight of a large fold change in metabolite intensities (84). **Study Assumption *Alpha* (SAα)**: all metabolites are equally important (75), and its disadvantage lines in its high sensitivity to the large fold changes (75). **Metabolomic Application**: it has been used to eliminate the mask effects in metabolomics and assisted the revealing of different responses to *Streptozotocin* and diet intervention in rat models (119).

POW	Power Scaling	**Method's Introduction**: this method aims at correcting the heteroscedasticity and pseudo-scaling through calculating the square root value of the metabolite intensity in different samples (75). **Study Assumption *Alpha* (SAα)**: all metabolites are equally important in the analyzed data, and it converts skewed metabolomics data to symmetric by non-linear transformation (75). **Metabolomic Application**: it has been used to facilitate the identification of serum metabolic changes and the investigation of their associations with colorectal cancers (120).
RAN	Range Scaling	**Method's Introduction**: this method scales the metabolite intensities for a systematic variance according to the intensity range of metabolites of all samples as the scaling factor (121). **Study Assumption *Alpha* (SAα)**: all metabolites are equally important (75), and it is usually applied for transforming the high concentration of metabolites to medium/small intensity (122). **Metabolomic Application**: it has helped to remove instrumental response factors from the metabolomics data and improve value comparability in prior to data fusion (121).
VAS	Vast Scaling	**Method's Introduction**: this method is an extension of auto scaling that focuses on stable variables and uses standard deviation and the so-called coefficient of variation as the scaling factor (75,123). **Study Assumption *Alpha* (SAα)**: all metabolites are equally important (75), and it is suitable for intensities of small fluctuations, but not suited for large variations without group structure (75). **Metabolomic Application**: it has been widely applied to both supervised and unsupervised learning from metabolomic data, and is discovered as a well-performing method for normalizing data and improving the multivariate models for metabolic feature selection and sample classification (114,124).
	Sample & metabolite-based Normalization	
VSN	Variance Stabilization Normalization	**Method's Introduction**: this method approaches the logarithm for large values to remove heteroscedasticity using the inverse hyperbolic sine (84), and keeps the variance constant over the entire data range (125). **Study Assumption *Gamma* (SAγ)**: most metabolites in different samples are not differentially expressed, and it makes the individual observations more directly comparable (126,127). **Metabolomic Application**: it has been adopted to metabolic profiling for processing the urine 1H NMR spectra signals with factors such as diseases, drugs and toxins (128).

Internal Standard-based Normalization		
CCM	Cross-contribution Compensating Multi-ISs Normalization	**Method's Introduction**: this method is capable of monitoring the systematic error and removing unwanted experimental variations based on multiple internal standards (129). **Study Assumption *Gamma* (SAγ)**: the intensities of most metabolites should not change under the studied conditions in the analyzed metabolomic dataset (129). **Metabolomic Application**: it has been applied to the MS-based metabolomics data from randomized and designed experiments that use internal standards to monitor the systematic error (129).
NOM	Normalization using Optimal Selection of Multiple ISs	**Method's Introduction**: this method removes unwanted systematic error via finding optimal normalization factor based on multiple internal standard compounds (130). **Study Assumption *Gamma* (SAγ)**: the intensities of most metabolites do not alter among samples, and it helps to remove unwanted systematic error in the analyzed data (129,131). **Metabolomic Application**: it has been applied to the UPLC/HRMS based mouse liver metabolomics data for removing the effect of systematic error across the full spectrum of metabolite peaks (130).
RUV	Remove Unwanted Variation-Random	**Method's Introduction**: this method utilizes the quality control metabolites (QCMs) that are only associated with unwanted variations to construct a linear mixed effects model for obtaining normalized data (111). **Study Assumption *Gamma* (SAγ)**: the intensities of most metabolites are assumed to be unchanged under the studied conditions in the analyzed metabolomic dataset (129). **Metabolomic Application**: it has facilitated the investigation of associations between metabolite patterns during late childhood and the exposure to maternal gestational diabetes mellitus (132).
SIS	Single Internal Standard	**Method's Introduction**: this method subtracts the log abundance of single internal standard from that of all metabolites in each sample of the analyzed dataset (101,133). **Study Assumption *Gamma* (SAγ)**: the intensities of most metabolites do not alter among samples, and this method is capable of removing unwanted variations in metabolomics data (129). **Metabolomic Application**: it has been integrated into a strategy which is applicable for dealing with large-scale human metabolomics studies, including data processing and validation (134).

Table S6. The time spent on applying NOREVA protocol measured by minute(s). Three metabolomics datasets of different sizes (PMID28528106, PMID21962342, and PMID22647087; as described in **Table 1**) are evaluated using four popular operating systems (CentOS, macOS, Ubuntu, and Windows). Hardware detail for time evaluation is shown under each system. Since the parallel computing together with its corresponding memory management are realized in NOREVA protocol, a comparison on time-cost between the application and non-application of 'parallel computing and memory management' is provided (indicated by 'YES' or 'NO', and measured by minutes).

Tested Datasets	**Application of Parallel Computing and Memory Management**	**CentOS Linux 7**	**macOS High Sierra**	**Ubuntu 20.04 LTS**	**Windows 10**
		2.5 GHz, 16 cores Xeon Platinum 8163	2.2 GHz, 4 cores Intel Core i5-8259U	3.0 GHz, 6 cores Intel Core i5-8500	2.9 GHz, 8 cores Intel Core i7-10700F
		64 GB RAM	8 GB RAM	16 GB RAM	16 GB RAM
PMID28528106 (135)	YES	~200 minutes	~600 minutes	~250 minutes	~200 minutes
	NO	~2,200 minutes	~3,300 minutes	~2,000 minutes	~2,100 minutes
PMID21962342 (4)	YES	~100 minutes	~400 minutes	~150 minutes	~100 minutes
	NO	~1,250 minutes	~2,100 minutes	~1,000 minutes	~1,100 minutes
PMID22647087 (136)	YES	~15 minutes	~60 minutes	~25 minutes	~15 minutes
	NO	~120 minutes	~160 minutes	~60 minutes	~70 minutes

Table S7. A comprehensive list of functions provided in NOREVA together with their descriptions. For each function, its argument names, default values and the allowable argument values are described. In total, 22 different functions are provided in NOREVA and discussed in this table.

(func1). Name of **Function 1**: ***PrepareInuputFiles***()

Description: this function enables the preparation and input of peak table which facilitate the subsequent application of other NOREVA functions. It could process not only a standardized format, but also the customized formats from available tools for peak table preparation (**Figure S1**).

Argument	Value Type	Description of the Argument and the Allowable Argument Values
dataformat	*numeric*	Allows the user to specify the FORMAT of their input peak table (default = *null*) “**1**” denotes a standardized format of peak table accepted by NOREVA “**2**” denotes the customized formats of peak table generated by 12 popular tools (such as XCMS)
rawdata	*character*	Allows the user to indicate the NAME of their input peak table file (default = *null*)
label	*character*	Allows the user to indicate the NAME of their input label file (default = *null*)

(func2). Name of **Function 2**: ***normulticlassqcall***()

Description: this function enables the performance assessment of metabolomic data processing for multi-class dataset (***with*** quality control sample but ***without*** internal standard) using four criteria, and can scan thousands of processing workflows and rank them based on their performances.

Argument	Value Type	Description of the Argument and the Allowable Argument Values
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)
SAalpha	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Alpha* (SAα, all metabolites are assumed to be equally important) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Alpha* (SAα) “**N**” denotes that the peak table does not satisfy the study assumption *Alpha* (SAα)

SAbeta	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Beta* (SAβ, the level of metabolite abundance is constant among all samples) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Beta* (SAβ) “**N**” denotes that the peak table does not satisfy the study assumption *Beta* (SAβ)
SAgamma	*character*	Allows the user to specify whether the input table satisfies study assumption *Gamma* (SAγ, the intensities of most metabolites are not changed under the studied conditions) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Gamma* (SAγ) “**N**” denotes that the peak table does not satisfy the study assumption *Gamma* (SAγ)

(*func3*). Name of **Function 3**: ***normulticlassnoall***()

Description: this function enables the performance assessment of metabolomic data processing for multi-class dataset (***without*** quality control sample and ***without*** internal standard) using four independent criteria, and can comprehensively scan thousands of processing workflows and rank all these workflows based on their performances (assessed from four different perspectives).

Argument	**Value Type**	**Description of the Argument and the Allowable Argument Values**
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)
SAalpha	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Alpha* (SAα, all metabolites are assumed to be equally important) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Alpha* (SAα) “**N**” denotes that the peak table does not satisfy the study assumption *Alpha* (SAα)
SAbeta	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Beta* (SAβ, the level of metabolite abundance is constant among all samples) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Beta* (SAβ) “**N**” denotes that the peak table does not satisfy the study assumption *Beta* (SAβ)

SAgamma	*character*	Allows the user to specify whether the input table satisfies study assumption *Gamma* (SAγ, the intensities of most metabolites are not changed under the studied conditions) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Gamma* (SAγ) “**N**” denotes that the peak table does not satisfy the study assumption *Gamma* (SAγ)

(*func4*). Name of **Function 4**: ***normulticlassisall***()

Description: this function enables the performance assessment of metabolomic data processing for multi-class dataset (***with*** internal standards but ***without*** quality control sample) using four criteria, and can scan thousands of processing workflows and rank them based on their performances.

Argument	**Value Type**	**Description of the Argument and the Allowable Argument Values**
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)
IS	*character*	Allows the user to indicate the column number(s) where the internal standard(s) locate (default = *null*) If there is only one internal standard (IS), the column number of this IS should be listed If there are multiple ISs, the column numbers of all ISs should be listed and separated using comma For example, the value of argument IS that is set to “2,6,8,n” indicates that the metabolites in the 3*rd*, 7*th*, 9*th*, and (n+1)*th* columns of your input peak table should be considered to be the IS metabolites.
SAalpha	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Alpha* (SAα, all metabolites are assumed to be equally important) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Alpha* (SAα) “**N**” denotes that the peak table does not satisfy the study assumption *Alpha* (SAα)
SAbeta	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Beta* (SAβ, the level of metabolite abundance is constant among all samples) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Beta* (SAβ) “**N**” denotes that the peak table does not satisfy the study assumption *Beta* (SAβ)

SAgamma	*character*	Allows the user to specify whether the input table satisfies study assumption *Gamma* (SAγ, the intensities of most metabolites are not changed under the studied conditions) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Gamma* (SAγ) “**N**” denotes that the peak table does not satisfy the study assumption *Gamma* (SAγ)

(*func5*). Name of **Function 5**: ***nortimecourseqcall***()

Description: this function enables the performance assessment of metabolomic data processing for the time-course dataset (***with*** quality control sample but ***without*** internal standard) using four independent criteria, and can comprehensively scan thousands of processing workflows and rank all these workflows based on their performances (assessed from four different perspectives).

Argument	**Value Type**	**Description of the Argument and the Allowable Argument Values**
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)
SAalpha	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Alpha* (SAα, all metabolites are assumed to be equally important) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Alpha* (SAα) “**N**” denotes that the peak table does not satisfy the study assumption *Alpha* (SAα)
SAbeta	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Beta* (SAβ, the level of metabolite abundance is constant among all samples) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Beta* (SAβ) “**N**” denotes that the peak table does not satisfy the study assumption *Beta* (SAβ)
SAgamma	*character*	Allows the user to specify whether the input table satisfies study assumption *Gamma* (SAγ, the intensities of most metabolites are not changed under the studied conditions) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Gamma* (SAγ) “**N**” denotes that the peak table does not satisfy the study assumption *Gamma* (SAγ)

(*func6*). Name of **Function 6**: ***nortimecoursenoall***()

Description: this function enables the performance assessment of metabolomic data processing for time-course dataset (***without*** quality control sample and ***without*** internal standard) using four independent criteria, and can comprehensively scan thousands of processing workflows and rank all these workflows based on their performances (assessed from four different perspectives).

Argument	**Value Type**	**Description of the Argument and the Allowable Argument Values**
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)
SAalpha	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Alpha* (SAα, all metabolites are assumed to be equally important) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Alpha* (SAα) “**N**” denotes that the peak table does not satisfy the study assumption *Alpha* (SAα)
SAbeta	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Beta* (SAβ, the level of metabolite abundance is constant among all samples) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Beta* (SAβ) “**N**” denotes that the peak table does not satisfy the study assumption *Beta* (SAβ)
SAgamma	*character*	Allows the user to specify whether the input table satisfies study assumption *Gamma* (SAγ, the intensities of most metabolites are not changed under the studied conditions) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Gamma* (SAγ) “**N**” denotes that the peak table does not satisfy the study assumption *Gamma* (SAγ)

(*func7*). Name of **Function 7**: ***nortimecourseisall***()

Description: this function enables the performance assessment of metabolomic data processing for time-course dataset (***with*** internal standards but ***without*** quality control sample) using four independent criteria, and can comprehensively scan thousands of processing workflows and rank all these workflows based on their performances (assessed from four different perspectives).

Argument	Value Type	Description of the Argument and the Allowable Argument Values
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)
IS	*character*	Allows the user to indicate the column number(s) where the internal standard(s) locate (default = *null*) If there is only one internal standard (IS), the column number of this IS should be listed If there are multiple ISs, the column numbers of all ISs should be listed and separated using comma For example, the value of argument IS that is set to “1,5,7,n” indicates that the metabolites in the 3*rd*, 7*th*, 9*th*, and (n+2)*th* columns of your input peak table should be considered to be the IS metabolites.
SAalpha	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Alpha* (SAα, all metabolites are assumed to be equally important) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Alpha* (SAα) “**N**” denotes that the peak table does not satisfy the study assumption *Alpha* (SAα)
SAbeta	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Beta* (SAβ, the level of metabolite abundance is constant among all samples) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Beta* (SAβ) “**N**” denotes that the peak table does not satisfy the study assumption *Beta* (SAβ)
SAgamma	*character*	Allows the user to specify whether the input table satisfies study assumption *Gamma* (SAγ, the intensities of most metabolites are not changed under the studied conditions) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Gamma* (SAγ) “**N**” denotes that the peak table does not satisfy the study assumption *Gamma* (SAγ)

(*func8*). Name of **Function 8**: ***normulticlassqcallgs***()

Description: this function enables the performance assessment of metabolomic data processing for multi-class dataset (***with*** quality control sample but ***without*** internal standard) using five independent criteria, and can comprehensively scan thousands of processing workflows and rank all these workflows based on their performances (assessed from five different perspectives).

Argument	Value Type	Description of the Argument and the Allowable Argument Values
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)
GS	*character*	Allows the user to indicate the name of the file that contains the spike-in compounds (default = *null*) The file should be in a .csv format, which provides the concentrations of spike-in compounds.
SAalpha	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Alpha* (SAα, all metabolites are assumed to be equally important) (default = "Y") "**Y**" denotes that the peak table satisfies the study assumption *Alpha* (SAα) "**N**" denotes that the peak table does not satisfy the study assumption *Alpha* (SAα)
SAbeta	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Beta* (SAβ, the level of metabolite abundance is constant among all samples) (default = "Y") "**Y**" denotes that the peak table satisfies the study assumption *Beta* (SAβ) "**N**" denotes that the peak table does not satisfy the study assumption *Beta* (SAβ)
SAgamma	*character*	Allows the user to specify whether the input table satisfies study assumption *Gamma* (SAγ, the intensities of most metabolites are not changed under the studied conditions) (default = "Y") "**Y**" denotes that the peak table satisfies the study assumption *Gamma* (SAγ) "**N**" denotes that the peak table does not satisfy the study assumption *Gamma* (SAγ)

(func9). Name of **Function 9**: ***normulticlassnoallgs***()

Description: this function enables the performance assessment of metabolomic data processing for multi-class dataset (***without*** quality control sample and ***without*** internal standard) using five criteria, and can scan thousands of workflows and rank them based on their performances.

Argument	Value Type	Description of the Argument and the Allowable Argument Values
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)

GS	*character*	Allows the user to indicate the name of the file that contains the spike-in compounds (default = *null*) The file should be in a .csv format, which provides the concentrations of spike-in compounds.
SAalpha	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Alpha* (SAα, all metabolites are assumed to be equally important) (default = "Y") "**Y**" denotes that the peak table satisfies the study assumption *Alpha* (SAα) "**N**" denotes that the peak table does not satisfy the study assumption *Alpha* (SAα)
SAbeta	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Beta* (SAβ, the level of metabolite abundance is constant among all samples) (default = "Y") "**Y**" denotes that the peak table satisfies the study assumption *Beta* (SAβ) "**N**" denotes that the peak table does not satisfy the study assumption *Beta* (SAβ)
SAgamma	*character*	Allows the user to specify whether the input table satisfies study assumption *Gamma* (SAγ, the intensities of most metabolites are not changed under the studied conditions) (default = "Y") "**Y**" denotes that the peak table satisfies the study assumption *Gamma* (SAγ) "**N**" denotes that the peak table does not satisfy the study assumption *Gamma* (SAγ)

(*func10*). Name of **Function 10**: ***normulticlassisallgs***()

Description: this function enables the performance assessment of metabolomic data processing for multi-class dataset (***with*** internal standards but ***without*** quality control sample) using five criteria, and can scan thousands of processing workflows and rank them based on their performances.

Argument	**Value Type**	**Description of the Argument and the Allowable Argument Values**
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)
GS	*character*	Allows the user to indicate the name of the file that contains the spike-in compounds (default = *null*) The file should be in a .csv format, which provides the concentrations of spike-in compounds.

IS	*character*	Allows the user to indicate the column number(s) where the internal standard(s) locate (default = *null*) If there is only one internal standard (IS), the column number of this IS should be listed If there are multiple ISs, the column numbers of all ISs should be listed and separated using comma For example, the value of argument IS that is set to "2,6,8,n" indicates that the metabolites in the 3*rd*, 7*th*, 9*th*, and (n+1)*th* columns of your input peak table should be considered to be the IS metabolites.
SAalpha	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Alpha* (SAα, all metabolites are assumed to be equally important) (default = "Y") "**Y**" denotes that the peak table satisfies the study assumption *Alpha* (SAα) "**N**" denotes that the peak table does not satisfy the study assumption *Alpha* (SAα)
SAbeta	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Beta* (SAβ, the level of metabolite abundance is constant among all samples) (default = "Y") "**Y**" denotes that the peak table satisfies the study assumption *Beta* (SAβ) "**N**" denotes that the peak table does not satisfy the study assumption *Beta* (SAβ)
SAgamma	*character*	Allows the user to specify whether the input table satisfies study assumption *Gamma* (SAγ, the intensities of most metabolites are not changed under the studied conditions) (default = "Y") "**Y**" denotes that the peak table satisfies the study assumption *Gamma* (SAγ) "**N**" denotes that the peak table does not satisfy the study assumption *Gamma* (SAγ)

(func11). Name of **Function 11**: ***nortimecourseqcallgs***()

Description: this function enables the performance assessment of metabolomic data processing for the time-course dataset (***with*** quality control sample but ***without*** internal standard) using five criteria, and can scan thousands of workflows and rank them based on their performances.

Argument	**Value Type**	**Description of the Argument and the Allowable Argument Values**
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)

GS	*character*	Allows the user to indicate the name of the file that contains the spike-in compounds (default = *null*) The file should be in a .csv format, which provides the concentrations of spike-in compounds.
SAalpha	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Alpha* (SAα, all metabolites are assumed to be equally important) (default = "Y") "**Y**" denotes that the peak table satisfies the study assumption *Alpha* (SAα) "**N**" denotes that the peak table does not satisfy the study assumption *Alpha* (SAα)
SAbeta	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Beta* (SAβ, the level of metabolite abundance is constant among all samples) (default = "Y") "**Y**" denotes that the peak table satisfies the study assumption *Beta* (SAβ) "**N**" denotes that the peak table does not satisfy the study assumption *Beta* (SAβ)
SAgamma	*character*	Allows the user to specify whether the input table satisfies study assumption *Gamma* (SAγ, the intensities of most metabolites are not changed under the studied conditions) (default = "Y") "**Y**" denotes that the peak table satisfies the study assumption *Gamma* (SAγ) "**N**" denotes that the peak table does not satisfy the study assumption *Gamma* (SAγ)

(*func12*). Name of **Function 12**: ***nortimecoursenoallgs***()

Description: this function enables the performance assessment of metabolomic data processing for time-course dataset (***without*** quality control sample and ***without*** internal standard) using five criteria, and can scan thousands of workflows and rank them based on their performances.

Argument	**Value Type**	**Description of the Argument and the Allowable Argument Values**
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)
GS	*character*	Allows the user to indicate the name of the file that contains the spike-in compounds (default = *null*) The file should be in a .csv format, which provides the concentrations of spike-in compounds.

SAalpha	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Alpha* (SAα, all metabolites are assumed to be equally important) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Alpha* (SAα) “**N**” denotes that the peak table does not satisfy the study assumption *Alpha* (SAα)
SAbeta	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Beta* (SAβ, the level of metabolite abundance is constant among all samples) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Beta* (SAβ) “**N**” denotes that the peak table does not satisfy the study assumption *Beta* (SAβ)
SAgamma	*character*	Allows the user to specify whether the input table satisfies study assumption *Gamma* (SAγ, the intensities of most metabolites are not changed under the studied conditions) (default = “Y”) “**Y**” denotes that the peak table satisfies the study assumption *Gamma* (SAγ) “**N**” denotes that the peak table does not satisfy the study assumption *Gamma* (SAγ)

(*func13*). Name of **Function 13**: ***nortimecourseisallgs***()

Description: this function enables the performance assessment of metabolomic data processing for time-course dataset (***with*** internal standards but ***without*** quality control sample) using five criteria, and can scan thousands of processing workflows and rank them based on their performances.

Argument	**Value Type**	**Description of the Argument and the Allowable Argument Values**
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)
IS	*character*	Allows the user to indicate the column number(s) where the internal standard(s) locate (default = *null*) If there is only one internal standard (IS), the column number of this IS should be listed If there are multiple ISs, the column numbers of all ISs should be listed and separated using comma For example, the value of argument IS that is set to “1,5,7,n” indicates that the metabolites in the 3*rd*, 7*th*, 9*th*, and (n+2)*th* columns of your input peak table should be considered to be the IS metabolites.

GS	*character*	Allows the user to indicate the name of the file that contains the spike-in compounds (default = *null*) The file should be in a .csv format, which provides the concentrations of spike-in compounds.
SAalpha	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Alpha* (SAα, all metabolites are assumed to be equally important) (default = "Y") "**Y**" denotes that the peak table satisfies the study assumption *Alpha* (SAα) "**N**" denotes that the peak table does not satisfy the study assumption *Alpha* (SAα)
SAbeta	*character*	Allows the user to specify whether the input peak table satisfies the study assumption *Beta* (SAβ, the level of metabolite abundance is constant among all samples) (default = "Y") "**Y**" denotes that the peak table satisfies the study assumption *Beta* (SAβ) "**N**" denotes that the peak table does not satisfy the study assumption *Beta* (SAβ)
SAgamma	*character*	Allows the user to specify whether the input table satisfies study assumption *Gamma* (SAγ, the intensities of most metabolites are not changed under the studied conditions) (default = "Y") "**Y**" denotes that the peak table satisfies the study assumption *Gamma* (SAγ) "**N**" denotes that the peak table does not satisfy the study assumption *Gamma* (SAγ)

(*func14*). Name of **Function 14**: ***normulticlassqcpart***()

Description: this function enables performance assessment of metabolomic data processing for multi-class dataset (***with*** quality control sample but ***without*** internal standard) using four criteria, and can scan the customized workflows selected by user and rank them based on their performances.

Argument	**Value Type**	**Description of the Argument and the Allowable Argument Values**
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)
selectedMethods	*character*	Allows the user to indicate the NAME of the file containing the customized workflows selected by user. The file should be in a .csv format, and the exemplar files are provided in the NOREVA *R* package and available for download at *https://idrblab.org/noreva/NOREVA_exampledata.zip*).

(*func15*). Name of **Function 15**: ***normulticlassnopart***()

Description: this function enables performance assessment of metabolomic data processing for multi-class dataset (***without*** quality control sample and ***without*** internal standard) using four criteria, and can scan the customized workflows selected by user and rank them based on performances.

Argument	Value Type	Description of the Argument and the Allowable Argument Values
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)
selectedMethods	*character*	Allows the user to indicate the NAME of the file containing the customized workflows selected by user. The file should be in a .csv format, and the exemplar files are provided in the NOREVA *R* package and available for download at *https://idrblab.org/noreva/NOREVA_exampledata.zip*).

(*func16*). Name of **Function 16**: ***normulticlassispart***()

Description: this function enables the performance assessment of metabolomic data processing for multi-class dataset (***with*** internal standards but ***without*** quality control sample) using four criteria, and can scan the customized workflows selected by user and rank them based on performances.

Argument	Value Type	Description of the Argument and the Allowable Argument Values
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)
IS	*character*	Allows the user to indicate the column number(s) where the internal standard(s) locate (default = *null*) If there is only one internal standard (IS), the column number of this IS should be listed If there are multiple ISs, the column numbers of all ISs should be listed and separated using comma For example, the value of argument IS that is set to “2,6,8,n” indicates that the metabolites in the 3*rd*, 7*th*, 9*th*, and (n+1)*th* columns of your input peak table should be considered to be the IS metabolites.
selectedMethods	*character*	Allows the user to indicate the NAME of the file containing the customized workflows selected by user. The file should be in a .csv format, and the exemplar files are provided in the NOREVA *R* package and available for download at *https://idrblab.org/noreva/NOREVA_exampledata.zip*).

(*func17*). Name of **Function 17**: ***nortimecourseqcpart***()

Description: this function enables performance assessment of metabolomic data processing for time-course data (***with*** quality control sample but ***without*** internal standard) using four criteria, and can scan the customized workflows selected by user and rank them based on their performances.

Argument	Value Type	Description of the Argument and the Allowable Argument Values
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)
selectedMethods	*character*	Allows the user to indicate the NAME of the file containing the customized workflows selected by user. The file should be in a .csv format, and the exemplar files are provided in the NOREVA *R* package.

(*func18*). Name of **Function 18**: ***nortimecoursenopart***()

Description: this function enables performance assessment of metabolomic data processing for time-course data (***without*** quality control sample and ***without*** internal standard) using four criteria, and can scan the customized workflows selected by user and rank them based on performances.

Argument	Value Type	Description of the Argument and the Allowable Argument Values
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)
selectedMethods	*character*	Allows the user to indicate the NAME of the file containing the customized workflows selected by user. The file should be in a .csv format, and the exemplar files are provided in the NOREVA *R* package.

(*func19*). Name of **Function 19**: ***nortimecourseispart***()

Description: this function enables the performance assessment of metabolomic data processing for time-course dataset (***with*** internal standards but ***without*** quality control sample) using four criteria, and can scan the customized workflows selected by user and rank them based on performances.

Argument	Value Type	Description of the Argument and the Allowable Argument Values
fileName	*character*	Allows the user to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)

IS	*character*	Allows the user to indicate the column number(s) where the internal standard(s) locate (default = *null*) If there is only one internal standard (IS), the column number of this IS should be listed If there are multiple ISs, the column numbers of all ISs should be listed and separated using comma For example, the value of argument IS that is set to “1,5,7,n” indicates that the metabolites in the 3*rd*, 7*th*, 9*th*, and (n+2)*th* columns of your input peak table should be considered to be the IS metabolites.
selectedMethods	*character*	Allows the user to indicate the NAME of the file containing the customized workflows selected by user. The file should be in a .csv format, and the exemplar files are provided in the NOREVA *R* package.

(*func20*). Name of **Function 20**: ***normulticlassmatrix***()

Description: based on a particular processing workflow (especially the one identified as well-performing for the studied metabolomic multi-class dataset), this function outputs the resulting levels of all metabolites among all samples after the data processing based on that workflow. Quality control sample (QCS) and internal standard (IS) are considered in this function.

Argument	**Value Type**	**Description of the Argument and the Allowable Argument Values**
datatype	*numeric*	Allows the users to specify the data type of their input peak table (default = *null*) “**1**” denotes the multi-class metabolomic dataset without QCSs and without ISs “**2**” denotes the multi-class metabolomic dataset with QCSs but without ISs “**3**” denotes the multi-class metabolomic dataset with ISs but without QCSs
fileName	*character*	Allows the users to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)
IS	*character*	Allows the user to indicate the column number(s) where the internal standard(s) locate (default = *null*) If there is only one internal standard (IS), the column number of this IS should be listed If there are multiple ISs, the column numbers of all ISs should be listed and separated using comma For example, the value of argument IS that is set to “2,6,8,n” indicates that the metabolites in the 3*rd*, 7*th*, 9*th*, and (n+1)*th* columns of your input peak table should be considered to be the IS metabolites.

impt	*numeric*	Allows the users to specify the NAME of the imputation method (default = 1) “**1**” denotes the method of MEI (*mean imputation*) “**2**” denotes the method of MDI (*median imputation*) “**3**” denotes the method of HAM (*half of the minimum imputation*) “**4**” denotes the method of KNN (*k-nearest neighbor imputation*)
qcsn	*numeric*	Allows the users to specify the NAME of the quality control sample correction method (default = 1) “**1**” denotes the method of NWE (*Nadaraya-Watson estimator*) “**2**” denotes the method of LLR (*local linear regression*) “**3**” denotes the method of LPF (*local polynomial fits*)
trsf	*numeric*	Allows the users to specify the NAME of the transformation method (default = 1) “**1**” denotes the method of CUT (*cube root transformation*) “**2**” denotes the method of LOG (*log transformation*) “**3**” denotes the NON application of any transformation method
nmal	*numeric*	Allows the users to specify the NAME of the normalization method (default = *null*) “**1**” denotes the NON application of any normalization method “**2**” denotes the sample-based method of PQN (*probabilistic quotient normalization*) “**3**” denotes the sample-based method of LOE (*cyclic loess*) “**4**” denotes the sample-based method of CON (*contrast*) “**5**” denotes the sample-based method of QUA (*quantile normalization*) “**6**” denotes the sample-based method of LIN (*linear baseline scaling*) “**7**” denotes the sample-based method of LIW (*Li-Wong*) “**8**” denotes the sample-based method of CUB (*cubic splines*) “**9**” denotes the metabolite-based method of AUT (*auto scaling*) “**10**” denotes the metabolite-based method of RAN (*range scaling*)

		“**11**” denotes the metabolite-based method of PAR (*pareto scaling*) “**12**” denotes the metabolite-based method of VAS (*vast scaling*) “**13**” denotes the metabolite-based method of LEV (*level scaling*) “**14**” denotes the sample & metabolite-based method of VSN (*variance stabilization normalization*) “**15**” denotes the metabolite-based method of POW (*power scaling*) “**16**” denotes the sample-based method of MST (*MS total useful signal*) “**17**” denotes the sample-based method of SUM (*total sum normalization*) “**18**” denotes the sample-based method of MED (*median normalization*) “**19**” denotes the sample-based method of MEA (*mean normalization*) “**20**” denotes the sample-based method of EIG (*EigenMS*)
nmal2	*numeric*	Allows the users to specify the NAME of the normalization method (default = *null*) “**1**” denotes the NON application of any normalization method “**2**” denotes the sample-based method of PQN (*probabilistic quotient normalization*) “**3**” denotes the sample-based method of LOE (*cyclic loess*) “**4**” denotes the sample-based method of CON (*contrast*) “**5**” denotes the sample-based method of QUA (*quantile normalization*) “**6**” denotes the sample-based method of LIN (*linear baseline scaling*) “**7**” denotes the sample-based method of LIW (*Li-Wong*) “**8**” denotes the sample-based method of CUB (*cubic splines*) “**9**” denotes the metabolite-based method of AUT (*auto scaling*) “**10**” denotes the metabolite-based method of RAN (*range scaling*) “**11**” denotes the metabolite-based method of PAR (*pareto scaling*) “**12**” denotes the metabolite-based method of VAS (*vast scaling*) “**13**” denotes the metabolite-based method of LEV (*level scaling*) “**14**” denotes the sample & metabolite-based method of VSN (*variance stabilization normalization*)

		"**15**" denotes the metabolite-based method of POW (*power scaling*) "**16**" denotes the sample-based method of MST (*MS total useful signal*) "**17**" denotes the sample-based method of SUM (*total sum normalization*) "**18**" denotes the sample-based method of MED (*median normalization*) "**19**" denotes the sample-based method of MEA (*mean normalization*) "**20**" denotes the sample-based method of EIG (*EigenMS*) In combination with the argument *nmal* above, this argument (*nmal2*) helps to enable a normalization of combination strategy. Particularly, if a sample-based normalization is selected for the argument *nmal*, a metabolite-based one should be chosen for this argument (*nmal2*). If a metabolite-based normalization is selected for the argument *nmal*, a sample-based one should be chosen for this argument (*nmal2*).
nmals	*numeric*	Allows the users to specify the NAME of the IS-based normalization method (default = *null*) "**1**" denotes the method of SIS (*single internal standard*) "**2**" denotes the method of NOM (*normalization using optimal selection of multiple ISs*) "**3**" denotes the method of CCM (*cross-contribution compensating multi-ISs normalization*) "**4**" denotes the method of RUV (*remove unwanted variation-random*)

(*func21*). Name of **Function 21**: ***nortimecoursematrix***()

Description: based on a particular processing workflow (especially the one identified as well-performing for the studied metabolomic time-course dataset), this function outputs the resulting levels of all metabolites among all samples after the data processing based on that workflow.

Argument	**Value Type**	**Description of the Argument and the Allowable Argument Values**
datatype	*numeric*	Allows the users to specify the data type of their input peak table (default = *null*) "**1**" denotes the multi-class metabolomic dataset without QCSs and without ISs "**2**" denotes the multi-class metabolomic dataset with QCSs but without ISs "**3**" denotes the multi-class metabolomic dataset with ISs but without QCSs

fileName	*character*	Allows the users to indicate the NAME of peak table resulted from *PrepareInuputFiles*() (default = *null*)
IS	*character*	Allows the user to indicate the column number(s) where the internal standard(s) locate (default = *null*) If there is only one internal standard (IS), the column number of this IS should be listed If there are multiple ISs, the column numbers of all ISs should be listed and separated using comma For example, the value of argument IS that is set to “1,5,7,n” indicates that the metabolites in the 3*rd*, 7*th*, 9*th*, and (n+2)*th* columns of your input peak table should be considered to be the IS metabolites.
impt	*numeric*	Allows the users to specify the NAME of the imputation method (default = 1) “**1**” denotes the method of MEI (*mean imputation*) “**2**” denotes the method of MDI (*median imputation*) “**3**” denotes the method of HAM (*half of the minimum imputation*) “**4**” denotes the method of KNN (*k-nearest neighbor imputation*)
qcsn	*numeric*	Allows the users to specify the NAME of the quality control sample correction method (default = 1) “**1**” denotes the method of NWE (*Nadaraya-Watson estimator*) “**2**” denotes the method of LLR (*local linear regression*) “**3**” denotes the method of LPF (*local polynomial fits*)
trsf	*numeric*	Allows the users to specify the NAME of the transformation method (default = 1) “**1**” denotes the method of CUT (*cube root transformation*) “**2**” denotes the method of LOG (*log transformation*) “**3**” denotes the NON application of any transformation method
nmal	*numeric*	Allows the users to specify the NAME of the normalization method (default = *null*) “**1**” denotes the NON application of any normalization method “**2**” denotes the sample-based method of PQN (*probabilistic quotient normalization*) “**3**” denotes the sample-based method of LOE (*cyclic loess*)

		“**4**” denotes the sample-based method of CON (*contrast*) “**5**” denotes the sample-based method of QUA (*quantile normalization*) “**6**” denotes the sample-based method of LIN (*linear baseline scaling*) “**7**” denotes the sample-based method of LIW (*Li-Wong*) “**8**” denotes the sample-based method of CUB (*cubic splines*) “**9**” denotes the metabolite-based method of AUT (*auto scaling*) “**10**” denotes the metabolite-based method of RAN (*range scaling*) “**11**” denotes the metabolite-based method of PAR (*pareto scaling*) “**12**” denotes the metabolite-based method of VAS (*vast scaling*) “**13**” denotes the metabolite-based method of LEV (*level scaling*) “**14**” denotes the sample & metabolite-based method of VSN (*variance stabilization normalization*) “**15**” denotes the metabolite-based method of POW (*power scaling*) “**16**” denotes the sample-based method of MST (*MS total useful signal*) “**17**” denotes the sample-based method of SUM (*total sum normalization*) “**18**” denotes the sample-based method of MED (*median normalization*) “**19**” denotes the sample-based method of MEA (*mean normalization*) “**20**” denotes the sample-based method of EIG (*EigenMS*)
nmal2	*numeric*	Allows the users to specify the NAME of the normalization method (default = *null*) “**1**” denotes the NON application of any normalization method “**2**” denotes the sample-based method of PQN (*probabilistic quotient normalization*) “**3**” denotes the sample-based method of LOE (*cyclic loess*) “**4**” denotes the sample-based method of CON (*contrast*) “**5**” denotes the sample-based method of QUA (*quantile normalization*) “**6**” denotes the sample-based method of LIN (*linear baseline scaling*) “**7**” denotes the sample-based method of LIW (*Li-Wong*)

		“**8**” denotes the sample-based method of CUB (*cubic splines*) “**9**” denotes the metabolite-based method of AUT (*auto scaling*) “**10**” denotes the metabolite-based method of RAN (*range scaling*) “**11**” denotes the metabolite-based method of PAR (*pareto scaling*) “**12**” denotes the metabolite-based method of VAS (*vast scaling*) “**13**” denotes the metabolite-based method of LEV (*level scaling*) “**14**” denotes the sample & metabolite-based method of VSN (*variance stabilization normalization*) “**15**” denotes the metabolite-based method of POW (*power scaling*) “**16**” denotes the sample-based method of MST (*MS total useful signal*) “**17**” denotes the sample-based method of SUM (*total sum normalization*) “**18**” denotes the sample-based method of MED (*median normalization*) “**19**” denotes the sample-based method of MEA (*mean normalization*) “**20**” denotes the sample-based method of EIG (*EigenMS*) In combination with the argument *nmal* above, this argument (*nmal2*) helps to enable a normalization of combination strategy. Particularly, if a sample-based normalization is selected for the argument *nmal*, a metabolite-based one should be chosen for this argument (*nmal2*), and vice versa.
nmals	*numeric*	Allows the users to specify the NAME of the IS-based normalization method (default = *null*) “**1**” denotes the method of SIS (*single internal standard*) “**2**” denotes the method of NOM (*normalization using optimal selection of multiple ISs*) “**3**” denotes the method of CCM (*cross-contribution compensating multi-ISs normalization*) “**4**” denotes the method of RUV (*remove unwanted variation-random*)

(*func22*). Name of **Function 22**: ***norvisualization***()

Description: this function enables the visualization of the overall ranking of all processing workflows using a circular bar plot.

Argument	**Value Type**	**Description of the Argument and the Allowable Argument Values**

data	*character*	Allows the users to specify the NAME of the resulting ranking file that is generated based on NOREVA functions. These functions include *normulticlassnoall*(), *normulticlassqcall*(), *nortimecourseqcall*(), *nortimecoursenoall*(), *et al* (default = *null*).
cutoff	*numeric*	Allows the users to specify the number of the top-ranked processing workflows that are selected by users to be displayed in the circular bar plot (default = 100).
outputtype	*character*	Allows the users to specify the type of output file, and a string indicating the file types of .pdf, .png, .jpg, and .eps are supported in NOREVA (default = "pdf").
outputfile	*character*	Allows the user to specify the NAME of output file (default = *NOREVA-Ranking-Top.%d.workflows.%s*). "%d" is the value of the argument *cutoff* above, and "%s" is the value of the argument *outputtype* above.
maxValue	*numeric*	Allows the users to specify the maximum length of multiple bars, which indicates the performances of the processing workflow under multiple criteria, in the inner layers of the bar plot (default = 40).
colorSet	*character*	Allows the users to specify the colors of the multiple bars that indicate the performances of the workflow under multiple criteria (default = "#FFE699", "#D3E8C7", "#B2B2FF", "#FFCACA").
totalAngle	*numeric*	Allows the users to specify the degree of total angle of rotation of the bar plot (default = 340).
bgColor	*character*	Allows the users to specify the background color of the bar plot (default = "#FFFFFF").
fontColor	*character*	Allows the users to specify the front color of the bar plot (default = "#000000").

Table S8. A comprehensive description on the output files of the software package developed in this study. In total, 13 outputs are generated after the running of NOREVA, which are all provided and described below.

File Name	File Type	Description of the Corresponding Output File
OUTPUT-NOREVA-Overall.Ranking.Data.csv	CSV File	A CSV file providing all results discovered by NOREVA which includes the details of the assessing results, selected criteria. overall ranking, and ranking under each criterion.
NOREVA-Ranking-Top.XXX.workflows.pdf	PDF/EPS/PNG/JPG File	A circular bar plot illustrating the performance and the overall ranking of all processing workflows based on the multiple criteria or a single criterion that are selected by user.
OUTPUT-NOREVA-All.Normalized.Data.Rdata	RDATA File	A RDATA file that provides the resulting outcomes of all processing workflows.
OUTPUT-NOREVA-Criteria.Ca	Folder	A folder of various PDF files that illustrate the performance (using PMAD plot showing intensities among replicates) of each processing workflow assessed by Criterion *Ca.*
OUTPUT-NOREVA-Criteria.Ca.Rdata	RDATA File	A RDATA file that provides the performance (illustrated by intensities among replicates of PMAD plot) of each processing workflow assessed by Criterion *Ca.*
OUTPUT-NOREVA-Criteria.Cb	Folder	A folder of various PDF files that illustrate the performance (using *k*-means clustering between distinct groups) of each processing workflow assessed by Criterion *Cb.*
OUTPUT-NOREVA-Criteria.Cb.Rdata	RDATA File	A RDATA file that provides the performance (shown by the purity of *k*-means clustering ability between distinct groups) of each processing workflow assessed by Criterion *Cb.*
OUTPUT-NOREVA-Criteria.Cc	Folder	A folder of various PDF files that illustrate the performance (using Venn diagram for the marker overlap) of each processing workflow assessed by Criterion *Cc.*
OUTPUT-NOREVA-Criteria.Cc.Rdata	RDATA File	A RDATA file that provides the performance (illustrated by the CWrel value of marker overlap) of each processing workflow assessed by Criterion *Cc.*
OUTPUT-NOREVA-Criteria.Cd	Folder	A folder of various PDF files that illustrate the performance (using the marker classification) of each processing workflow assessed by Criterion *Cd.*

OUTPUT-NOREVA-Criteria.Cd.Rdata	RDATA File	A RDATA file that provides the performance (illustrated by the AUC value for marker classification) of each processing workflow assessed by Criterion *Cd.*
OUTPUT-NOREVA-Criteria.Ce	Folder	A folder of various PDF files that illustrate the performance (using the concentrations of known spike-in compounds) of each processing workflow assessed by Criterion *Ce.*
OUTPUT-NOREVA-Criteria.Ce.Rdata	RDATA File	A RDATA file that provides the performance (illustrated by difference between data and expected logFCs) of each processing workflow assessed by Criterion *Ce*.

Method S1. Processing workflows generated by method combination.

1. ***For the Metabolomic Data with Quality Control (QC) Samples***

Each processing workflow is composed of four sequential steps (S1-S4, as shown in **Figure 1**). For metabolomic data with QC samples, a random, comprehensive, and sequential integration among 4 imputation, 3 QC sample correction, 3 transformation (taking non-transformation into consideration), and 164 normalization (144 combined normalization between sample-based and metabolite-based ones, 19 sample/metabolite/sample & metabolite-based normalization shown in **Table S5**, and 1 non-normalization), can result in a total of **5,904** processing workflows.

2. ***For the Metabolomic Data with Internal Standards (ISs)***

Each processing workflow is composed of four sequential steps (S1-S4, as shown in **Figure 1**). For metabolomic data with ISs, a random, comprehensive, and sequential integration among 4 imputation, non-QC sample correction, 3 transformation (including non-transformation), and 4 IS-based normalization (**Table S5**), can result in a total of **48** processing workflows.

3. ***For the Metabolomic Data without QC Samples and ISs***

Each processing workflow is composed of four sequential steps (S1-S4, as shown in **Figure 1**). For metabolomic data without QC samples and ISs, a random, comprehensive, and sequential integration among 4 imputation, non-QC sample correction, 3 transformation (including non-transformation), and 164 normalization (144 combined normalization between sample-based and metabolite-based ones, 19 sample/metabolite/sample & metabolite-based normalization shown in **Table S5**, and 1 non-normalization), can result in a total of **1,968** workflows.

Method S2. Key steps in metabolomic data processing.

1. ***Data Filtering***

Data filtering aims to remove technical/mathematical uninformative features from the datasets of metabolomics (137). It is realized by setting a pre-specified cutoff based on the intrinsic properties of the metabolomic data, such as the ratio of missing values within the analyzed data (138) and the feature variability among all samples (139). In this protocol, 2 representative data filtering methods are provided, including the TPMV (tolerable percent of missing values) and TRSD (tolerance of relative standard deviation). The former filters metabolites when the ratio of missing values of that metabolite is larger than pre-specified cutoff (138). The latter exerts its effects when the relative standard deviation (the absolute measurement of batch-to-batch variation) of the metabolite among all samples is larger than pre-specified cutoff (139).

2. ***Data Imputation***

Data imputation tends to replace the missing values arising from biological or technical reasons with specific values based on the existing information (e.g., baseline signals or zero values). It accounts for the missingness and reduces the bias while keeping the data structure unchanged (140). In this protocol, 4 representative data imputation methods are provided, including the HAM (half of the minimum imputation), KNN (k-nearest neighbor imputation), MDI (column median imputation) and MEI (column mean imputation). HAM replaces the missing values with the half of the minimum positive values (58). KNN performs imputation by determining k metabolites of interest that are similar to the metabolites containing missing values (141). MDI and MEI replace the missing values with the median and mean value of non-missing values of the corresponding metabolite, respectively (142).

3. ***QC Sample Correction***

QC sample correction targets to correct for signal intensity variations, quality accuracy drifts, and intra- and inter-batch variability based on QC samples (pooled sample mixture by mixing small and equal aliquots from the real samples of interest and are dispersed evenly across the multiple batches to ensure the data quality) (143). It reduces the interference of harmful and uncontrollable signals within metabolomic data and ensures data consistency (144). In this protocol, a representative QC correction strategy named ‘QC-RLSC’ is provided, which is powerful in evaluating signal drifts and other systematic noise using mathematical algorithms (70). Particularly, 3 distinct regression models of the QC-RLSC algorithm including Nadaraya-Watson estimator, local linear regression and local polynomial fits are available here.

4. *Data Transformation*

Data transformation performs the nonlinear conversions of the metabolomic data for correcting heteroscedasticity, for converting multiplicative relations to additive relations, and for making skewed distributions more symmetric (75). It reduces the influence of disturbing factors such as measurement noise by converting the data into different scales (75). In this protocol, 2 representative transformation methods are provided, including CUT (cube root transformation) and LOG (log transformation). The former can improve normality distribution of simple count data by increasing the weight of metabolites of relatively lower concentration and compressing the weight of metabolites of higher ones (145). The latter performs nonlinear data conversions to decrease heteroscedasticity and get symmetric distribution prior to statistical analysis (146).

5. *Data Normalization*

Data normalization adjusts values or statistical distributions of the metabolomic data to remove unwanted variations while preserving biological variability (130). In this protocol, 4 categories of data normalization methods are provided, including sample-based, metabolite-based, sample & metabolite-based and IS-based normalization. Sample-based normalization tends to reduce systematic biases among samples and to make the data from all samples directly comparable to each other (147), while metabolite-based method aims to eliminate the effect of very large metabolite intensities and to make metabolites more comparable or normally distributed (109). Particularly, VSN (variance stabilization normalization), which is a sample & metabolite-based normalization, combines variance stabilization and between-sample normalization by transforming the data through a nonlinear approach to keep the variance of the dataset constant (84). Finally, IS-based normalization removes undesired data fluctuations by using IS(s), which is ideally stable isotopically labelled compounds introduced while sample processing and can be easily distinguished from endogenous metabolites (148). All in all, 12 sample-based, 6 metabolite-based, 1 sample & metabolite-based and 4 IS-based methods are provided here.

Method S3. Performance assessment from multiple perspectives.

As the processing performance of different processing workflows varies considerably and the fact that single criterion is not feasible and sufficient to ensure performance assessment from various perspectives (102), performance assessment conducts quantitative/qualitative evaluation of workflow using multiple criteria (131). In this protocol, five well-established criteria (131) with independent underlying theories are employed. Under each criterion, one specific measure is selected as representative, and a variety of well-defined cutoffs of this measure are used to categorize the processing performance into *Good*, *Fair* and *Poor*.

Criterion C*a*: workflow's ability to reduce intragroup variation among samples.

The PMAD (pooled median absolute deviation) is a frequently used metric under this criterion, which is integrated in several popular pipelines such as NormalyzerDE (149). The lower value of PMAD, the larger biological/experimental induced intragroup variation among samples that is removed by the workflow (150). As reported, PMAD values that fall in the range of ≤0.3, ≤0.7 & >0.3 and >0.7 denote *Good, Fair and Poor* performance, respectively (78,151,152).

Criterion C*b*: workflow's effect on differential metabolic analysis.

To meet the needs of time-course/multi-class metabolomics, the multivariate empirical Bayes statistics and OPLS-DA (orthogonal partial least squares-discriminant analysis) are employed. Specifically, the multivariate empirical Bayes statistics (153) selects time-course metabolic markers using HotellingT2 statistics (154). For multi-class metabolomics, OPLS-DA calculates the number of orthogonal components via cross-validation (155). The parameter of 'crossvalI', 'orthoI' and 'predI' in opls function are set to 'two', 'NA' and 'one', respectively. That is to say, opls function automatically computes the number of orthogonal components and optimizes it based on two-fold cross-validation and one predictive component (155). As a result, differential metabolic markers are selected by setting the VIP (variable influence on projection) value >1 (156) and considered as markers among classes in *K*-means clustering to indicate differential classes among samples (157). Finally, the well-defined metric (*purity*) is measured based on the differential metabolic markers and is chosen as the representative metric for this criterion. *Purity* is a direct metric for assessing clustering performance whose value falls in the range of 0 to 1. The closer the values of *purity* to 0, the poorer the clustering outcome. The closer the value to 1, the better the clustering performance provided by the studied workflow (158,159). *Purity* value of a specific processing workflow within the range of >0.8, ≤0.8 & >0.5 and ≤0.5 are generally accepted as *Good, Fair and Poor* performances, respectively (159).

Criterion C*c*: workflow's consistency in markers discovered from different datasets.

In consideration of the low reproducibility among different sets of markers selected from the same metabolomic dataset by different workflows (160,161), the evaluation of the consistency among workflows is thus adopted in NOREVA and considered as essential assessing criterion (131). First, time-course/multi-class metabolomic data are evenly divided into three subsets using stratified random selection (162,163) and then the selection of differential metabolic markers provided in Criterion C*b* is implemented to each sub-dataset. As expected, three sets of markers are selected from the corresponding sub-datasets and are always somewhat at variance. In this case, CW*rel*, a well-established metric which was discovered powerful in evaluating the consistency among subset-size-biased subsets (164), is now integrated in this criterion. To be more specific, CW*rel* calculates the number of times each feature appears in each single set of markers, which denotes the robustness and reproducibility among selected markers from an overall perspective (164). The value of CW*rel* fluctuates between 0 and 1 in which a larger value refers higher robustness and reproducibility of the different sets of the selected markers (164). Particularly, the value of CW*rel* within the range of >0.3, ≤0.3 & >0.15 and ≤0.15 represent *Good, Fair and Poor* performances, respectively (164).

Criterion C*d*: workflow's influence on classification accuracy.

The general goal of analyzing time-course/multi-class metabolomic data is to identify a variety of markers that can be validated as accurate for revealing biological dynamics or differentiating diverse classes (165,166). In this case, the workflow's influence on classification accuracy is thus evaluated using area under the curve (AUC) and receiver operating characteristic (ROC) analysis (167). This evaluation metric under Criterion C*d* is implemented by three steps, which involves: (1) the identification of the differential metabolic markers as described in Criterion C*b*; (2) the construction of a multiple classification model which is based on SVM (support vector machine) in e1071 *R* package; (3) the calculation of AUC value by running the multi roc function employed in multiROC *R* package. The parameters of 'type', 'kernel' and 'cross' are set as 'classification', 'radial basis' and '5', respectively. In this situation, an RBF-kernel and five-fold cross-validation are implemented to avoid overfitting (168). The 'cost' and 'gamma' parameters are optimized using tune in e1071 R package (169). The output of ROC curve is a graph where x-axis and y-axis represent the 'specificity' and '1-sensitivity', respectively. The higher these two values, the larger the value of AUC. The value of AUC closer to 1 denotes higher classification performance of the classifier, and the AUC within the range of >0.9, ≤0.9 & >0.7 and ≤0.7 refers to *Good, Fair and Poor* performances, respectively (170,171).

Criterion C*e*: level of correspondence between processed and reference data.

The metric calculated under this criterion is the log value of fold changes (logFC) between the concentrations of any two groups in the analyzed dataset. The level of correspondence between the processed and reference data is calculated. Specifically, the level of correspondence can be evaluated by calculating the logFC between processed data and references (relative intensities of various spike-in metabolites) for performance assessment (78,167). The closer the logFC of the means of normalized data corresponds to that of reference data, the better the performance is. Moreover, Criterion C*e* utilizes boxplots for demonstrating the variations between any two groups, and it is desirable that the medians in boxplot would equal to zero with the narrowed variations (131). The logFC of the standard deviations can be calculated as a supplement.

Method S4. The user manual of NOREVA protocol.

Introduction

The NOREVA package not only enables the pre-processing and assessment of multi-class/time-series metabolomic data but also realize a high-throughput discovery of the well-performing pre-processing workflows. Particularly, five well-established criteria, each with a distinct underlying theory, are integrated to ensure a much more comprehensive evaluation than any single criterion. This study provides guidelines for researchers who will engage in biomarker discovery or other differential profiling “omics” studies with respect selecting the most appropriate preprocessing method for a given dataset. For function descriptions and analysis of sample datasets you can also use “??NOREVA” command in *R*.

Installation

```
# download the source package of NOREVA_0.1.0.tar.gz and install it.

install.packages ("NOREVA_0.1.0.tar.gz", repos = NULL, type = "source",
INSTALL_opts = "--no-multiarch")

# Or the development version from GitHub:

install.packages("devtools")

devtools::install_github("idrblab/NOREVA")

# NOREVA package depends on several packages, which can be installed using the
below commands:

if (!requireNamespace("BiocManager", quietly = TRUE))

install.packages("BiocManager")
BiocManager::install("Biobase")
BiocManager::install("pcaMethods")
BiocManager::install("multtest")
BiocManager::install("limma")
BiocManager::install("impute")
BiocManager::install("statTarget")
BiocManager::install("ProteoMM")
BiocManager::install("timecourse")
BiocManager::install("ropls")
BiocManager::install("vsn")
BiocManager::install("affy")
devtools::install_github("metabolomicstats/NormalizeMets")
devtools::install_github("fawda123/ggord")
install.packages(c('rJava', 'DiffCorr', 'MetNorm', 'ggsci', 'multiROC', 'dummies',
'ggfortify', 'ggpubr', 'sampling', 'VennDiagram', 'RcmdrMisc', 'reshape2',
'futile.logger', 'foreach', 'data.table', 'parallel', 'doSNOW', 'tidyverse',
'iterators'))
```

Usage

```
library(NOREVA)
```

1. This function enables the preparation and input of peak table which facilitate the subsequent application of other NOREVA functions.

```
PrepareInuputFiles(dataformat, rawdata, label)
```

`dataformat` This variable allows the user to specify the FORMAT of their input peak table.

“1” denotes the standardized format of peak table accepted by NOREVA;

“2” denotes the customized format of peak table generated by 12 available software tools.

`rawdata` This variable allows the user to indicate the NAME of their input peak table file.

`label` This variable allows the user to indicate the NAME of their input label file for time-course/multi-class.

2. This function enables the performance assessment of time-course metabolomic study with dataset with QCSs based on 4 distinct criteria. It could automatically assess the performance of all processing workflows from different criteria.

```
nortimecourseqcall(fileName, SAalpha="Y", SAbeta="Y", SAgamma ="Y")
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`SAalpha` This variable allows the user to specify the study assumption of their input peak table.

“Y” denotes the peak table satisfies the study assumption.

“N” denotes the peak table satisfies the study assumption.

Study assumption alpha represents that all metabolites are assumed to be equally important.

`SAbeta` This variable allows the user to specify the study assumption of their input peak table.

“Y” denotes the peak table satisfies the study assumption.

“N” denotes the peak table satisfies the study assumption.

Study assumption beta represents that the level of metabolite abundance is constant among all samples.

`SAgamma` This variable allows the user to specify the study assumption of their input peak table.

“Y” denotes the peak table satisfies the study assumption.

“N” denotes the peak table satisfies the study assumption.

Study assumption gamma represents that the intensities of the majority of the metabolites are not changed under the studied conditions.

3. This function enables the performance assessment of time-course metabolomic study with dataset with ISs based on 4 distinct criteria. It could automatically assess the performance of all processing workflows from different criteria.

```
nortimecourseisall(fileName, IS)
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`IS` This variable allows the user to indicate the column of IS.

If there is only one IS, the column number of this IS should be listed.

If there are multiple ISs, the column number of all ISs should be listed and separated by comma (,).

For example, the replacement of IS to 2,6,9,n indicates that the metabolites in the 2st, 6th, 9th, and nth columns of in your peak table should be considered as the ISs metabolites.

4. This function enables the performance assessment of time-course metabolomic study with dataset without QCSs and ISs based on 4 distinct criteria. It could automatically assess the performance of all processing workflows from different criteria.

```
nortimecoursenoall(fileName, SAalpha="Y", SAbeta ="Y", SAgamma ="Y")
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function. Sample data of this data type can be downloaded as the following section “Welcome to Download the Sample Data for Testing and for File Format Correcting”.

`SAalpha` This variable allows the user to specify the study assumption of their input peak table.

“Y” denotes the peak table satisfies the study assumption.

“N” denotes the peak table satisfies the study assumption.

Study assumption alpha represents that all metabolites are assumed to be equally important.

`SAbeta` This variable allows the user to specify the study assumption of their input peak table.

“Y” denotes the peak table satisfies the study assumption.

“N” denotes the peak table satisfies the study assumption.

Study assumption beta represents that the level of metabolite abundance is constant among all samples.

`SAgamma` This variable allows the user to specify the study assumption of their input peak table.

“Y” denotes the peak table satisfies the study assumption.

“N” denotes the peak table satisfies the study assumption.

Study assumption gamma represents that the intensities of the majority of the metabolites are not changed under the studied conditions.

5. This function enables the performance assessment of multi-class (N>1) metabolomic study with dataset with QCSs based on 4 distinct criteria. It could automatically assess the performance of all processing workflows from different criteria.

```
normulticlassqcall(fileName, SAalpha="Y", SAbeta ="Y", SAgamma ="Y")
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`SAalpha` This variable allows the user to specify the study assumption of their input peak table.

“Y” denotes the peak table satisfies the study assumption.

“N” denotes the peak table satisfies the study assumption.

Study assumption alpha represents that all metabolites are assumed to be equally important.

`SAbeta` This variable allows the user to specify the study assumption of their input peak table.

“Y” denotes the peak table satisfies the study assumption.

“N” denotes the peak table satisfies the study assumption.

Study assumption beta represents that the level of metabolite abundance is constant among all samples.

`SAgamma` This variable allows the user to specify the study assumption of their input peak table.

“Y” denotes the peak table satisfies the study assumption.

“N” denotes the peak table satisfies the study assumption.

Study assumption gamma represents that the intensities of the majority of the metabolites are not changed under the studied conditions.

6. This function enables the performance assessment of multi-class (N>1) metabolomic study with dataset with ISs based on 4 distinct criteria. It could automatically assess the performance of all processing workflows from different criteria.

```
normulticlassisall(fileName, IS)
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`IS` This variable allows the user to indicate the column of IS.

If there is only one IS, the column number of this IS should be listed.

If there are multiple ISs, the column number of all ISs should be listed and separated by comma (,).

For example, the replacement of IS to 2,6,9,n indicates that the metabolites in the 2st, 6th, 9th, and nth columns of in your peak table should be considered as the ISs metabolites.

7. This function enables the performance assessment of multi-class (N>1) metabolomic study with dataset without QCSs and ISs based on 4 distinct criteria. It could automatically assess the performance of all processing workflows from different criteria.

```
normulticlassnoall(fileName, SAalpha="Y", SAbeta="Y", SAgamma ="Y")
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`SAalpha` This variable allows the user to specify the study assumption of their peak table.

“Y” denotes the peak table satisfies the study assumption.

“N” denotes the peak table satisfies the study assumption.

Study assumption alpha represents that all metabolites are assumed to be equally important.

`SAbeta` This variable allows the user to specify the study assumption of their input peak table.

“Y” denotes the peak table satisfies the study assumption.

“N” denotes the peak table satisfies the study assumption.

Study assumption beta represents that the level of metabolite abundance is constant among all samples.

`SAgamma` This variable allows the user to specify the study assumption of their input peak table.

“Y” denotes the peak table satisfies the study assumption.

“N” denotes the peak table satisfies the study assumption.

Study assumption gamma represents that the intensities of the majority of the metabolites are not changed under the studied conditions.

8. This function enables the performance assessment of time-course metabolomic study with dataset with QCSs based on 5 distinct criteria including the performance of processing workflows could be assessed using criteria (e).

```
nortimecourseqcallgs(fileName, GS, SAalpha="Y", SAbeta="Y", SAgamma="Y")
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`GS` The corresponding data of golden standards for performance evaluation using Criterion e. For the detailed information of the correct file format, please use “??NOREVA” and download sample data in the corresponding section “Welcome to Download the Sample Data for Testing and for File Format Correcting”.

`SAalpha` This variable allows the user to specify the study assumption of their input peak table.

“Y” denotes the peak table satisfies the study assumption.

“N” denotes the peak table satisfies the study assumption.

Study assumption alpha represents that all metabolites are assumed to be equally important.

`SAbeta` This variable allows the user to specify the study assumption of their input peak table.

“Y” denotes the peak table satisfies the study assumption.

“N” denotes the peak table satisfies the study assumption.

Study assumption beta represents that the level of metabolite abundance is constant among all samples.

`SAgamma` This variable allows the user to specify the study assumption of their input peak table.

“Y” denotes the peak table satisfies the study assumption.

“N” denotes the peak table satisfies the study assumption.

Study assumption gamma represents that the intensities of the majority of the metabolites are not changed under the studied conditions.

9. This function enables the performance assessment of time-course metabolomic study with dataset with ISs based on 5 distinct criteria including the performance of processing workflows could be assessed using criteria (e).

```
nortimecourseisallgs(fileName, IS, GS)
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`IS` This variable allows the user to indicate the column of IS.

If there is only one IS, the column number of this IS should be listed.

If there are multiple ISs, the column number of all ISs should be listed and separated by comma (,).

For example, the replacement of IS to 2,6,9,n indicates that the metabolites in the 2st, 6th, 9th, and nth columns of in your peak table should be considered as the ISs metabolites.

`GS` The corresponding data of golden standards for performance evaluation using Criterion e. For the detailed information of the correct file format, please use “??NOREVA” and download sample data in the corresponding section “Welcome to Download the Sample Data for Testing and for File Format Correcting”

10. This function enables the performance assessment of time-course metabolomic study with dataset without QCSs and ISs based on 5 distinct criteria including the performance of processing workflows could be assessed using criteria (e).

```
nortimecoursenoallgs(fileName, GS, SAalpha="Y", SAbeta ="Y", SAgamma ="Y")
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`GS` The corresponding data of golden standards for performance evaluation using Criterion e. For the detailed information of the correct file format, please use “??NOREVA” and download sample data in the corresponding section “Welcome to Download the Sample Data for Testing and for File Format Correcting”.

`SAalpha` This variable allows the user to specify the study assumption of their input peak table.

“Y” denotes the peak table satisfies the study assumption.

“N” denotes the peak table satisfies the study assumption.

Study assumption alpha represents that all metabolites are assumed to be equally important.

`SAbeta` This variable allows the user to specify the study assumption of their input peak table.

“Y” denotes the peak table satisfies the study assumption.

“N” denotes the peak table satisfies the study assumption.

Study assumption beta represents that the level of metabolite abundance is constant among all samples.

`SAgamma` This variable allows the user to specify the study assumption of their input peak table.

"Y" denotes the peak table satisfies the study assumption.

"N" denotes the peak table satisfies the study assumption.

Study assumption gamma represents that the intensities of the majority of the metabolites are not changed under the studied conditions.

11. This function enables the performance assessment of multi-class (N>1) metabolomic study with dataset with QCSs based on 5 distinct criteria including the performance of processing workflows could be assessed using criteria (e).

```
normulticlassqcallgs(fileName, GS, SAalpha="Y", SAbeta ="Y", SAgamma ="Y")
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`GS` The corresponding data of golden standards for performance evaluation using Criterion e. For the detailed information of the correct file format, please use "??NOREVA" and download sample data in the corresponding section "Welcome to Download the Sample Data for Testing and for File Format Correcting".

`SAalpha` This variable allows the user to specify the study assumption of their input peak table.

"Y" denotes the peak table satisfies the study assumption.

"N" denotes the peak table satisfies the study assumption.

Study assumption alpha represents that all metabolites are assumed to be equally important.

`SAbeta` This variable allows the user to specify the study assumption of their input peak table.

"Y" denotes the peak table satisfies the study assumption.

"N" denotes the peak table satisfies the study assumption.

Study assumption beta represents that the level of metabolite abundance is constant among all samples.

`SAgamma` This variable allows the user to specify the study assumption of their input peak table.

"Y" denotes the peak table satisfies the study assumption.

"N" denotes the peak table satisfies the study assumption.

Study assumption gamma represents that the intensities of the majority of the metabolites are not changed under the studied conditions.

12. This function enables the performance assessment of multi-class (N>1) metabolomic study with dataset with ISs based on 5 distinct criteria including the performance of processing workflows could be assessed using criteria (e).

```
normulticlassisallgs(fileName, IS, GS)
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`GS` The corresponding data of golden standards for performance evaluation using Criterion e. For the correct file format, please use "??NOREVA" and download sample data from the section "Welcome to Download the Sample Data for Testing and for File Format Correcting".

`IS` This variable allows the user to indicate the column of IS.

If there is only one IS, the column number of this IS should be listed.

If there are multiple ISs, the column number of all ISs should be listed and separated by comma (,).

For example, the replacement of IS to 2,6,9,n indicates that the metabolites in the 2st, 6th, 9th, and nth columns of in your peak table should be considered as the ISs metabolites.

13. This function enables the performance assessment of multi-class (N>1) metabolomic study with dataset without QCSs and ISs based on 5 distinct criteria including the performance of processing workflows could be assessed using criteria (e).

```
normulticlassnoallgs(fileName, GS, SAalpha="Y", SAbeta="Y", SAgamma ="Y")
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`GS` The corresponding data of golden standards for performance evaluation using Criterion e. For the detailed information of the correct file format, please use "??NOREVA" and download sample data in the corresponding section "Welcome to Download the Sample Data for Testing and for File Format Correcting".

`SAalpha` This variable allows the user to specify the study assumption of their peak table.

"Y" denotes the peak table satisfies the study assumption.

"N" denotes the peak table satisfies the study assumption.

Study assumption alpha represents that all metabolites are assumed to be equally important.

`SAbeta` This variable allows the user to specify the study assumption of their input peak table.

"Y" denotes the peak table satisfies the study assumption.

"N" denotes the peak table satisfies the study assumption.

Study assumption beta represents that the level of metabolite abundance is constant among all samples.

`SAgamma` This variable allows the user to specify the study assumption of their input peak table.

"Y" denotes the peak table satisfies the study assumption.

"N" denotes the peak table satisfies the study assumption.

Study assumption gamma represents that the intensities of the majority of the metabolites are not changed under the studied conditions.

14. This function enables the performance assessment of the processing workflows defined by the preference of NOREVA users based on time-course metabolomic study with dataset with QCSs.

```
nortimecourseqcpart(fileName, selectedMethods)
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`selectedMethods` This variable allows the user to indicate the NAME of the file containing processing workflows defined by the preference of NOREVA users. The file should be in the format of Comma-Separated Values (CSV). Exemplar files are provided by NOREVA and available for download at https://idrblab.org/noreva/NOREVA_exampledata.zip).

15. This function enables the performance assessment of processing workflows defined by the preference of NOREVA users based on time-course metabolomics with ISs.

```
nortimecourseispart(fileName, IS, selectedMethods)
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`IS` This variable allows the user to indicate the column of IS.

If there is only one IS, the column number of this IS should be listed.

If there are multiple ISs, the column number of all ISs are listed and separated by comma (,).

For example, the replacement of IS to 2,6,9,n indicates that the metabolites in the 2st, 6th, 9th, and nth columns of in your peak table should be considered as the ISs metabolites.

`selectedMethods` This variable allows the user to indicate the NAME of the file containing processing workflows defined by the preference of NOREVA users. The file should be in the format of Comma-Separated Values (CSV). Exemplar files are provided by NOREVA and available for download at https://idrblab.org/noreva/NOREVA_exampledata.zip).

16. This function enables the performance assessment of processing workflows defined by users' preference based on time-course metabolomics without QCSs and ISs.

```
nortimecoursenopart(fileName, selectedMethods)
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`selectedMethods` This variable allows the user to indicate the NAME of the file containing processing workflows defined by the preference of NOREVA users. The file should be in the format of Comma-Separated Values (CSV). Exemplar files are provided by NOREVA and available for download at https://idrblab.org/noreva/NOREVA_exampledata.zip).

17. This function enables the performance assessment of processing workflows defined by users' preference based on multi-class (N>1) metabolomic study with dataset with QCSs.

```
normulticlassqcpart(fileName, selectedMethods)
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`selectedMethods` This variable allows the user to indicate the NAME of the file containing processing workflows defined by the preference of NOREVA users. The file should be in the format of Comma-Separated Values (CSV). Exemplar files are provided by NOREVA and available for download at https://idrblab.org/noreva/NOREVA_exampledata.zip).

18. This function enables the performance assessment of processing workflows defined by users' preference based on multi-class (N>1) metabolomic study with dataset with ISs.

```
normulticlassispart(fileName, IS, selectedMethods)
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`IS` This variable allows the user to indicate the column of IS.

If there is only one IS, the column number of this IS should be listed.

If there are multiple ISs, the column number of all ISs should be listed and separated by comma (,).

For example, the replacement of IS to 2,6,9,n indicates that the metabolites in the 2st, 6th, 9th, and nth columns of in your peak table should be considered as the ISs metabolites.

`selectedMethods` This variable allows the user to indicate the NAME of the file containing processing workflows defined by the preference of NOREVA users. The file should be in the format of Comma-Separated Values (CSV). Exemplar files are provided by NOREVA and available for download at https://idrblab.org/noreva/NOREVA_exampledata.zip).

19. This function enables the performance assessment of the processing workflows defined by the preference of NOREVA users based on multi-class (N>1) metabolomic study with dataset without QCSs and ISs.

```
seleranks_non <- normulticlassnopart(fileName, selectedMethods)
```

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`selectedMethods` This variable allows the user to indicate the NAME of the file containing processing workflows defined by the preference of NOREVA users. The file should be in the format of Comma-Separated Values (CSV). Exemplar files are provided by NOREVA and available for download at https://idrblab.org/noreva/NOREVA_exampledata.zip).

20. This function will output a processed peak table of time-course metabolomic study according to the choice of users must provide the processing workflow.

```
nortimecoursematrix(datatype, fileName, IS, impt=NULL, trsf=NULL, nmal=NULL,
nmal2=NULL, nmals=NULL)
```

`datatype` This variable allows the user to specify the Type of their input peak table.

“1” denotes time-course metabolomic study without QCSs and ISs.

“2” denotes time-course metabolomic study with QCSs.

“3” denotes time-course metabolomic study with ISs.

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`IS` This variable allows the user to indicate the column of IS.

If there is only one IS, the column number of this IS should be listed.

If there are multiple ISs, the column number of all ISs are listed and separated by comma (,).

For example, the replacement of IS to 2,6,9,n indicates that the metabolites in the 2st, 6th, 9th, and nth columns of in your peak table should be considered as the ISs metabolites.

`impt` This variable allows the user to specify the Type of imputation method.

“1” denotes method of column mean imputation.

“2” denotes method of column median imputation.

“3” denotes method of half of the minimum positive value.

“4” denote method of K-nearest neighbor imputation.

`trsf` This variable allows the user to specify the Type of transformation method.

“1” denotes method of cube root transformation.

“2” denotes method of log transformation.

“3” denotes none transformation method.

`nmal` This variable allows the user to specify the Type of normalization method.

“1” denotes none normalization method.

Metabolite-based Normalization:

"2" denotes method of probabilistic quotient normalization.

"3" denotes method of cyclic loess.

"4" denotes method of contrast.

"5" denotes method of quantile.

"6" denotes method of linear baseline.

"7" denotes method of Li-Wong.

"8" denotes method of cubic splines.

"16" denotes method of MS total useful signal.

"17" denotes method of total sum normalization.

"18" denotes method of median normalization.

"19" denotes method of mean normalization.

"20" denotes method of EigenMS.

Sample-based Normalization:

"9" denotes method of auto scaling.

"10" denotes method of range scaling.

"11" denotes method of pareto scaling

"12" denotes method of vast scaling

"13" denotes method of level scaling

"15" denotes method of power scaling

Sample & Metabolite-based Normalization:

"14" denotes method of variance stabilization normalization.

`nmal2` This variable allows the user to specify the Type of normalization method. According to the normalization methods of combination strategy, if you choose sample-based normalization methods for argument “nmal”, “nmal2” should select metabolite-based normalization methods. Similarly, if you choose metabolite-based normalization methods for argument “nmal”, “nmal2” should select sample-based normalization methods. The VSN method you selected is a sample & metabolite-based normalization, which should be applied alone to remove the unwanted signal variations.

`nmals` This variable allows the user to specify the Type of IS-based normalization method.

“1” denotes method of Single Internal Standard.

“2” denotes method of Normalization using Optimal Selection of Multiple ISs

“3” denotes method of Cross-contribution Compensating Multi-ISs Normalization

“4” denotes method of Remove Unwanted Variation-Random.

21. This function will output a processed peak table of multi-class metabolomic study according to the choice of users must provide the processing workflow.

```
normulticlassmatrix(datatype, fileName, IS, impt=NULL, trsf=NULL, nmal=NULL,
nmal2=NULL, nmals=NULL)
```

`datatype` Input the number of data type.

If set 1, the dataset of multi-class (N>1) metabolomic study without QCSs and ISs.

If set 2, the dataset of multi-class (N>1) metabolomic study with QC samples (QCSs).

If set 3, the dataset of multi-class (N>1) metabolomic study with dataset with (ISs).

`fileName` This variable allows the user to indicate the NAME of result obtained from PrepareInuputFiles function.

`IS` This variable allows the user to indicate the column of IS.

If there is only one IS, the column number of this IS should be listed.

If there are multiple ISs, the column number of all ISs are listed and separated by comma (,).

For example, the replacement of IS to 2,6,9,n indicates that the metabolites in the 2st, 6th, 9th, and nth columns of in your peak table should be considered as the ISs metabolites.

`impt` This variable allows the user to specify the Type of imputation method.

“1” denotes method of column mean imputation.

“2” denotes method of column median imputation.

“3” denotes method of half of the minimum positive value.

“4” denote method of K-nearest neighbor imputation.

`trsf` This variable allows the user to specify the Type of transformation method.

“1” denotes method of cube root transformation.

“2” denotes method of log transformation.

“3” denotes none transformation method.

`nmal` This variable allows the user to specify the Type of normalization method.

“1” denotes none normalization method.

Metabolite-based Normalization:

"2" denotes method of probabilistic quotient normalization.

"3" denotes method of cyclic loess.

"4" denotes method of contrast.

"5" denotes method of quantile.

"6" denotes method of linear baseline.

"7" denotes method of Li-Wong.

"8" denotes method of cubic splines.

"16" denotes method of MS total useful signal.

"17" denotes method of total sum normalization.

"18" denotes method of median normalization.

"19" denotes method of mean normalization.

"20" denotes method of EigenMS.

Sample-based Normalization:

"9" denotes method of auto scaling.

"10" denotes method of range scaling.

"11" denotes method of pareto scaling

"12" denotes method of vast scaling

"13" denotes method of level scaling

"15" denotes method of power scaling

Sample & Metabolite-based Normalization:

"14" denotes method of variance stabilization normalization.

`nmal2` This variable allows the user to specify the Type of normalization method. According to the normalization methods of combination strategy, if you choose sample-based normalization methods for argument “nmal”, “nmal2” should select metabolite-based normalization methods. Similarly, if you choose metabolite-based normalization methods for argument “nmal”, “nmal2” should select sample-based normalization methods. The VSN method you selected is a sample & metabolite-based normalization, which should be applied alone to remove the unwanted signal variations.

`nmals` This variable allows the user to specify the Type of IS-based normalization method.

“1” denotes method of Single Internal Standard.

“2” denotes method of Normalization using Optimal Selection of Multiple ISs

“3” denotes method of Cross-contribution Compensating Multi-ISs Normalization

“4” denotes method of Remove Unwanted Variation-Random.

22. Plot circular barplot of overall ranking results. A circular barplot illustrating the performance level and the overall ranking of all calculatable processing workflows based on the multiple criteria or a single criterion that are selected by user..

```
norvisualization(data, outputfile="NOREVA-Ranking-
Top.%d.workflows.%s",cutoff="100", outputtype="pdf", maxValue="40", colorSet =
c("#EA4335", "#4285F4", "#FBBC05", "#800080"), totalAngle = "340", bgColor =
"#FFFFFF", fontColor="#000000")
```

`data` This variable allows the user to specify the NAME of the file (.csv) containing the names of processing workflows, their ranking value and representative measurement values under differential criteria, which is obtained from the functions such as the “normulticlassnoall”, “normulticlassqcall”, “nortimecourseqcall”, or “nortimecoursenoall” et al.

`outputfile` This variable allows the user to specify the NAME of the output file. A format string containing the cutoff value and data type to generate formatted file name.

`cutoff` This variable allows the user to specify the cutoff value. Integer for the number of strategies ranking at the top of the list, which is used to filter the results. Integer, which means to filter the results, the default is 100.

`outputtype` String, indicating the output type, support pdf, eps, default is pdf.

`MaxValue` Double-precision floating-point number, representing the characteristic value represented by the maximum length of the rectangle, the default is 40.

`colorSet` Hexadecimal color string group, representing the four-layer color setting of the graphics from the inside to the outside, the default is "#FFE699", "#D3E8C7", "#B2B2FF", "#FFCACA".

`totalAngle` Double-precision floating-point number, representing the total angle of rotation of the drawing, in degrees, the default value is 340.

`bgColor` Hexadecimal color string, representing the background color of the graphic drawing, the default is white (#FFFFFF).

`FontColor` Hexadecimal color string, representing the font color, the default is black (#000000).

Welcome to Download the Sample Data for Testing and for File Format Correcting

```
# Time-course Metabolomic Study
```

Dataset with quality control samples (QCSs) could be downloaded and the corresponding data of golden standards for performance evaluation using Criterion e could be downloaded.

Dataset with internal standards (ISs) could be downloaded and the corresponding data of golden standards for performance evaluation using Criterion e could be downloaded.

Dataset without QCSs and ISs could be downloaded and the corresponding data of golden standards for performance evaluation using Criterion e could be downloaded.

```
# Multi-class Metabolomic Study
```

Dataset with quality control samples (QCSs) could be downloaded and the corresponding data of golden standards for performance evaluation using Criterion e could be downloaded.

Dataset with internal standards (ISs) could be downloaded and the corresponding data of golden standards for performance evaluation using Criterion e could be downloaded.

Dataset without QCSs and ISs could be downloaded and the corresponding data of golden standards for performance evaluation using Criterion e could be downloaded.

Sequential step for the performance assessment of time-course metabolomic study with dataset with QCSs. For other types of study, replace the function related to the types.

```
Step 1: time_qcs_data <- PrepareInuputFiles(dataformat = 1, rawdata =
"Timecourse-QC-XXX.csv")

Step 2: normulticlassqcall(fileName = time_qcs_data, SAalpha = "Y", SAbeta = "Y",
SAgamma = "Y")

Step 3: norvisualization(data = "OUTPUT-NOREVA-Overall.Ranking.Data.csv", cutoff
= "100")
```

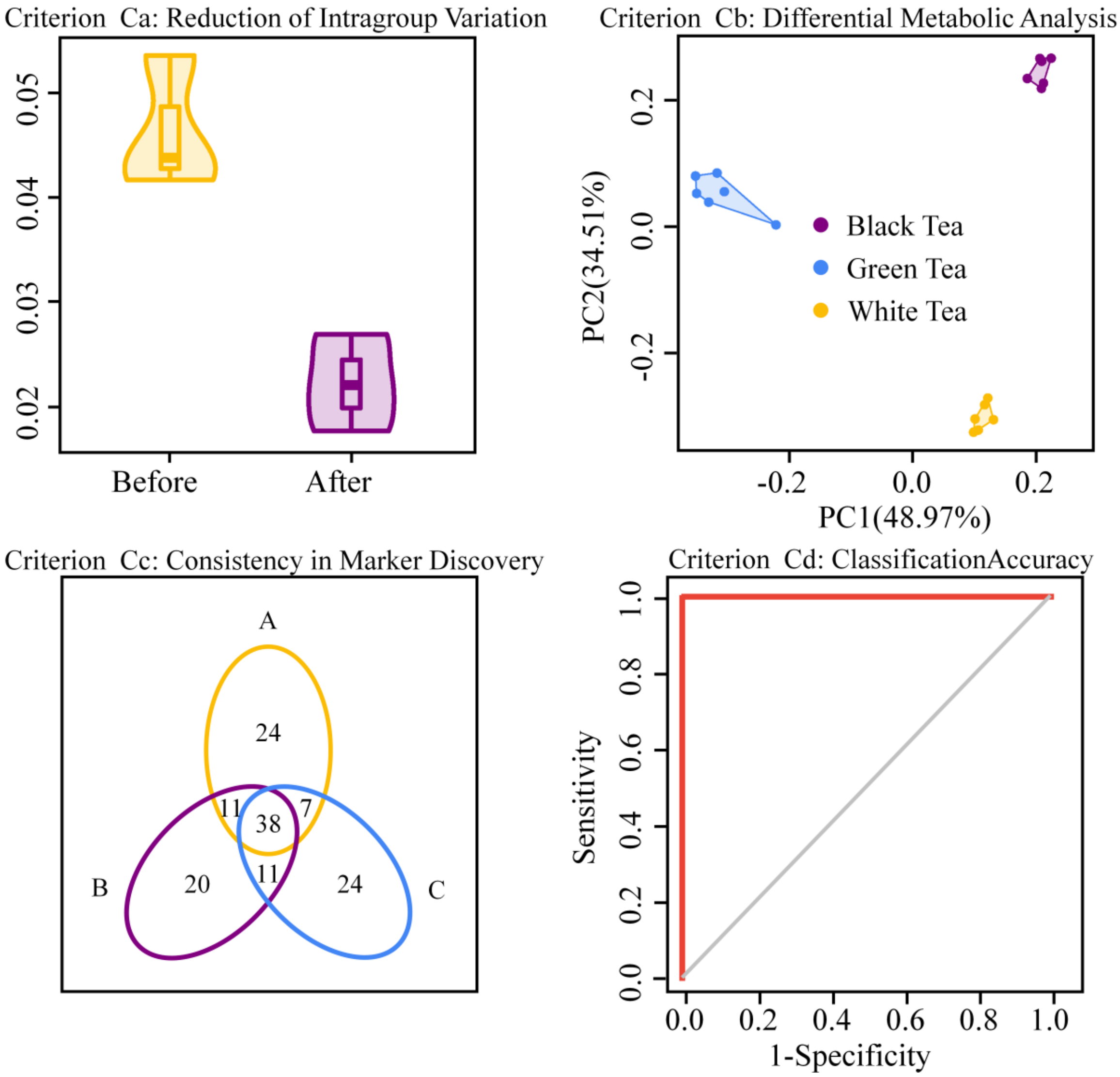
Sequential step for the performance assessment of multi-class metabolomic study with dataset with QCSs. For other types of study, replace the function related to the types.

```
Step 1: multi_qcs_data <- PrepareInuputFiles(dataformat = 1, rawdata =
"Multiclass-QC-XXX.csv")

Step 2: normulticlassqcall(fileName = multi_qcs_data, SAalpha = "Y", SAbeta =
"Y", SAgamma = "Y")

Step 3: norvisualization(data = "OUTPUT-NOREVA-Overall.Ranking.Data.csv", cutoff
= "100")
```

Examples

Step 1: Prepare input of peak table for assessing normalization for metabolomic data

```
multi_qcs_data <- PrepareInuputFiles(dataformat = 1, rawdata = "Multiclass-QC-
MTSL403.csv")
```



NOREVA-All-Criteria-Output-Figures

Note: the file should be in the format of Comma-Separated Values (CSV), which provides the intensity data of metablites. Different functions require different data types. Please refer to the section “Welcome to Download the Sample Data for Testing and for File Format Correcting”.

Sample data MTBLS403: untargeted metabolomic dataset of 3 classes with quality control samples (QCSs), which contained 3805 metabolites from 3 types of tea samples (green tea, black tea and white tea), could be downloaded. *Food Res Int*. 96:40-45,2017.

Step 2: Assessing all processing workflows for multi-class metabolomic data

Multi-class (N>1) Metabolomic Study with dataset with Quality Control Samples (QCSs)

```
normulticlassqcall(fileName=multi_qcs_data,SAalpha="Y",SAbeta="Y",SAgamma="Y")
```

```
allrankings <- read.csv(file = "./sampledata/OUTPUT-NOREVA-
verall.Ranking.Data.csv",header = T)
head(allrankings)
```

```
##                X Overall.Rank Criteria.Ca.Rank Criteria.Cb.Rank
## 1 HAM+CUT+SUM+LEV            1              164                1
## 2 HAM+LOG+MST+NON            2               78                1
## 3 HAM+LOG+MST+PAR            3              166                1
## 4 HAM+LOG+SUM+RAN            4              298                1
## 5 KNN+LOG+MST+NON            5               85                1
## 6 HAM+CUT+SUM+RAN            6              305                1
##   Criteria.Cc.Rank Criteria.Cd.Rank Criteria.Ca.Value Criteria.Cb.Value
## 1               50                1            0.0076                 1
## 2              140                1            0.0003                 1
## 3              139                1            0.0079                 1
## 4               13                1            0.0344                 1
## 5              232                1            0.0012                 1
## 6               51                1            0.0365                 1
##   Criteria.Cc.Value Criteria.Cd.Value
## 1            0.7208                 1
## 2            0.6167                 1
## 3            0.6167                 1
## 4            0.7500                 1
## 5            0.5667                 1
## 6            0.7208                 1
```

Step 3: a circular barplot illustrating the performance ranking of all processing workflows.

```
norvisualization(data = "OUTPUT-NOREVA-Overall.Ranking.Data.csv", cutoff = "100")
```

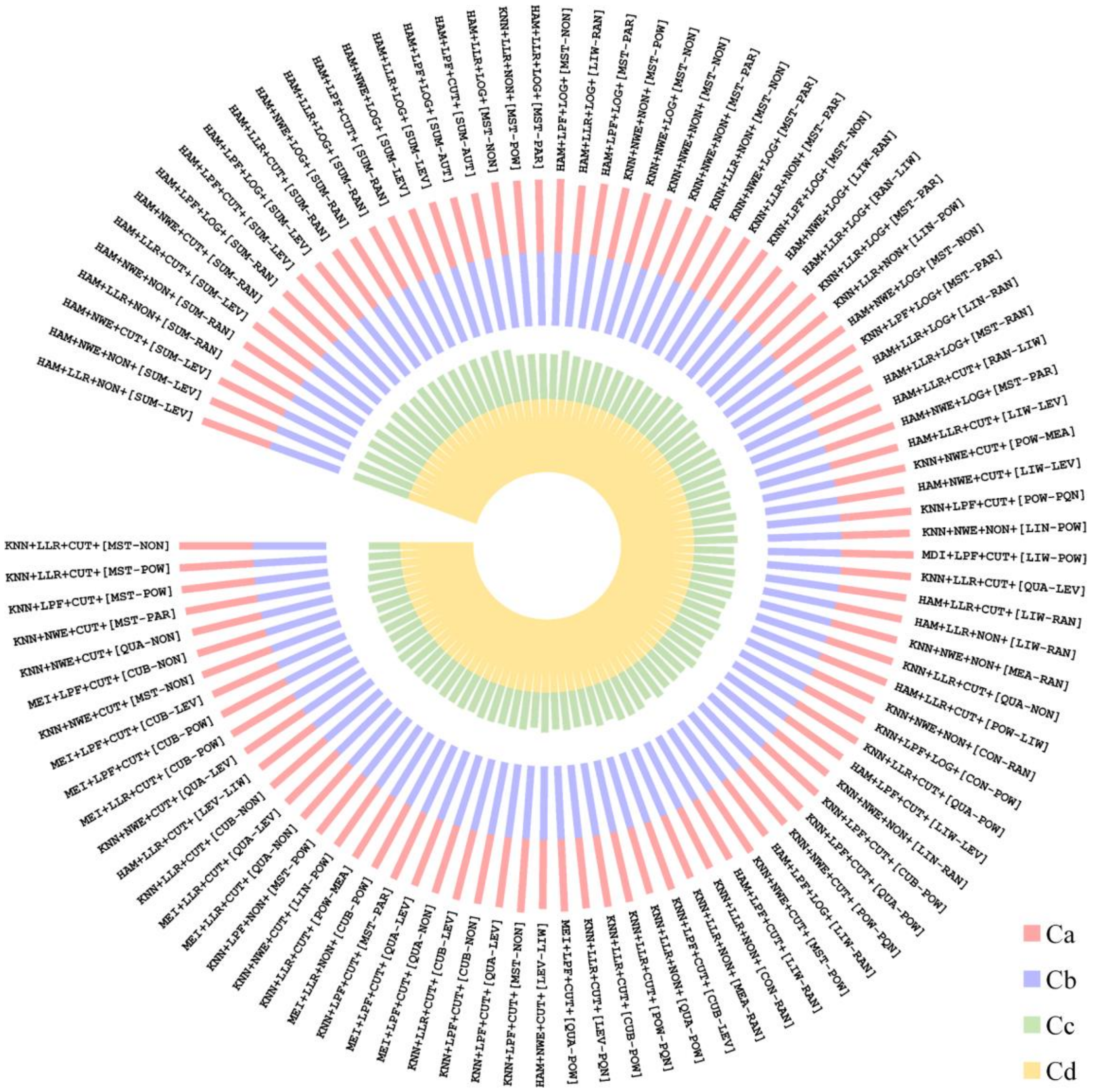
Comprehensive assessment among all processing workflows (the top-100 were shown) based on the collective evaluations using four different criteria.

Step 4: Processing datasets using the processing workflow based on the results of assessment.

Normalization with datasets of multi-class (N>1) metabolomic study

```
nordata <- normulticlassmatrix(datatype = 2, fileName = multi_qcs_data, impt = "3",
trsf = "1", nmal = "17", nmal2 = "13")
```

Note: please select the appropriate number code represents imputation, transformation, normalization methods (See above details).



NOREVA-Ranking-Top.100.workflows

Step 5: Users can also use NOREVA for accessing the part of normalization methods/strategies which you preferred.

Multi-class (N>1) Metabolomic Study with dataset with Quality Control Samples (QCSs)

```
normulticlassqcpart(fileName = multi_qcs_data, selectedMethods =
"selectedMethods.csv")
```

Note: please select the appropriate number code represents imputation, transformation, normalization methods (see above details).

References

1. Skarke, C., Lahens, N.F., Rhoades, S.D., Campbell, A., Bittinger, K., Bailey, A., Hoffmann, C., Olson, R.S., Chen, L., Yang, G. *et al.* (2017) A pilot characterization of the human chronobiome. *Sci Rep*, **7**, 17141.

2. Dos Santos, R.O., Goncalves-Lopes, R.M., Lima, N.F., Scopel, K.K.G., Ferreira, M.U. and Lalwani, P. (2020) Kynurenine elevation correlates with T regulatory cells increase in acute Plasmodium vivax infection: a pilot study. *Parasite Immunol*, **42**, e12689.

3. Hunt, N.H., Too, L.K., Khaw, L.T., Guo, J., Hee, L., Mitchell, A.J., Grau, G.E. and Ball, H.J. (2017) The kynurenine pathway and parasitic infections that affect CNS function. *Neuropharmacology*, **112**, 389-398.

4. Wehrens, R., Franceschi, P., Vrhovsek, U. and Mattivi, F. (2011) Stability-based biomarker selection. *Anal Chim Acta*, **705**, 15-23.

5. Hao, J., Liebeke, M., Astle, W., De Iorio, M., Bundy, J.G. and Ebbels, T.M. (2014) Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nat Protoc*, **9**, 1416-1427.

6. Kasalica, V., Schwammle, V., Palmblad, M., Ison, J. and Lamprecht, A.L. (2021) APE in the wild: automated exploration of proteomics workflows in the bio.tools registry. *J Proteome Res*, **10.102**, 0c00983.

7. Hohrenk, L.L., Itzel, F., Baetz, N., Tuerk, J., Vosough, M. and Schmidt, T.C. (2020) Comparison of software tools for liquid chromatography-high-resolution mass spectrometry data processing in nontarget screening of environmental samples. *Anal Chem*, **92**, 1898-1907.

8. Haimi, P., Uphoff, A., Hermansson, M. and Somerharju, P. (2006) Software tools for analysis of mass spectrometric lipidome data. *Anal Chem*, **78**, 8324-8331.

9. Zhou, W., Su, S.L., Duan, J.A., Guo, J.M., Qian, D.W., Shang, E.X. and Zhang, J. (2010) Characterization of the active constituents in shixiao san using bioactivity evaluation followed by UPLC-QTOF and Markerlynx analysis. *Molecules*, **15**, 6217-6230.

10. Fernandez-Ochoa, A., Quirantes-Pine, R., Borras-Linares, I., Cadiz-Gurrea, M.L., Alarcon Riquelme, M.E., Brunius, C. and Segura-Carretero, A. (2020) A case report of switching from specific vendor-based to R-based pipelines for untargeted LC-MS metabolomics. *Metabolites*, **10**, 28.

11. Wang, X., Zhang, A., Han, Y., Wang, P., Sun, H., Song, G., Dong, T., Yuan, Y., Yuan, X., Zhang, M. *et al.* (2012) Urine metabolomics analysis for biomarker discovery and

detection of jaundice syndrome in patients with liver disease. *Mol Cell Proteomics*, **11**, 370-380.

12. Clasquin, M.F., Melamud, E. and Rabinowitz, J.D. (2012) LC-MS data processing with MAVEN: a metabolomic analysis and visualization engine. *Curr Protoc Bioinformatics*, **14**, Unit14.11.

13. Claridge, T. (2009) MNova: NMR data processing, analysis, and prediction software. *J Chem Inf Model*, **49**, 1136-1137.

14. Pluskal, T., Castillo, S., Villar-Briones, A. and Oresic, M. (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, **11**, 395.

15. Schober, D., Jacob, D., Wilson, M., Cruz, J.A., Marcu, A., Grant, J.R., Moing, A., Deborde, C., de Figueiredo, L.F., Haug, K. *et al.* (2018) nmrML: a community supported open data standard for the description, storage, and exchange of NMR data. *Anal Chem*, **90**, 649-656.

16. Jacob, D., Deborde, C., Lefebvre, M., Maucourt, M. and Moing, A. (2017) NMRProcFlow: a graphical and interactive tool dedicated to 1D spectra processing for NMR-based metabolomics. *Metabolomics*, **13**, 36.

17. Rost, H.L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., Andreotti, S., Ehrlich, H.C., Gutenbrunner, P., Kenar, E. *et al.* (2016) OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Methods*, **13**, 741-748.

18. Zhang, J., Yang, W., Li, S., Yao, S., Qi, P., Yang, Z., Feng, Z., Hou, J., Cai, L., Yang, M. *et al.* (2016) An intelligentized strategy for endogenous small molecules characterization and quality evaluation of earthworm from two geographic origins by ultra-high performance HILIC/QTOF MS(E) and Progenesis QI. *Anal Bioanal Chem*, **408**, 3881-3890.

19. Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J. *et al.* (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol*, **30**, 918-920.

20. Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. and Siuzdak, G. (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem*, **78**, 779-787.

21. Verhoeven, A., Giera, M. and Mayboroda, O.A. (2018) KIMBLE: a versatile visual NMR metabolomics workbench in KNIME. *Anal Chim Acta*, **1044**, 66-76.

22. Neuweger, H., Albaum, S.P., Dondrup, M., Persicke, M., Watt, T., Niehaus, K., Stoye, J.

and Goesmann, A. (2008) MeltDB: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics*, **24**, 2726-2732.

23. Xia, J., Mandal, R., Sinelnikov, I.V., Broadhurst, D. and Wishart, D.S. (2012) MetaboAnalyst 2.0: a comprehensive server for metabolomic data analysis. *Nucleic Acids Res*, **40**, W127-W133.

24. Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., Edison, A., Fiehn, O., Higashi, R., Nair, K.S. *et al.* (2016) Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res*, **44**, D463-D470.

25. Kastenmuller, G., Romisch-Margl, W., Wagele, B., Altmaier, E. and Suhre, K. (2011) metaP-server: a web-based metabolomics data analysis tool. *J Biomed Biotechnol*, **2011**, 839862.

26. Wen, B., Mei, Z., Zeng, C. and Liu, S. (2017) metaX: a flexible and comprehensive software for processing metabolomics data. *BMC Bioinformatics*, **18**, 183.

27. Biswas, A., Mynampati, K.C., Umashankar, S., Reuben, S., Parab, G., Rao, R., Kannan, V.S. and Swarup, S. (2010) MetDAT: a modular and workflow-based free online pipeline for mass spectrometry data processing, analysis and interpretation. *Bioinformatics*, **26**, 2639-2640.

28. Shen, X. and Zhu, Z.J. (2019) MetFlow: an interactive and integrated workflow for metabolomics data cleaning and differential metabolite discovery. *Bioinformatics*, **35**, 2870-2872.

29. Yang, Q., Li, B., Chen, S., Tang, J., Li, Y., Li, Y., Zhang, S., Shi, C., Zhang, Y., Mou, M. *et al.* (2021) MMEASE: online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis. *J Proteomics*, **232**, 104023.

30. Seo, J.W., Han, K., Lee, J., Kim, E.K., Moon, H.J., Yoon, J.H., Park, V.Y., Baek, H.M. and Kwak, J.Y. (2018) Application of metabolomics in prediction of lymph node metastasis in papillary thyroid carcinoma. *PLoS One*, **13**, e0193883.

31. Guitton, Y., Tremblay-Franco, M., Le Corguille, G., Martin, J.F., Petera, M., Roger-Mele, P., Delabriere, A., Goulitquer, S., Monsoor, M., Duperier, C. *et al.* (2017) Create, run, share, publish, and reference your LC-MS, FIA-MS, GC-MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 galaxy online infrastructure for metabolomics. *Int J Biochem Cell Biol*, **93**, 89-101.

32. Cardoso, S., Afonso, T., Maraschin, M. and Rocha, M. (2019) WebSpecmine: a website for

metabolomics data analysis and mining. *Metabolites*, **9**, 237.

33. Lee, N.Y., Yoon, S.J., Han, D.H., Gupta, H., Youn, G.S., Shin, M.J., Ham, Y.L., Kwak, M.J., Kim, B.Y., Yu, J.S. *et al.* (2020) Lactobacillus and Pediococcus ameliorate progression of non-alcoholic fatty liver disease through modulation of the gut microbiome. *Gut Microbes*, **11**, 882-899.

34. Ayoola, M.B., Shack, L.A., Nakamya, M.F., Thornton, J.A., Swiatlo, E. and Nanduri, B. (2019) Polyamine synthesis effects capsule expression by reduction of precursors in Streptococcus pneumoniae. *Front Microbiol*, **10**, 1996.

35. Franciosi, E., Narduzzi, L., Paradiso, A., Carlin, S., Tuohy, K., Beretta, A. and Mattivi, F. (2020) Microbial community dynamics in phyto-thermotherapy baths viewed through next generation sequencing and metabolomics approach. *Sci Rep*, **10**, 17931.

36. Taverna, F., Goveia, J., Karakach, T.K., Khan, S., Rohlenova, K., Treps, L., Subramanian, A., Schoonjans, L., Dewerchin, M., Eelen, G. *et al.* (2020) BIOMEX: an interactive workflow for (single cell) omics data interpretation and visualization. *Nucleic Acids Res*, **48**, W385-W394.

37. Liu, R. and Yang, Z. (2021) Single cell metabolomics using mass spectrometry: Techniques and data analysis. *Anal Chim Acta*, **1143**, 124-134.

38. Whitson, J.A., Bitto, A., Zhang, H., Sweetwyne, M.T., Coig, R., Bhayana, S., Shankland, E.G., Wang, L., Bammler, T.K., Mills, K.F. *et al.* (2020) SS-31 and NMN: two paths to improve metabolism and function in aged hearts. *Aging Cell*, **19**, e13213.

39. Cui, X., Yang, Q., Li, B., Tang, J., Zhang, X., Li, S., Li, F., Hu, J., Lou, Y., Qiu, Y. *et al.* (2019) Assessing the effectiveness of direct data merging strategy in long-term and large-scale pharmacometabonomics. *Front Pharmacol*, **10**, 127.

40. Woollam, M., Teli, M., Liu, S., Daneshkhah, A., Siegel, A.P., Yokota, H. and Agarwal, M. (2020) Urinary volatile terpenes analyzed by gas chromatography-mass spectrometry to monitor breast cancer treatment efficacy in mice. *J Proteome Res*, **19**, 1913-1922.

41. Lee, S.M., Kang, Y., Lee, E.M., Jung, Y.M., Hong, S., Park, S.J., Park, C.W., Norwitz, E.R., Lee, D.Y. and Park, J.S. (2020) Metabolomic biomarkers in midtrimester maternal plasma can accurately predict the development of preeclampsia. *Sci Rep*, **10**, 16142.

42. Lee, C.W., Lee, D., Lee, E.M., Park, S.J., Ji, D.Y., Lee, D.Y. and Jung, Y.C. (2019) Lipidomic profiles disturbed by the internet gaming disorder in young Korean males. *J Chromatogr B Analyt Technol Biomed Life Sci*, **1114-1115**, 119-124.

43. Park, S.J., Lee, J., Lee, S., Lim, S., Noh, J., Cho, S.Y., Ha, J., Kim, H., Kim, C., Park, S. *et*

al. (2020) Exposure of ultrafine particulate matter causes glutathione redox imbalance in the hippocampus: a neurometabolic susceptibility to Alzheimer's pathology. *Sci Total Environ*, **718**, 137267.

44. Lee, B.M., Lee, E.M., Kang, D.J., Seo, J., Choi, H., Kim, Y. and Lee, D.Y. (2020) Discovery study of integrative metabolic profiles of sesame seeds cultivated in different countries. *LWT-Food Sci Technol*, **129**, 109454.

45. Gonzalez-Riano, C., Dudzik, D., Garcia, A., Gil-de-la-Fuente, A., Gradillas, A., Godzien, J., Lopez-Gonzalvez, A., Rey-Stolle, F., Rojo, D., Ruperez, F.J. *et al.* (2020) Recent developments along the analytical process for metabolomics workflows. *Anal Chem*, **92**, 203-226.

46. Deng, K., Zhang, F., Tan, Q., Huang, Y., Song, W., Rong, Z., Zhu, Z.J., Li, K. and Li, Z. (2019) WaveICA: a novel algorithm to remove batch effects for large-scale untargeted metabolomics data based on wavelet analysis. *Anal Chim Acta*, **1061**, 60-69.

47. De Livera, A.M., Olshansky, G., Simpson, J.A. and Creek, D.J. (2018) NormalizeMets: assessing, selecting and implementing statistical methods for normalizing metabolomics data. *Metabolomics*, **14**, 54.

48. Drotleff, B. and Lammerhofer, M. (2019) Guidelines for selection of internal standard-based normalization strategies in untargeted lipidomic profiling by LC-HR-MS/MS. *Anal Chem*, **91**, 9836-9843.

49. Noonan, M.J., Tinnesand, H.V. and Buesching, C.D. (2018) Normalizing gas-chromatography-mass spectrometry data: method choice can alter biological inference. *Bioessays*, **40**, e1700210.

50. Han, W. and Li, L. (2020) Evaluating and minimizing batch effects in metabolomics. *Mass Spectrom Rev*, **10.1002**, mas.21672.

51. Zullig, T. and Kofeler, H.C. (2021) High resolution mass spectrometry in lipidomics. *Mass Spectrom Rev*, **40**, 162-176.

52. Narduzzi, L., Royer, A.L., Bichon, E., Guitton, Y., Buisson, C., Le Bizec, B. and Dervilly-Pinel, G. (2019) Ammonium fluoride as suitable additive for HILIC-based LC-HRMS metabolomics. *Metabolites*, **9**, 292.

53. Liggi, S., Hinz, C., Hall, Z., Santoru, M.L., Poddighe, S., Fjeldsted, J., Atzori, L. and Griffin, J.L. (2018) KniMet: a pipeline for the processing of chromatography-mass spectrometry metabolomics data. *Metabolomics*, **14**, 52.

54. Lin, S., Liu, H., Kanawati, B., Liu, L., Dong, J., Li, M., Huang, J., Schmitt-Kopplin, P. and

Cai, Z. (2013) Hippocampal metabolomics using ultrahigh-resolution mass spectrometry reveals neuroinflammation from Alzheimer's disease in CRND8 mice. *Anal Bioanal Chem*, **405**, 5105-5117.

55. Oakes, J.M., Scadeng, M., Breen, E.C., Marsden, A.L. and Darquenne, C. (2012) Rat airway morphometry measured from in situ MRI-based geometric models. *J Appl Physiol*, **112**, 1921-1931.

56. Dai, W., Wei, C., Kong, H., Jia, Z., Han, J., Zhang, F., Wu, Z., Gu, Y., Chen, S., Gu, Q. *et al.* (2011) Effect of the traditional Chinese medicine tongxinluo on endothelial dysfunction rats studied by using urinary metabonomics based on liquid chromatography-mass spectrometry. *J Pharm Biomed Anal*, **56**, 86-92.

57. Xia, J., Psychogios, N., Young, N. and Wishart, D.S. (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res*, **37**, W652-W660.

58. Taylor, S.L., Ruhaak, L.R., Kelly, K., Weiss, R.H. and Kim, K. (2017) Effects of imputation on correlation: implications for analysis of mass spectrometry data from multiple biological matrices. *Brief Bioinform*, **18**, 312-320.

59. Willmann, L., Schlimpert, M., Hirschfeld, M., Erbes, T., Neubauer, H., Stickeler, E. and Kammerer, B. (2016) Alterations of the exo- and endometabolite profiles in breast cancer cell lines: a mass spectrometry-based metabolomics approach. *Anal Chim Acta*, **925**, 34-42.

60. Chai, L.E., Law, C.K., Mohamad, M.S., Chong, C.K., Choon, Y.W., Deris, S. and Illias, R.M. (2014) Investigating the effects of imputation methods for modelling gene networks using a dynamic bayesian network from gene expression data. *Malays J Med Sci*, **21**, 20-27.

61. Verma, P., Devaraj, J., Skiles, J.L., Sajdyk, T., Ho, R.H., Hutchinson, R., Wells, E., Li, L., Renbarger, J., Cooper, B. *et al.* (2020) A metabolomics approach for early prediction of vincristine-induced peripheral neuropathy. *Sci Rep*, **10**, 9659.

62. Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T. and Ni, Y. (2018) Missing value imputation approach for mass spectrometry-based metabolomics data. *Sci Rep*, **8**, 663.

63. Rotroff, D.M., Oki, N.O., Liang, X., Yee, S.W., Stocker, S.L., Corum, D.G., Meisner, M., Fiehn, O., Motsinger-Reif, A.A., Giacomini, K.M. *et al.* (2016) Pharmacometabolomic assessment of metformin in non-diabetic, African Americans. *Front Pharmacol*, **7**, 135.

64. Rotroff, D.M., Shahin, M.H., Gurley, S.B., Zhu, H., Motsinger-Reif, A., Meisner, M., Beitelshees, A.L., Fiehn, O., Johnson, J.A., Elbadawi-Sidhu, M. *et al.* (2015)

Pharmacometabolomic assessments of atenolol and hydrochlorothiazide treatment reveal novel drug response phenotypes. *Pharmacometrics Syst Pharmacol*, **4**, 669-679.

65. Cervellera, C. and Maccio, D. (2014) Local linear regression for function learning: an analysis based on sample discrepancy. *IEEE Trans Neural Netw Learn Syst*, **25**, 2086-2098.

66. Su, B., Luo, P., Yang, Z., Yu, P., Li, Z., Yin, P., Zhou, L., Fan, J., Huang, X., Lin, X. *et al.* (2019) A novel analysis method for biomarker identification based on horizontal relationship: identifying potential biomarkers from large-scale hepatocellular carcinoma metabolomics data. *Anal Bioanal Chem*, **411**, 6377-6386.

67. Gamst, A., Wolfson, T. and Parry, B. (2004) Local polynomial regression modeling of human plasma melatonin levels. *J Biol Rhythms*, **19**, 164-174.

68. Narduzzi, L., Dervilly, G., Marchand, A., Audran, M., Le Bizec, B. and Buisson, C. (2020) Applying metabolomics to detect growth hormone administration in athletes: Proof of concept. *Drug Test Anal*, **12**, 887-899.

69. Meinicke, P., Klanke, S., Memisevic, R. and Ritter, H. (2005) Principal surfaces from unsupervised kernel regression. *IEEE Trans Pattern Anal Mach Intell*, **27**, 1379-1391.

70. Luan, H., Ji, F., Chen, Y. and Cai, Z. (2018) statTarget: a streamlined tool for signal drift correction and interpretations of quantitative mass spectrometry-based omics data. *Anal Chim Acta*, **1036**, 66-72.

71. Zheng, H., Cai, A., Zhou, Q., Xu, P., Zhao, L., Li, C., Dong, B. and Gao, H. (2017) Optimal preprocessing of serum and urine metabolomic data fusion for staging prostate cancer through design of experiment. *Anal Chim Acta*, **991**, 68-75.

72. More, T.H., RoyChoudhury, S., Christie, J., Taunk, K., Mane, A., Santra, M.K., Chaudhury, K. and Rapole, S. (2018) Metabolomic alterations in invasive ductal carcinoma of breast: a comprehensive metabolomic study using tissue and serum samples. *Oncotarget*, **9**, 2678-2696.

73. Callister, S.J., Barry, R.C., Adkins, J.N., Johnson, E.T., Qian, W.J., Webb-Robertson, B.J., Smith, R.D. and Lipton, M.S. (2006) Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J Proteome Res*, **5**, 277-286.

74. Tiziani, S., Lopes, V. and Gunther, U.L. (2009) Early stage diagnosis of oral cancer using 1H NMR-based metabolomics. *Neoplasia*, **11**, 269-276.

75. van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K. and van der Werf,

M.J. (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, **7**, 142.

76. Benito, S., Sanchez-Ortega, A., Unceta, N., Jansen, J.J., Postma, G., Andrade, F., Aldamiz-Echevarria, L., Buydens, L.M.C., Goicolea, M.A. and Barrio, R.J. (2018) Plasma biomarker discovery for early chronic kidney disease diagnosis based on chemometric approaches using LC-QTOF targeted metabolomics data. *J Pharm Biomed Anal*, **149**, 46-56.

77. Brennan, L. (2014) NMR-based metabolomics: from sample preparation to applications in nutrition research. *Prog Nucl Magn Reson Spectrosc*, **83**, 42-49.

78. Valikangas, T., Suomi, T. and Elo, L.L. (2018) A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief Bioinform*, **19**, 1-11.

79. Fundel, K., Haag, J., Gebhard, P.M., Zimmer, R. and Aigner, T. (2008) Normalization strategies for mRNA expression data in cartilage research. *Osteoarthritis Cartilage*, **16**, 947-955.

80. Tobin, J., Walach, J., de Beer, D., Williams, P.J., Filzmoser, P. and Walczak, B. (2017) Untargeted analysis of chromatographic data for green and fermented rooibos: Problem with size effect removal. *J Chromatogr A*, **1525**, 109-115.

81. Parca, L., Beck, M., Bork, P. and Ori, A. (2018) Quantifying compartment-associated variations of protein abundance in proteomics data. *Mol Syst Biol*, **14**, e8131.

82. Cox, J., Hein, M.Y., Luber, C.A., Paron, I., Nagaraj, N. and Mann, M. (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics*, **13**, 2513-2526.

83. Astrand, M. (2003) Contrast normalization of oligonucleotide arrays. *J Comput Biol*, **10**, 95-102.

84. Kohl, S.M., Klein, M.S., Hochrein, J., Oefner, P.J., Spang, R. and Gronwald, W. (2012) State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics*, **8**, 146-160.

85. Hochrein, J., Zacharias, H.U., Taruttis, F., Samol, C., Engelmann, J.C., Spang, R., Oefner, P.J. and Gronwald, W. (2015) Data normalization of (1)H NMR metabolite fingerprinting data sets in the presence of unbalanced metabolite regulation. *J Proteome Res*, **14**, 3217-3228.

86. Saccenti, E. (2017) Correlation patterns in experimental data are affected by normalization procedures: consequences for data analysis and network inference. *J Proteome Res*, **16**,

619-634.

87. Workman, C., Jensen, L.J., Jarmer, H., Berka, R., Gautier, L., Nielser, H.B., Saxild, H.H., Nielsen, C., Brunak, S. and Knudsen, S. (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol*, **3**, research0048.

88. Puchades-Carrasco, L., Palomino-Schatzlein, M., Perez-Rambla, C. and Pineda-Lucena, A. (2016) Bioinformatics tools for the analysis of NMR metabolomics studies focused on the identification of clinically relevant biomarkers. *Brief Bioinform*, **17**, 541-552.

89. Karpievitch, Y.V., Nikolic, S.B., Wilson, R., Sharman, J.E. and Edwards, L.M. (2014) Metabolomics data normalization with EigenMS. *PLoS One*, **9**, e116221.

90. Karpievitch, Y.V., Dabney, A.R. and Smith, R.D. (2012) Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics*, **13**, 5.

91. Ejigu, B.A., Valkenborg, D., Baggerman, G., Vanaerschot, M., Witters, E., Dujardin, J.C., Burzykowski, T. and Berg, M. (2013) Evaluation of normalization methods to pave the way towards large-scale LC-MS-based metabolomics profiling experiments. *OMICS*, **17**, 473-485.

92. Karpievitch, Y.V., Taverner, T., Adkins, J.N., Callister, S.J., Anderson, G.A., Smith, R.D. and Dabney, A.R. (2009) Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition. *Bioinformatics*, **25**, 2573-2580.

93. Chen, X., de Seymour, J.V., Han, T.L., Xia, Y., Chen, C., Zhang, T., Zhang, H. and Baker, P.N. (2018) Metabolomic biomarkers and novel dietary factors associated with gestational diabetes in China. *Metabolomics*, **14**, 149.

94. Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185-193.

95. Ballman, K.V., Grill, D.E., Oberg, A.L. and Therneau, T.M. (2004) Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics*, **20**, 2778-2786.

96. Adriaens, M.E., Jaillard, M., Eijssen, L.M., Mayer, C.D. and Evelo, C.T. (2012) An evaluation of two-channel ChIP-on-chip and DNA methylation microarray normalization strategies. *BMC Genomics*, **13**, 42.

97. Backshall, A., Sharma, R., Clarke, S.J. and Keun, H.C. (2011) Pharmacometabonomic profiling as a predictor of toxicity in patients with inoperable colorectal cancer treated with capecitabine. *Clin Cancer Res*, **17**, 3019-3028.

98. Webb-Robertson, B.J., Kim, Y.M., Zink, E.M., Hallaian, K.A., Zhang, Q., Madupu, R., Waters, K.M. and Metz, T.O. (2014) A statistical analysis of the effects of urease pre-treatment on the measurement of the urinary metabolome by gas chromatography-mass spectrometry. *Metabolomics*, **10**, 897-908.

99. Gonzalez-Dominguez, R., Garcia-Barrera, T., Vitorica, J. and Gomez-Ariza, J.L. (2014) Region-specific metabolic alterations in the brain of the APP/PS1 transgenic mice of Alzheimer's disease. *Biochim Biophys Acta*, **1842**, 2395-2402.

100. Paredi, G., Mori, F., de Marino, M.G., Raboni, S., Marchi, L., Galati, S., Buschini, A., Lo Fiego, D.P. and Mozzarelli, A. (2019) Is the protein profile of pig Longissimus dorsi affected by gender and diet? *J Proteomics*, **206**, 103437.

101. De Livera, A.M., Dias, D.A., De Souza, D., Rupasinghe, T., Pyke, J., Tull, D., Roessner, U., McConville, M. and Speed, T.P. (2012) Normalizing and integrating metabolomics data. *Anal Chem*, **84**, 10768-10776.

102. Craig, A., Cloarec, O., Holmes, E., Nicholson, J.K. and Lindon, J.C. (2006) Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Anal Chem*, **78**, 2262-2267.

103. Puskarich, M.A., Finkel, M.A., Karnovsky, A., Jones, A.E., Trexel, J., Harris, B.N. and Stringer, K.A. (2015) Pharmacometabolomics of l-carnitine treatment response phenotypes in patients with septic shock. *Ann Am Thorac Soc*, **12**, 46-56.

104. Martinez-Lozano Sinues, P., Kohler, M. and Zenobi, R. (2013) Human breath analysis may support the existence of individual metabolic phenotypes. *PLoS One*, **8**, e59909.

105. Warrack, B.M., Hnatyshyn, S., Ott, K.H., Reily, M.D., Sanders, M., Zhang, H. and Drexler, D.M. (2009) Normalization strategies for metabonomic analysis of urine samples. *J Chromatogr B Analyt Technol Biomed Life Sci*, **877**, 547-552.

106. Jacob, C.C., Dervilly-Pinel, G., Biancotto, G. and Le Bizec, B. (2014) Evaluation of specific gravity as normalization strategy for cattle urinary metabolome analysis. *Metabolomics*, **10**, 627-637.

107. Mizuno, H., Ueda, K., Kobayashi, Y., Tsuyama, N., Todoroki, K., Min, J.Z. and Toyo'oka, T. (2017) The great importance of normalization of LC-MS data for highly-accurate non-targeted metabolomics. *Biomed Chromatogr*, **31**, e3864.

108. Dieterle, F., Ross, A., Schlotterbeck, G. and Senn, H. (2006) Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal Chem*, **78**, 4281-4290.

109. Emwas, A.H., Saccenti, E., Gao, X., McKay, R.T., Dos Santos, V., Roy, R. and Wishart, D.S. (2018) Recommended strategies for spectral processing and post-processing of 1D (1)H-NMR data of biofluids with a particular focus on urine. *Metabolomics*, **14**, 31.

110. Liang, Y.J., Lin, Y.T., Chen, C.W., Lin, C.W., Chao, K.M., Pan, W.H. and Yang, H.C. (2016) SMART: statistical metabolomics analysis-an R tool. *Anal Chem*, **88**, 6334-6341.

111. De Livera, A.M., Sysi-Aho, M., Jacob, L., Gagnon-Bartsch, J.A., Castillo, S., Simpson, J.A. and Speed, T.P. (2015) Statistical methods for handling unwanted variation in metabolomics data. *Anal Chem*, **87**, 3606-3615.

112. Smolinska, A., Hauschild, A.C., Fijten, R.R., Dallinga, J.W., Baumbach, J. and van Schooten, F.J. (2014) Current breathomics: a review on data pre-processing techniques and machine learning in metabolomics breath analysis. *J Breath Res*, **8**, 027105.

113. Jiang, L., Lee, S.C. and Ng, T.C. (2018) Pharmacometabonomics analysis reveals serum formate and acetate potentially associated with varying response to gemcitabine-carboplatin chemotherapy in metastatic breast cancer patients. *J Proteome Res*, **17**, 1248-1257.

114. Chung, R.H. and Kang, C.Y. (2019) A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification. *Gigascience*, **8**, giz045.

115. Xi, B., Gu, H., Baniasadi, H. and Raftery, D. (2014) Statistical analysis and modeling of mass spectrometry-based metabolomics data. *Methods Mol Biol*, **1198**, 333-353.

116. Wang, S.Y., Kuo, C.H. and Tseng, Y.J. (2013) Batch Normalizer: a fast total abundance regression calibration method to simultaneously adjust batch and injection order effects in liquid chromatography/time-of-flight mass spectrometry-based metabolomics data and comparison with current calibration methods. *Anal Chem*, **85**, 1037-1046.

117. Struck, W., Siluk, D., Yumba-Mpanga, A., Markuszewski, M., Kaliszan, R. and Markuszewski, M.J. (2013) Liquid chromatography tandem mass spectrometry study of urinary nucleosides as potential cancer markers. *J Chromatogr A*, **1283**, 122-131.

118. Masson, P., Spagou, K., Nicholson, J.K. and Want, E.J. (2011) Technical and biological variation in UPLC-MS-based untargeted metabolic profiling of liver extracts: application in an experimental toxicity study on galactosamine. *Anal Chem*, **83**, 1116-1123.

119. Li, H., Ni, Y., Su, M., Qiu, Y., Zhou, M., Qiu, M., Zhao, A., Zhao, L. and Jia, W. (2007) Pharmacometabonomic phenotyping reveals different responses to xenobiotic intervention in rats. *J Proteome Res*, **6**, 1364-1370.

120. Leichtle, A.B., Nuoffer, J.M., Ceglarek, U., Kase, J., Conrad, T., Witzigmann, H., Thiery, J. and Fiedler, G.M. (2012) Serum amino acid profiles and their alterations in colorectal cancer. *Metabolomics*, **8**, 643-653.

121. Smilde, A.K., van der Werf, M.J., Bijlsma, S., van der Werff-van der Vat, B.J. and Jellema, R.H. (2005) Fusion of mass spectrometry-based metabolomics data. *Anal Chem*, **77**, 6729-6736.

122. Parastar, H. and Bazrafshan, A. (2016) Fuzzy c-means clustering for chromatographic fingerprints analysis: a gas chromatography-mass spectrometry case study. *J Chromatogr A*, **1438**, 236-243.

123. Di Guida, R., Engel, J., Allwood, J.W., Weber, R.J., Jones, M.R., Sommer, U., Viant, M.R. and Dunn, W.B. (2016) Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics*, **12**, 93.

124. Keun, H.C., Ebbels, T.M.D., Antti, H., Bollard, M.E., Beckonert, O., Holmes, E., Lindon, J.C. and Nicholson, J.K. (2003) Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling. *Anal Chim Acta*, **490**, 265-276.

125. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, 96-104.

126. Rausch, T.K., Schillert, A., Ziegler, A., Luking, A., Zucht, H.D. and Schulz-Knappe, P. (2016) Comparison of pre-processing methods for multiplex bead-based immunoassays. *BMC Genomics*, **17**, 601.

127. Lin, S.M., Du, P., Huber, W. and Kibbe, W.A. (2008) Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res*, **36**, e11.

128. Zhang, S., Zheng, C., Lanza, I.R., Nair, K.S., Raftery, D. and Vitek, O. (2009) Interdependence of signal processing and analysis of urine 1H NMR spectra for metabolic profiling. *Anal Chem*, **81**, 6080-6088.

129. Redestig, H., Fukushima, A., Stenlund, H., Moritz, T., Arita, M., Saito, K. and Kusano, M. (2009) Compensation for systematic cross-contribution improves normalization of mass spectrometry based metabolomics data. *Anal Chem*, **81**, 7974-7980.

130. Sysi-Aho, M., Katajamaa, M., Yetukuri, L. and Oresic, M. (2007) Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC*

Bioinformatics, **8**, 93.

131.Li, B., Tang, J., Yang, Q., Li, S., Cui, X., Li, Y., Chen, Y., Xue, W., Li, X. and Zhu, F. (2017) NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Res*, **45**, W162-W170.

132.Perng, W., Ringham, B.M., Smith, H.A., Michelotti, G., Kechris, K.M. and Dabelea, D. (2020) A prospective study of associations between in utero exposure to gestational diabetes mellitus and metabolomic profiles during late childhood and adolescence. *Diabetologia*, **63**, 296-312.

133.Gullberg, J., Jonsson, P., Nordstrom, A., Sjostrom, M. and Moritz, T. (2004) Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of Arabidopsis thaliana samples in metabolomic studies with gas chromatography/mass spectrometry. *Anal Biochem*, **331**, 283-295.

134.Bijlsma, S., Bobeldijk, I., Verheij, E.R., Ramaker, R., Kochhar, S., Macdonald, I.A., van Ommen, B. and Smilde, A.K. (2006) Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Anal Chem*, **78**, 567-574.

135.Dai, W., Xie, D., Lu, M., Li, P., Lv, H., Yang, C., Peng, Q., Zhu, Y., Guo, L., Zhang, Y. *et al.* (2017) Characterization of white tea metabolome: comparison against green and black tea by a nontargeted metabolomics approach. *Food Res Int*, **96**, 40-45.

136.O'Callaghan, S., De Souza, D.P., Isaac, A., Wang, Q., Hodkinson, L., Olshansky, M., Erwin, T., Appelbe, B., Tull, D.L., Roessner, U. *et al.* (2012) PyMS: a python toolkit for processing of gas chromatography-mass spectrometry (GC-MS) data. *BMC Bioinformatics*, **13**, 115.

137.Schiffman, C., Petrick, L., Perttula, K., Yano, Y., Carlsson, H., Whitehead, T., Metayer, C., Hayes, J., Rappaport, S. and Dudoit, S. (2019) Filtering procedures for untargeted LC-MS metabolomics data. *BMC Bioinformatics*, **20**, 334.

138.Chen, J., Zhang, P., Lv, M., Guo, H., Huang, Y., Zhang, Z. and Xu, F. (2017) Influences of normalization method on biomarker discovery in gas chromatography-mass spectrometry-based untargeted metabolomics: what should be considered? *Anal Chem*, **89**, 5342-5348.

139.Dunn, W.B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., Brown, M., Knowles, J.D., Halsall, A., Haselden, J.N. *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.

140.Kokla, M., Virtanen, J., Kolehmainen, M., Paananen, J. and Hanhineva, K. (2019) Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. *BMC Bioinformatics*, **20**, 492.

141. Tang, J., Fu, J., Wang, Y., Li, B., Li, Y., Yang, Q., Cui, X., Hong, J., Li, X., Chen, Y. *et al.* (2020) ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Brief Bioinform*, **21**, 621-636.

142. Huan, T. and Li, L. (2015) Counting missing values in a metabolite-intensity data set for measuring the analytical performance of a metabolomics platform. *Anal Chem*, **87**, 1306-1313.

143. Kirwan, J.A., Weber, R.J., Broadhurst, D.I. and Viant, M.R. (2014) Direct infusion mass spectrometry metabolomics dataset: a benchmark for data processing and quality control. *Sci Data*, **1**, 140012.

144. Wang, S. and Yang, H. (2019) pseudoQC: a regression-based simulation software for correction and normalization of complex metabolomics and proteomics datasets. *Proteomics*, **19**, e1900264.

145. Ho, E.N., Kwok, W.H., Leung, D.K., Riggs, C.M., Sidlow, G., Stewart, B.D., Wong, A.S. and Wan, T.S. (2015) Control of the misuse of testosterone in castrated horses based on an international threshold in plasma. *Drug Test Anal*, **7**, 414-419.

146. Purohit, P.V., Rocke, D.M., Viant, M.R. and Woodruff, D.L. (2004) Discrimination models using variance-stabilizing transformation of metabolomic NMR data. *OMICS*, **8**, 118-130.

147. Xia, J. and Wishart, D.S. (2011) Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nat Protoc*, **6**, 743-760.

148. Trezzi, J.P., Jager, C., Galozzi, S., Barkovits, K., Marcus, K., Mollenhauer, B. and Hiller, K. (2017) Metabolic profiling of body fluids and multivariate data analysis. *MethodsX*, **4**, 95-103.

149. Willforss, J., Chawade, A. and Levander, F. (2019) NormalyzerDE: online tool for improved normalization of omics expression data and high-sensitivity differential expression analysis. *J Proteome Res*, **18**, 732-740.

150. Puhka, M., Takatalo, M., Nordberg, M.E., Valkonen, S., Nandania, J., Aatonen, M., Yliperttula, M., Laitinen, S., Velagapudi, V., Mirtti, T. *et al.* (2017) Metabolomic profiling of extracellular vesicles and alternative normalization methods reveal enriched metabolites and strategies to study prostate cancer-related changes. *Theranostics*, **7**, 3824-3841.

151. Chawade, A., Alexandersson, E. and Levander, F. (2014) Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. *J Proteome Res*, **13**, 3114-3120.

152. Fu, J., Tang, J., Wang, Y., Cui, X., Yang, Q., Hong, J., Li, X., Li, S., Chen, Y., Xue, W. *et al.* (2018) Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification. *Front Pharmacol*, **9**, 681.

153. Tai, Y.C. and Speed, T.P. (2009) On gene ranking using replicated microarray time course data. *Biometrics*, **65**, 40-51.

154. Li, P., Tang, H., Shi, C., Xie, Y., Zhou, H., Xia, B., Zhang, C., Chen, L. and Jiang, L. (2019) Untargeted metabolomics analysis of Mucor racemosus Douchi fermentation process by gas chromatography with time-of-flight mass spectrometry. *Food Sci Nutr*, **7**, 1865-1874.

155. Thevenot, E.A., Roux, A., Xu, Y., Ezan, E. and Junot, C. (2015) Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. *J Proteome Res*, **14**, 3322-3335.

156. Yang, Q., Xu, L., Tang, L.J., Yang, J.T., Wu, B.Q., Chen, N., Jiang, J.H. and Yu, R.Q. (2018) Simultaneous detection of multiple inherited metabolic diseases using GC-MS urinary metabolomics by chemometrics multi-class classification strategies. *Talanta*, **186**, 489-496.

157. Jacob, S., Nodzenski, M., Reisetter, A.C., Bain, J.R., Muehlbauer, M.J., Stevens, R.D., Ilkayeva, O.R., Lowe, L.P., Metzger, B.E., Newgard, C.B. *et al.* (2017) Targeted metabolomics demonstrates distinct and overlapping maternal metabolites associated with BMI, glucose, and insulin sensitivity during pregnancy across four ancestry groups. *Diabetes Care*, **40**, 911-919.

158. Jiang, H., Sohn, L.L., Huang, H. and Chen, L. (2018) Single cell clustering based on cell-pair differentiability correlation and variance analysis. *Bioinformatics*, **34**, 3684-3694.

159. Huang, S., Cheng, Y., Lang, D., Chi, R. and Liu, G. (2014) A formal algorithm for verifying the validity of clustering results based on model checking. *PLoS One*, **9**, e90109.

160. Wang, X., Gardiner, E.J. and Cairns, M.J. (2015) Optimal consistency in microRNA expression analysis using reference-gene-based normalization. *Mol Biosyst*, **11**, 1235-1240.

161. Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J. *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671-679.

162. Onderwater, G.L.J., Ligthart, L., Bot, M., Demirkan, A., Fu, J., van der Kallen, C.J.H., Vijfhuizen, L.S., Pool, R., Liu, J., Vanmolkot, F.H.M. *et al.* (2019) Large-scale plasma metabolome analysis reveals alterations in HDL metabolism in migraine. *Neurology*, **92**, e1899-e1911.

163. Setia, M.S. (2016) Methodology series module 5: sampling strategies. *Indian J Dermatol*, **61**, 505-509.

164. Somol, P. and Novovicova, J. (2010) Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. *IEEE Trans Pattern Anal Mach Intell*, **32**, 1921-1939.

165. Peeters, L., Beirnaert, C., Van der Auwera, A., Bijttebier, S., De Bruyne, T., Laukens, K., Pieters, L., Hermans, N. and Foubert, K. (2019) Revelation of the metabolic pathway of hederacoside C using an innovative data analysis strategy for dynamic multiclass biotransformation experiments. *J Chromatogr A*, **1595**, 240-247.

166. Peters, S., Janssen, H.G. and Vivo-Truyols, G. (2010) Trend analysis of time-series data: a novel method for untargeted metabolite discovery. *Anal Chim Acta*, **663**, 98-104.

167. Risso, D., Ngai, J., Speed, T.P. and Dudoit, S. (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol*, **32**, 896-902.

168. Jia, Z. (2017) Controlling the overfitting of heritability in genomic selection through cross validation. *Sci Rep*, **7**, 13678.

169. Cinelli, M., Sun, Y., Best, K., Heather, J.M., Reich-Zeliger, S., Shifrut, E., Friedman, N., Shawe-Taylor, J. and Chain, B. (2017) Feature selection using a one dimensional naive Bayes' classifier increases the accuracy of support vector machine classification of CDR3 repertoires. *Bioinformatics*, **33**, 951-955.

170. Jiang, J., Yin, X.Y., Song, X.W., Xie, D., Xu, H.J., Yang, J. and Sun, L.R. (2018) EgoNet identifies differential ego-modules and pathways related to prednisolone resistance in childhood acute lymphoblastic leukemia. *Hematology*, **23**, 221-227.

171. Gillis, J. and Pavlidis, P. (2011) The role of indirect connections in gene networks in predicting function. *Bioinformatics*, **27**, 1860-1866.