Briefings in Bioinformatics, 22(5), 2021, 1–25

https://doi.org/10.1093/bib/bbab138
Method Review

Pharmacometabonomics: data processing and statistical analysis

Jianbo Fu†, Ying Zhang†, Jin Liu, Xichen Lian, Jing Tang and Feng Zhu

Corresponding author. Prof. Feng Zhu, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China.
Tel.: +86-0571-88208444; E-mail: zhufeng@zju.edu.cn
†These authors contributed equally to the paper as co-first authors.

Abstract

Individual variations in drug efficacy, side effects and adverse drug reactions are still challenging that cannot be ignored in drug research and development. The aim of pharmacometabonomics is to better understand the pharmacokinetic properties of drugs and monitor the drug effects on specific metabolic pathways. Here, we systematically reviewed the recent technological advances in pharmacometabonomics for better understanding the pathophysiological mechanisms of diseases as well as the metabolic effects of drugs on bodies. First, the advantages and disadvantages of all mainstream analytical techniques were compared. Second, many data processing strategies including filtering, missing value imputation, quality control-based correction, transformation, normalization together with the methods implemented in each step were discussed. Third, various feature selection and feature extraction algorithms commonly applied in pharmacometabonomics were described. Finally, the databases that facilitate current pharmacometabonomics were collected and discussed. All in all, this review provided guidance for researchers engaged in pharmacometabonomics and metabolomics, and it would promote the wide application of metabolomics in drug research and personalized medicine.

Key words: pharmacometabonomics; precision medicine; analytical technique; data processing; statistical analysis

Introduction

Diseases are commonly complicated and concerned with dysregulation of multiple biological pathways, including neuropsychiatric disorders such as depressive disorders [1–3], cardiovascular diseases such as atherosclerosis [4], metabolic diseases such as diabetes mellitus [5], and cancers [6–11]. Despite the development and clinical application of various drugs with pharmacotherapeutic potential, the intrinsic diversity in disease subtypes [12, 13] coupled with individual variability in efficacy, side effects [14] and adverse drug reactions (ADRs) of drugs [15–18] are still non-negligible challenges in pharmaceutical researches. Severe shortcomings of empiric therapy reinforce the urgent need to discover and validate biomarkers which are beneficial for understanding the mechanisms of diseases, detecting and diagnosing diseases, guiding drug development and clinical selection, and stably predicting individual differences in response to the same drug treatment [19–24]. Consequently, precision medicine and individualized treatment have emerged recently, which utilize high throughput omics technologies and computational resources to overcome these challenges [25–27].

Numerous factors collectively contributed to differential drug response phenotypes, mainly including genetics and environmental influences [28]. And multiple omics-based

Jianbo Fu is a PhD student in the College of Pharmaceutical Sciences in Zhejiang University, China. He is interested in bioinformatics and metabolomics.
Ying Zhang is a PhD student in the College of Pharmaceutical Sciences in Zhejiang University, China. She is interested in bioinformatics and metabolomics.
Jin Liu is a PhD student in the College of Pharmaceutical Sciences in Zhejiang University, China. She is interested in bioinformatics and molecular biology.
Xichen Lian is a PhD student in the College of Pharmaceutical Sciences in Zhejiang University, China. She is interested in bioinformatics.
Jing Tang is an Associate Professor of the Department of Bioinformatics in Chongqing Medical University, China. She is interested in the area of metabolomics and system biology.
Feng Zhu is a Professor at the College of Pharmaceutical Sciences in Zhejiang University, China. His research laboratory (https://idrblab.org/) has been working in the fields of bioinformatics, OMIC-based drug discovery, system biology and medicinal chemistry.
Submitted: 17 January 2021; **Received (in revised form):** 9 February 2021

Figure 1. Summary and classification of pharmacometabonomics applications of all bioinformatics tools discussed in this review.

approaches such as pharmacogenomics, pharmacotranscriptomics, pharmacoproteomics and pharmacometabonomics have been put forward in succession [29–32]. Different from the other three omics techniques, pharmacometabonomics identifies something that is actually happening rather than something may happen [33]. Pharmacometabonomics, first proposed by Clayton *et al.* in 2006 [32], was described as the prediction and evaluation of the response (for instance, therapeutic efficacy or toxicity) of pharmaceutical compounds individually based on statistical models of pre-treatment metabolic signatures. Broadly speaking, pharmacometabonomics refers to the quantitative measurement and analysis of metabolites produced by the body in pre-, during- and post-treatment, with the aim of understanding the pharmacokinetic properties of drugs better and monitoring the effects of drugs on specific metabolic pathways (pharmacodynamics) [34–39]. As shown in Figure 1, main applications in the field of pharmacometabonomics fall into the following categories [34, 35, 40]: (i) Drug discovery: the investigation into the drug and target for identifying the disease biomarkers, understanding the pharmacokinetics (PK)/pharmacodynamics (PD) properties of drugs on the human body and identifying/validating therapeutic targets; (ii) Clinical research: the detection of metabolic changes resulted from drug exposure as well as the variations in metabolic features of a drug among different conditions (e.g. drug versus placebo), the selection of key features which reliably distinguish drug response phenotypes (good responders versus poor responders, therapeutic effects versus side effects or ADRs) based on variations in baseline metabolism, and the utilization in drug safety assessment; (iii) Personalized medicine: the measurement of metabolite concentration changes in human to determine disease status, the identification of diagnostic/prognostic biomarkers, and the prediction of individualized drug responses.

As a discipline stems from metabolomics, the general analytical workflow of pharmacometabonomics studies [34, 41–47] schematically shown in Figure 2 includes: (i) Sample preparation: the collection and extraction of samples of interest (tissue biopsies, biofluids, etc.), (ii) Data acquisition: the separation and quantification of the molecules of interest based on analytical platforms, including the nuclear magnetic resonance spectrometry (NMR) and/or hyphenated mass spectrometry (MS) such as Liquid Chromatography coupled with Mass Spectrometry (LC–MS) and Gas Chromatography coupled with Mass Spectrometry (GC–MS), (iii) Data preprocessing: the collection and curation of raw instrumental signals acquired from analytical platforms, and the format conversion of these original data into sample-by-feature tables by commercial or open source software, which facilitate the subsequent data processing and statistical analysis, (iv) Data processing: the employment of multiple processing procedures to transform the raw data matrix with the aim of improving the quality of data, such as normality and comparability, (v) Statistical analysis: the comprehensive and flexible application of various univariate and/or multivariate statistics to reveal discriminant metabolites, (vi) Metabolite identification: the putative metabolites searching based on metabolic databases followed by the validation of them and (vii) Interpretation: the biological interpretations based on the association of altered metabolites to corresponding metabolic pathways based on

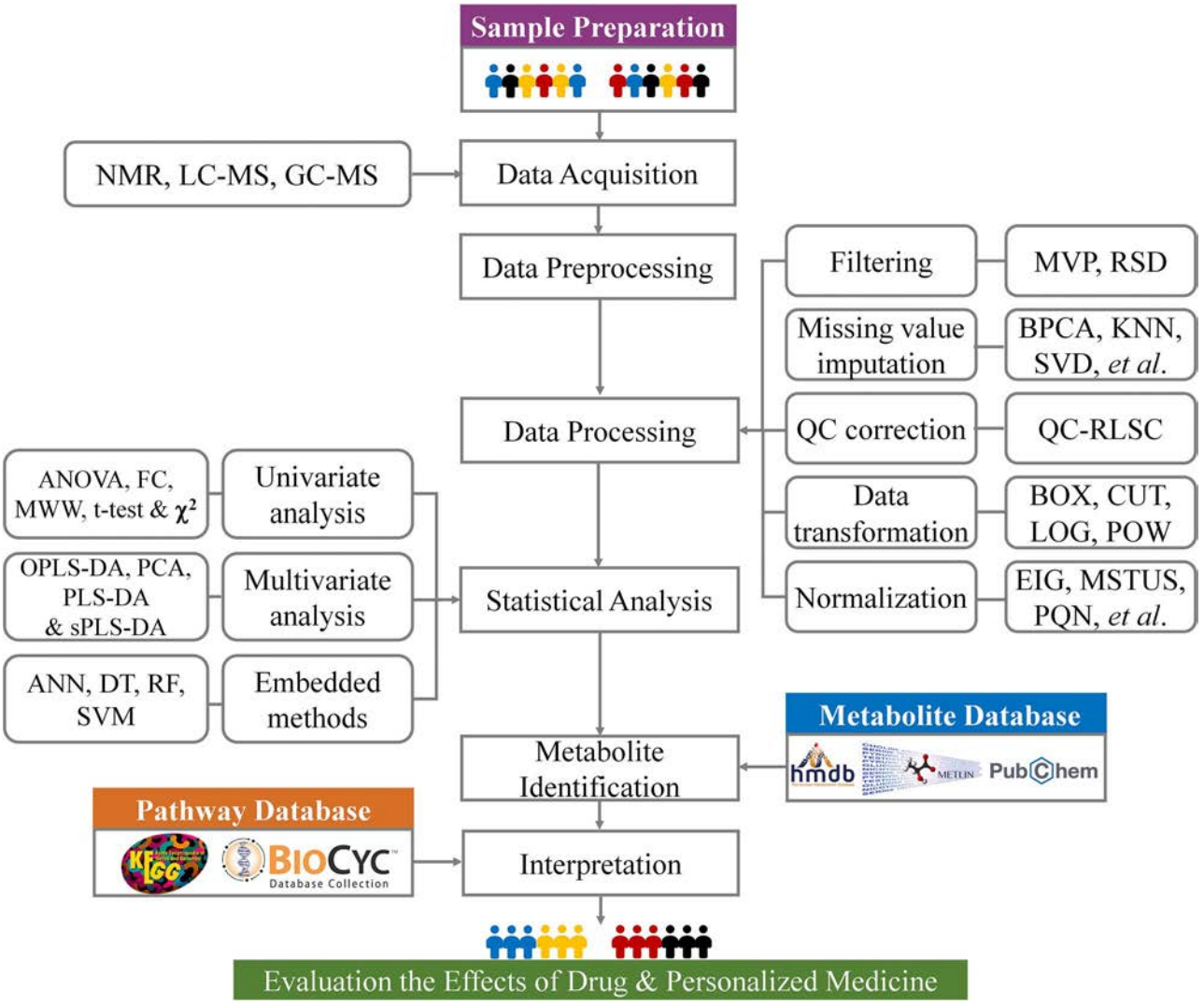

Figure 2. The schematic representing general workflow of pharmacometabonomics studies, from sample preparation to biological interpretation.

metabolic databases. In this study, we systematically reviewed recent technological advances in pharmacometabonomics based on the main analytical steps as mentioned above.

The field of separation and quantitative analysis of compounds has grown by leaps and bounds in recent years, which significantly promotes the development of analysis techniques in pharmacometabonomics. The mainstream analytical platforms commonly utilized for data acquisition in pharmacometabonomics are NMR and MS [21, 48–50]. For example, a latest study used the NMR technique to identify novel biomarkers of warfarin, which could distinguish warfarin responses based on the international normalized ratio with good accuracy [51]. Another study based on the ultra-high-performance liquid chromatography coupled with high-resolution mass spectrometry (UPLC-HRMS) performed targeted neurotransmitter quantitative analysis and non-targeted metabolic profiling for pharmacometabonomics analysis of olanzapine, and identified significantly down-regulated/upregulated metabolites for providing insights on interpreting the pharmacodynamic effect and mechanism of olanzapine [52]. Although these technologies have served as the primary workhorses and been widely used in multiple aspects of pharmacometabonomics studies, each technique has its own strengths and weaknesses. Inevitably, many technical challenges of curation and analyzing of the obtained data still remain under consideration. High-dimensional information [53, 54] generated by various sophisticated analytical techniques [55, 56] usually contains large proportions of uninformative features [57] and non-negligible amounts of missing values [58], accompanied with properties such as heteroscedasticity [59], skewness [60] and biological variability [61]. To cope with these pivotal technical challenges which hamper the discovery of biomarkers in pharmacometabonomics, the quantitative performances of the analytical platforms have been improved and multiple computational algorithms and software have been developed [62–66]. Various data processing procedures and methods, each with their own underlying theory, have been proposed in succession and extensively used in pharmacometabonomics studies to determine significantly altered metabolites as biomarkers. In this case, it is of increasing importance to review features and underlying algorithms of these procedures and methods for proper selection and application on specific datasets with different intrinsic properties.

Both feature selection and feature extraction are powerful strategies for reducing dimensionality [67], which significantly simplified the analysis of pharmacometabonomics data with high dimensionality. Feature selection methods can pick out markedly altered features from hundreds or thousands of

metabolites among distinct conditions (e.g. healthy control versus patients, drug treatment versus placebo treatment, good versus poor responders). And the feature extraction methods can reduce dimensionality and facilitate the classification of samples. So far, a variety of feature selection approaches have been successfully and extensively applied in pharmacometabonomics to identify biomarkers and to further provide corresponding hidden biological interpretations [36, 37, 39, 68, 69]. Kaddurah-Daouk *et al.* [70] constructed the partial least squares discriminant analysis model (PLS-DA) to investigate the baseline metabolic predictors of response to placebo or sertraline in depressed outpatients, which distinguished outpatients between who did and did not respond to treatment with placebo or sertraline. In another work, Trupp *et al.* [38] made use of the orthogonal partial least squares discriminant analysis (OPLS-DA) model to analyze baseline metabolic signatures of good and poor low-density lipoprotein cholesterol responders to simvastatin for the determination of biomarkers that best defined distinct metabotypes. However, the consistency and reproducibility of the biomarkers selected by feature selection methods remain ambiguous owing to the absence of robustness [71]. Consequently, the performance of several popular feature selection algorithms used in metabolomics studies were comprehensively assessed and compared to provide guidelines on better determining proper methods for specific metabolomics datasets [72–74]. Nevertheless, no such review has been reported yet on pharmacometabonomics studies.

Moreover, accompanied with the advent of the era of big data as well as the booming development of computational statistical software and algorithms, numerous computerized databases have been constructed and further used throughout pharmacometabonomics studies [75–77]. And it makes sense to review and highlight the concepts and applications of these technical tools against the background of rapid multiplication of multifarious tools.

In this review, the recent technological advances in pharmacometabonomics for better understanding the pathophysiological mechanisms of diseases as well as the metabolic effects of drugs on bodies were systematically reviewed. First, the advantages and disadvantages of all mainstream analytical techniques were compared. Second, many data processing strategies including filtering, missing value imputation, quality control (QC)-based correction, transformation, normalization together with the methods implemented in each step were discussed. Third, various feature selection and feature extraction algorithms commonly applied in pharmacometabonomics were described. Finally, the databases that facilitate current pharmacometabonomics were collected and discussed. To the best of our knowledge, this review was the first and the most comprehensive one providing exhaustive discussion, systematic classification and rational advice on applications of popular data manipulation methods in pharmacometabonomics studies, while most of the other reviews focused on reviewing the history, discussing multiple applications of pharmacometabonomics, and briefly introducing the general analytical workflow of pharmacometabonomics studies based on MS and/or NMR [15, 78–81].

Mainstream analytical techniques for pharmacometabonomics

Researches on the application of various separation and quantification techniques to metabolic profiling of biological samples could be traced back to the work of Dalgliesh *et al.* [82] carried out in last century, in which metabolites in urine were separated by two-dimensional paper chromatography to produce the so called 'map of spots'. With the increasing advancement of pulse-Fourier transform proton NMR spectroscopy [80] as well as a variety of hyphenated MS techniques [39, 83] which were capable of analyzing hundreds or thousands of metabolites in a single experimental run [84, 85], the use of these analytical platforms in pharmacometabonomics significantly increased. Up to now, there are three mainstream analytical techniques (NMR, LC–MS and GC–MS), which provide convenience for efficient pharmacometabonomics studies. The comprehensive assessments of strengths and weaknesses in pharmacometabonomics of these quantitative techniques were briefly shown in Table 1.

NMR

The basic theory of NMR is the induction of the energy level transitions of NMR-active nuclei of certain compounds with a highly homogeneous and strong magnetic field [86, 87]. Take the representative one-dimensional NMR experiment as an example, the biofluid samples are kept in NMR tubes and further inserted into probes under a generated magnetic field, where the absorption of electromagnetic radiation and the energy level transitions occur based on the variations across magnetic levels. Then, the excited nuclei relax to ground state and cause the decaying oscillating current in receiver coils that last for several seconds, which can be detected by the spectrometer. Finally, the free induction decay signals are transformed into classical frequency-domain NMR spectrums which contain the correspondence of nuclear resonance frequencies and signal intensities.

NMR is one of the most dominant methods used for illuminating the structure of small molecular compounds and has played a crucial role in detecting, identifying and quantifying metabolites in biological samples, especially in biofluids for pharmacometabonomics studies. Particularly, the first demonstration of NMR in pharmacometabonomics was conducted by Clayton *et al.* [32] for the metabolic prediction of rats administered with paracetamol (acetaminophen). In addition to pharmacometabonomics studies being carried on preclinically, a host of clinical studies have been published. Clayton *et al.* [88] continued their research for the metabolic prediction of paracetamol in humans and finally identified a host-microbiome co-metabolite predictor by using ^{1}H-NMR spectroscopy. Moreover, the applications of NMR in pharmacometabonomics studies also included the prediction of drug efficacy [51, 89–91], side effects and ADRs [32, 92–98].

Hyphenated MS (LC–MS and GC–MS)

MS is another powerful quantification technology for metabolic profiling, which quantitively measures the mass-to-charge ratios of charged ions. In general, metabolites to be analyzed are first broken down into multiple fragments and ionized in the ion source to generate charged ions. Then, the ion beam is formed by the accelerating electric field and detected by mass analyzer. Due to the excellent capabilities of separation for complex samples and highly selective and sensitive detection belonged to chromatography and MS respectively, MS is mostly commonly combined with separative chromatography such as gas chromatography (GC), high/ultra-performance liquid chromatography. In this case, information of retention time, mass-to-charge and peak intensities are provided simultaneously, which can be utilized for metabolite identification and further quantitative analysis [99–102].

Table 1. Comprehensive comparisons and assessments of mainstream analytical techniques applied in pharmacometabonomics studies

Analytical techniques	Strengths	Weaknesses
NMR [296–301]	1. Non-invasive measurement without sample destruction so that samples can be recovered and analyzed for multiple times. 2. High selectivity and resolution. 3. Superiorities in illuminating the structure of small molecular compounds. 4. The permission of quantification of all metabolites containing NMR-active nuclei with detectable concentration levels by an NMR spectrum based on only one reference compound. 5. 5. The use of crude extracts without purification of samples and/or separation of metabolites.	1. Low sensitivity with detection limits in the range from mM to μM, but can be improved by higher magnetic fields, low temperature, microprobes and the dynamic nuclear polarization. 2. High investment of the instrument and equipment.
LC–MS [302–306]	1. High sensitivity with detection limits in the range from mM to pM. 2. High selectivity and resolution. 3. Suitable for the analysis of metabolites which are unstable, hard to derivatize, not easy to volatilize and/or with large molecular weights.	1. Destructive measurement, but only a few samples are required. 2. With relatively few databases and limitations in metabolite identification.
GC–MS [43, 104, 307–309]	1. High sensitivity with detection limits in the range from mM to pM. 2. High selectivity and resolution. 3. Suitable for the analysis of metabolites which are easy to derivatize, with low polarity and/or with small molecular weights. 4. With relatively sound databases for metabolite identification.	1. Destructive measurement, but only a few samples are required. 2. Require somewhat more complicated processes of sample preparation. For instance, non-volatile or semi-volatile compounds need to be derivatized in prior to further analysis.

Due to the volatilization process at the beginning of GC, the types of biological samples suitable for GC–MS and LC–MS analysis differ considerably [102]. LC–MS is suitable for the study of metabolites which are unstable, hard to derivatize, not easy to volatilize and/or with large molecular weights, whereas GC–MS is more applicable for metabolites which are easy to derivatize, with low polarity and/or with small molecular weights. Therefore, the separation analysis by GC–MS often requires additional derivatization of metabolites so as to ensure the volatilization before going through the chromatographic column [102–104]. However, variable derivatization may occur owing to the different derivatization efficiency between different metabolites, and the existence of mono-derivatization, di-derivatization and/or poly-derivatization of partial metabolites will generate various fragment ions and significantly make it harder to interpret the spectrum. Except for differences in the types of samples suitable for analysis, another difference between LC–MS and GC–MS lies in the ionization methods used. Electron impact ionization is usually used in GC–MS experiments, which belongs to hard ionization technologies resulting in fragmentations of metabolites [105, 106]. In contrast, the most commonly used ionization technology in LC–MS mode is electrospray ionization in positive and/or negative ionization modes, which belongs to soft ionization technologies and generates fewer fragments [107, 108].

Plenty of publications have proved the superior capabilities of separation and detection of hyphenated MS in pharmacometabonomics studies [109–119]. The first application of pharmacometabonomics-informed pharmacogenomics research strategy to precision medicine used the GC–MS technique to identify outcome biomarkers of citalopram/escitalopram treatment for patients with major depressive disorder [68]. For the prediction of pharmacokinetics and drug metabolism, Navarro *et al.* [120] successfully identified the predictive biomarkers of intravenous busulfan clearance of hematopoietic cell transplant recipients relying on the targeted LC–MS/MS analytical platform against 200 standard metabolites. Muhrez *et al.* [121] determined urine metabolites by GC–MS and tried to evaluate whether the baseline metabolic profiles of high-dose-methotrexate administration were predictive of clearance and/or toxicity in adult patients with lymphoid malignancies. Moreover, for the prediction of side effects and ADRs, an integrated LC–MS and GC–MS pharmacometabonomics analysis was conducted to determine the associations between the variability in toxic response of rats to lipopolysaccharide treatment and the predose serum metabolic profiles [122]. In another work, untargeted LC–MS combined with GC–MS metabolomics analysis were performed for recognizing the individual metabolic differences in rats treated with cisplatin and predicted nephrotoxicity with accuracy of 85%, which brought insights into nephrotoxicity and personalized medicine of cisplatin in clinical studies [123].

Data processing methods for pharmacometabonomics datasets

After data acquisition and corresponding preprocessing methods, a raw data matrix containing rows as observations (samples)

and columns as variables (features) is generated. Due to the escalation of data complexity and the intrinsic properties of the data, such as the inclusion of uninformative features [57] and missing values [58], heteroscedasticity [59], skewness [60] and biological variability [61], processing of the data is required and employed to improve the quality of the data by transforming the raw data matrix into a more 'cleaner' one [124]. As various data processing procedures and corresponding methods with different underlying theories have been proposed and widely used in pharmacometabonomics, the choice of robust and accurate methods is crucial to prevent possible errors during multiple processing steps between the original matrix and the statistical output. Even though some earlier literatures have discussed and compared part of available processing methods, most of them were not comprehensive with focus on only a particular processing step or only a small number of methods [125–127]. Therefore, key procedures and as many methods of pharmacometabonomics data processing were fully discussed below for providing guidance on making proper decisions in each step of the processing pipeline with respect to specific datasets. Five sequential data processing procedures as suggested by the Metabolomics Standards Initiative [53] and some other literatures were discussed as follows. And the Table 2 summarized these procedures with their corresponding methods that are commonly used in pharmacometabonomics studies.

Methods for filtering the pharmacometabonomics data

Filtering was advised to be implemented first by many literatures [128–130]. It refers to selectively removing data points that are uninformative or with low quality. The filtering procedure is considered to improve the data quality and to help in lowering the false discovery rate during downstream statistical analysis [124]. Here, two common filtering methods that are usually used in pharmacometabonomics studies were introduced.

The Percent of Missing Values (MVP) method calculates the percentage of missing values of each feature based on the precondition that a representative quantitative metabolomics dataset usually involves a large number of missing values [131]. Features are directly discarded if they are missing in more than a user-set percentage (default 20% by experience, which is called '80% rule') of samples [125]. This method was employed to remove missing peaks in a pharmacometabonomics study for revealing the therapeutic mechanism of HuangQi injections in rats with cisplatin-induced nephrotoxicity [132].

The Relative Standard Deviation (RSD) method calculates the value absolutely representing the inter-batch variations (i.e. the coefficient of variation) [133]. The lower the RSD of a specific feature, the better analytical reproducibility of the feature among batches indicated. And the metabolic feature will be removed from the data matrix if the RSD value of it among QCs is higher than the threshold value predefined. The predetermined threshold value (default 30%) is determined based on the experience for metabolomics studies, which is considered to be stable enough for prolonged analysis [134]. This method has been utilized to filter chromatograms with the threshold of RSD (less than 30% in QCs) in a pharmacometabonomics study of identifying predictors of cytarabine and anthracycline-treated chemosensitivity in patients with acute myeloid leukemia [135]. In addition to performing the RSD method, the chromatograms obtained from the real samples were also filtered by the MVP method in this study.

Methods for imputing missing values of the pharmacometabonomics data

Due to technical and/or biological reasons, metabolomics datasets usually contain ~20–30% missing values [131]. Given the integrality of the dataset that some transformation and normalization methods require [136, 137], it is common to apply missing value imputation followed by transformation and normalization. Moreover, imputation is performed to obtain coherent and complete dataset, which is usually recognized as a prerequisite for reducing the bias and ensuring robust and accurate statistical analysis [138]. Therefore, seven missing value imputation methods were described for proper selection for pharmacometabonomics datasets.

By obtaining the values generated from the Bayesian principal component analysis (BPCA) regression, the algorithm of the BPCA imputation fills the missing values within the data matrix. To be more specific, it combines Bayesian estimation and an expectation maximization algorithm with PCA regression [139]. In this case, each missing value imputed will not occur multiple times among the dataset, either across the samples nor across the metabolites [140]. BPCA outperforms the K-nearest Neighbor Imputation (KNN) and Singular Value Decomposition (SVD) imputation methods due to its ability to select the estimation parameters automatically [141]. BPCA has been applied in processing non-targeted ultra-high performance liquid chromatography coupled with mass spectrometry (UHPLC–MS) metabolomics data [142] and handling missing covariates in epidemiologic studies [143].

The Half of the Minimum Positive Value (HAM)/Background Imputation method substitutes missing values with the half of the minimum positive values with respect to the corresponding variable [141, 144], which is likely to reduce differences between diverse experimental groups and result in less statistical power [58]. The HAM method has helped to predict different lipopolysaccharide-induced lipid metabolomic profiles in survival and non-survival rats [122]. In another pharmacometabonomics study, this method acted as an imputation step to demonstrate the synergistic killing of the combination of polymyxin B and mitotane against multidrug-resistant Acinetobacter baumannii [145].

The KNN algorithm is designed to search k metabolites of interest that are close to the metabolites containing missing values. The similarity between metabolites is determined by the Euclidean distance, and the metabolite with missing values are imputed with the weighted mean of k-nearest metabolites [146]. To be more specific, if we process metabolite A that contains a missing value in experiment I, KNN aims to select k metabolites whose intensity in experiment I is non-missing and in other experiments is most similar to metabolite A [147]. KNN appears to provide a more superior performance than SVD and BPCA and has been applied in multi-omics integrative analysis [146, 147]. Furthermore, the KNN method also has been used to process pharmacometabonomics data of early prediction of vincristine-induced peripheral neuropathy [148].

The algorithm of Mean Imputation (MDI) fills the specific feature containing missing values with the average value of remaining positive values. Compared with HAM method which reduces differences between diverse experimental groups, mean imputation tends to increase differences between diverse experimental groups and reduce variance simultaneously, which results in more rejections of the null hypothesis of no difference. This MDI has been used to invesitigate biomarkers of metformin exposure

Table 2. Summary of five sequential data processing procedures (filtering, missing value imputation, QC-based correction, transformation and normalization based on samples, metabolites or ISs) available for pharmacometabonomics studies, coupled with the accessibility in R packages and applications in pharmacometabonomics of various methods in each procedure

Procedure and method (abbr.)		R package (function)	Applications in pharmacometabonomics study
Filtering	MVP	base packages	Removing missing peaks for discovering the therapeutic mechanism of HuangQi injections in cisplatin-induced nephrotoxic rats [132].
	RSD	goeveg (cv)	Filtering chromatograms to find dodecanamide and leukotriene B4 dimethylamide (LTB4-DMA) as a predictor of chemosensitivity to cytarabine plus anthracycline chemotherapy for patients with acute myeloid leukemia [135].
Missing value imputation	BPCA	pcaMethods (bpca)	Processing untargeted UHPLC–MS datasets acquired for different samples of mouse serum, placental tissue extracts, human urine and mammalian cellular extracts [142].
	HAM	base packages	Replacing missing values with a half of the minimum value to help to select sphingosine, sphinganine, palmitic acid, oleic acid and cholesterol as predictors for variable LPS responses in survival and non-survival rats [122].
	KNN	impute (impute.knn)	Enabling early prediction of peripheral neuropathy in patients with pediatric leukemia [148].
	MDI	base packages	Imputing missing values of metabolite data in a pharmacometabonomic assessment research of metformin administration in non-diabetic African Americans [149].
	MEI	base packages	Imputing missing values in prior to hierarchical clustering for analyzing significant metabolites of participants treated with atenolol and hydrochlorothiazide [150].
	SVD	pcaMethods (svdImpute)	Using urine metabolomics to understand the pathogenesis of infants infected with respiratory syncytial virus and the role of respiratory syncytial virus in childhood wheezing [153].
	ZER	base packages	Dealing with missing values in hepatocellular carcinoma metabolomics dataset and helping to identify four upregulated and two downregulated metabolites as potential biomarkers of hepatocellular carcinoma [154].
QC correction	QC-RLSC	statTarget (shiftCor)	Correcting for signal shifts and batch effects in a study of identifying the serum metabolomic alterations in Beagle dogs with Toxocara canis infection [164].
Transformation	BOX	AID (boxcoxfr)	Helping to identify serum biomarkers of cholangiocarcinoma, hepatocellular carcinoma and primary sclerosing cholangitis for diagnosis [179].
	CUT	pamr (pamr.cube.roo)	Utilized with the combination of other processing and statistical tools in MetaboAnalyst 3.0 for revealing the role for histidine, phenylalanine and threonine in the development of paclitaxel-caused peripheral neuropathy [177].
	LOG	metabolomics (LogTransform)	Correcting for the positively skewed distribution of the pharmacometabonomics data for discovering time-dependent alternations in urinary metabolome induced by intensive phase tuberculosis therapy [174].
	SRT	base packages	Reducing the influence of the skewed distribution and heteroscedasticity of the data, and thus facilitating the maternal serum metabolomic fingerprint-based study of diagnostic performance evaluation of a machine learning ensemble model [180].
Sample-based normalization	CON	affy (normalize.AffyBatch.contrasts)	Correcting for unwanted experimental or biological variations in LC/MS-based untargeted metabolomics analysis [197].
	CUB	affy (normalize.qspline)	Reducing batch variation and help to correctly classify samples with the aim of identifying clinically relevant biomarkers [198].
	CYC	affy (normalize.loess)	Removing the systematic effect in a study where the metabolic alterations in the brain of the APP/PS1 mice with Alzheimer's disease was observed [199].
	EIG	ProteoMM (eig_norm1, eig_norm2)	Detecting and correcting for systematic bias to assist the identification of metabolomic biomarkers and novel dietary factors related to gestational diabetes in China [200].
	LIN	affy (normalize.scaling)	Correcting for systematic variation for predicting capecitabine-induced toxicity in patients with inoperable colorectal cancer by pharmacometabonomic profiling [95].
	LIW	affy (normalize.AffyBatch.invariantset)	Removing unwanted sample-to-sample variation in LC–MS-based untargeted metabolomics analysis [197].

(*Continued*)

Table 2. Continued

Procedure and method (abbr.)		R package (function)	Applications in pharmacometabonomics study
	MEA	metabolomics (Normalise)	Eliminating background effects for differentiating L-carnitine response phenotypes in patients with vasopressor-dependent septic shock [90].
	MED	metabolomics (Normalise)	Normalizing the breath mass spectra in a pharmacometabonomics analysis of the confirmation of the existence of highly individual phenotypes [201].
	MSTUS	base packages	Overcoming the sample variability in long-term and large-scale pharmacometabonomics studies and identifying diagnostic and prognostic biomarkers [202].
	PQN	AlpsNMR (nmr_normalize)	Assisting the prediction of capecitabine-induced toxicity in patients with inoperable colorectal cancer using the processing serum metabolic profiles [95].
	QUA	affy (normalize.quantile)	Integrated in a metabolomic analysis tool to process a pharmacometabolomics dataset of antihypertensive medication [204].
	TSN	metabolomics (Normalise)	Helping to demonstrate the potential association of serum formate and acetate with varying responses to gemcitabine-carboplatin chemotherapy in patients with metastatic breast cancer [205].
Metabolite-based normalization	ATO	DiffCorr (scalingMethods)	Used together with the cube root transformation for processing the secondary whole blood pharmacometabolomics dataset [177].
	LEV	DiffCorr (scalingMethods)	Scaling the data and helping to differentiating liver metabolic profiles between sample groups in toxicology studies and clinical investigations of liver disease [212].
	PAR	DiffCorr (scalingMethods)	Reducing the weight of the large FCs in metabolite signals for pharmacometabonomic phenotyping of different responses to xenobiotic intervention in rats [119].
	POW	DiffCorr (scalingMethods)	Correcting for heteroscedasticity and pseudo scaling in MS-based serum metabolic profiling and investigating their alterations in colorectal cancer [213].
	RAN	DiffCorr (scalingMethods)	Scaling important indicators with different order of magnitude in a pharmacometabonomics study of predicting individual differences of cisplatin nephrotoxicity in rats [209].
	VAS	DiffCorr (scalingMethods)	Enhancing multivariate models used for disease classification and biomarker identification in unsupervised and supervised metabolomics analysis [210, 214].
Sample and metabolite-based normalization	VSN	vsn (vsn2)	Processing urine ^{1}H NMR spectra with factors such as diseases, drugs and toxins for metabolic profiling [219].
IS-based normalization	CCMN	metabolomics (Normalise)	Used in metabolomics and integrative omics for facilitating the development of Thai traditional medicine [224].
	NOMIS	metabolomics (Normalise)	Utilizing multiple ISs to remove overall unwanted experimental and biological variations in untargeted metabolomics data for Thai traditional medicine [224].
	RUV-random	MetNorm (NormalizeRUVRand)	Dealing with multivariate noise between samples in blood metabolomics data from maintenance hemodialysis patients with chronic kidney disease [225].
	RUV-2	ruv (RUV2)	Detecting and correcting for unwanted variation in drug metabolomics data [226].

Note: Abbreviation (abbr.) was assigned to each processing method and was described in section of the 'Data processing methods for pharmacometabonomics datasets'.

and response in non-diabetic volunteers by the non-targeted pharmacometabonomics approach [149].

The Median Imputation (MEI) method is similar to the MDI method, except, rather than the average value of non-missing values, it uses the median value to replace missing values. Median value is used in MDI to improve reliability since the mean value is easily affected by outliers. Both MDI and MEI method is frequently used because they are quite easier to implement compared with other methods. In a pharmacometabonomic assessment research on discovering metabolic phenotypes of atenolol and hydrochlorothiazide, the MEI method was applied to impute missing values in prior to hierarchical clustering [150].

The SVD method is also referred as principal-component analysis in statistics and Karhunen–Loève expansion in pattern recognition, respectively [151]. It is an imputation method that estimates the missing values based on a linear consideration [146]. Furthermore, the fundamental principle of this algorithm is estimating the missing values by regression against the principal components representing the whole data matrix information [152]. Both the SVD and KNN imputation methods outperform the Zero imputation [147], and SVD has been adopted to understand the pathogenesis of respiratory syncytial virus infection in infants and its association of childhood wheezing [153].

Zero Imputation (ZER) is easier to understand compared with the methods mentioned above, which simply replaces all missing values with zero. Zero imputation does not utilize any information from the dataset and may lead to biases such as altered distribution of missing variables and lower level of

the standard deviation [125]. This ZER has been successfully used to impute missing values in hepatocellular carcinoma metabolomics dataset for biomarker identification [154].

Method for QC-based correction of the pharmacometabonomics data

The QC samples refer to the pooled sample mixtures mixed by small and equal aliquots from the real samples of interest and dispersed evenly across the multiple batches to ensure the data quality for metabolic profiling [155, 156]. QC samples are frequently involved in large-scale metabolomics studies for signal drifts correction, intra- and inter-batch variations removal, where a great number of samples are impossible to be analyzed in a single run [157]. From this point of view, this data processing strategy based on the QCS for correcting for signal drifts and batch variations could be termed as 'QC-based normalization' [158, 159]. In this case, the available Quality Control–based Robust LOESS Signal Correction (QC-RLSC) method intergrated in the statTarget R package was proposed [160]. By specifying the parameter 'degree', the QC-RLSC method provides three distinct regression models (i.e. degree = 0/1/2 represent the Nadaraya–Watson estimator, locally linear regression and local polynomial regression fitting, respectively). The regression model of Nadaraya–Watson estimator is the classical one of QC-RLSC, which estimates the regression function by using the weighted average of the original data [161]. The locally linear regression is a non-parametric model where input-space looks linear if the function has sufficient smoothness [162]. And the local polynomial regression fitting model is also non-parametric for smoothing scatter plots and modeling functions [163]. This QC sample-based processing procedure has been applied in large-scale pharmacometabonomics study for signal correction and batch effects removal [164].

Methods for transforming the pharmacometabonomics data

The next step of the processing pipeline is the data transformation for reducing heteroscedasticity and correcting skewness [165–169]. The raw metabolomics matrix is typically heteroscedastic and right-skewed/positively skewed [137]. The heteroscedasticity is reflected as unequal variance of the data across samples, while the latter refers to the increase of variance accompanied with the value of the measurement. Given the assumption (the intensities of most metabolites are assumed to be unchanged considerably across samples) that most normalization methods hold for [137, 170], data are required to go through transformation in prior to data normalization. Here, four data transformation methods that are widely used in pharmacometabonomics studies were discussed.

The Box-Cox transformation, cube root transformation and square root transformation (SRT) are all transformation methods which utilize the power function for reducing heteroscedasticity and correcting skewness. Particularly, the corresponding transformation method is called as 'Box-Cox transformation', 'cube root transformation' or 'SRT' when the exponent of the power function equaled to [−5, 5], 3 and 2, respectively [126, 171, 172]. For instance, the 'SRT' calculates the square root of each element in the data matrix and replaces it with the original data.

In order to reduce heteroscedasticity of the data and to make the distribution more symmetrically previous to statistical analysis, the Log Transformation (LOG) converts multiplicative relations into addictive relations nonlinearly. The log transformation is often used due to its ability of removing heteroscedasticity based on the precondition that the RSD is constant. Two typical defects of the log transformation are that it is not applicable to value zero and it excessively emphasizes metabolites with relatively lower concentrations [173]. This method has been performed in a pharmacometabonomics study of describing urinary metabolomic alterations reflecting time-dependent alterations in response to intensive phase tuberculosis treatment, where the LOG method significantly corrected for the positively skewness of the metabolomics dataset [174].

The Cube Root Transformation (CUT) method employs the *n*th power transformation by replacing *n* with 1/3, which is based on the probability density function, the mean and the variance of the distribution [175]. The cube root transformation is often carried out to increase the weight of metabolites with relatively lower concentrations and compress the weight of metabolites with relatively higher concentrations to approximate a normal distribution [176]. With the help of the CUT method integrated in the MetaboAnalyst, Sun *et al.* [177] carried out a secondary whole blood pharmacometabonomics analysis and revealed the role of threonine, histidine and phenylalanine playing during the progression of peripheral neuropathy caused by paclitaxel.

As another means allowing parametric power transformation, the Box-Cox Transformation (BOX) enhances its performance with the ability of breaking away from multiple anomalies [178]. Nowadays, this method has been applied to correct for skewness of the serum metabolomics data and helped to find biomarkers of cholangiocarcinoma, hepatocellular carcinoma and primary sclerosing cholangitis for diagnosis [179].

The SRT is a method with the aim of transforming the distribution of metabolomics data to be more normal. This method is realized by calculating the square root of each element in the data matrix and replacing it with the original data [126, 172]. For the maternal serum metabolomic fingerprint-based study of diagnostic performance evaluation of a machine learning ensemble model, the SRT was adopted to transform the data and to correct for the skewness and heteroscedasticity [180].

Methods for normalizing the pharmacometabonomics data

Derived from the technical and/or biological errors during sample preparation and data acquisition, various forms of unwanted variations in the raw data matrix may bias the subsequent identification of significantly altered metabolites among different conditions [124, 181]. Therefore, data normalization strategy was proposed for eliminating the unwanted systematic variations while preserving the 'genuine' biological variability, which enhanced the reliability and interpretability of the subsequent statistical analysis [64, 65]. Recent literatures have categorized the normalization methods into four groups, including the sample-based, metabolite-based, sample and metabolite-based, and internal standard (IS)-based normalization [182]. And the first two categories could be distinguished by their aim for reducing systematic biases and making data more comparable among samples or metabolites. At present, a variety of normalization algorithms have been gradually developed and frequently used in pharmacometabonomics studies.

Sample-based normalization algorithms

With the aim of eliminating technical and/or biological variations across samples, a variety of sample-based normalization

methods have been developed and extensively utilized in the field of pharmacometabonomics, which include the Contrast (CON) [183, 184], Cubic Splines (CUB) [184, 185], Cyclic Loess (CYC) [184], EigenMS (EIG) [186–189], Linear Baseline Scaling (LIN) [184], Li-Wong (LIW) [183, 184], Mean Normalization (MEA) [186, 190], Median Normalization (MED) [191, 192], MS Total Useful Signal (MSTUS) [185, 193], Probabilistic Quotient Normalization (PQN) [194], Quantile (QUA) [184, 195] and Total Sum Normalization (TSN) [196].

Nowadays, the Contrast has been adopted in non-targeted metabolomics studies based on the LC/MS analytical technique to reduce the unwanted variations induced by biological or experimental factors [197]. The CUB method has been utilized to classify samples correctly regardless of the data set size and thus helped to identify clinically relevant biomarkers [198]. The CYC algorithm has been adopted in a high-throughput metabolomic study and assisted to observe significantly altered intracephalic metabolic features of the APP/PS1 model constructed for Alzheimer's disease [199]. This EigenMS has been used to process LC/MS-based metabolomics dataset for identifying and reducing systematic variations [187] and has also been utilized in the characterization of the maternal metabolome with gestational diabetes in China [200]. With its capacity of correcting for systematic variations, this LIN method has been conducted in a pharmacometabonomics study for predicting capecitabine-induced toxicity in patients with inoperable colorectal cancer [95]. LIW has been utilized to eliminate unwanted sample-to-sample bias in LC–MS based non-targeted metabolomics as far as possible [197]. Mean Normalization was applied to process the pharmacometabonomics data in prior to statistical analysis in a study of differentiating L-carnitine outcomes in patients for the treatment of septic shock [90]. Median Normalization has been adopted in a pharmacometabonomics analysis of human breath, which suggested the existence of highly individual phenotypes [201]. Given the superiority of improving the differentiation across sample groups and facilitating determination of statistically significant alterations of the urine samples, the MSTUS has been advised to serve as a normalization step in the analysis of urine samples [193]. Moreover, this method has been widely applied in pharmacometabonomics studies [202]. In the ^{1}H NMR-based metabolomics analysis, the Probabilistic Quotient Normalization has been discovered to perform as a robust algorithm for complicated biological mixtures containing various dilution concentration levels [203]. Furthermore, it has been applied to predict capecitabine-induced toxicity based on the baseline serum metabolic profiles [95]. The Quantile has been considered as another well-performing normalization method for the 1D ^{1}H metabolomics analysis of urinary samples [194] and has been integrated in an analysis tool with the capacity of dealing with pharmacometabonomics dataset [204]. So far, the TSN method has been extensively applied to act as a normalization strategy with the aim of removing unwanted variations among samples in pharmacometabonomics studies [205].

Metabolite-based normalization algorithms

Six metabolite-based normalization algorithms for minimizing the metabolite-to-metabolite variations and making data more comparable among metabolites were comprehensively reviewed, which included the Auto Scaling (ATO) [184, 188, 206, 207], Level Scaling (LEV) [173, 188], Pareto Scaling (PAR) [173, 184, 208], Power Transformation (POW) [173], Range Scaling (RAN) [173, 209] and Vast Scaling (VAS) [173, 210].

In current pharmacometabonomics, the Auto scaling has been adopted in MS-based analysis for facilitating the identification and diagnosis of patients with bladder cancers and assisting the feature selection in patients with urogenital cancers [211]. Moreover, it has been utilized to process the secondary whole blood pharmacometabolomics dataset [177]. Level scaling has been applied to process the (UPLC–MS) dataset in prior to PCA for classifying pre-designated classes of samples in toxicology studies and clinical investigations of liver disease [212].

And Pareto scaling has been utilized to perform as a normalization algorithm for eliminating the mask effects in current pharmacometabonomics analysis [119, 174]. The Power Transformation has been adopted to facilitate the identification of serum metabolic changes in patients with colorectal cancers [213]. Range scaling has been used to scale important indicators with different order of magnitude for improving value comparability in a pharmacometabonomics study [209]. Moreover, the Vast Scaling is reported to be a well-performing normalization method and has been widely utilized in the supervised or unsupervised metabolomics analysis, where the performance multivariate models for feature selection and sample classification were significantly improved [210, 214].

Both sample-based and metabolite-based normalization algorithm

Originally proposed to process single or two-channel microarray dataset [215], the Variance Stabilization Normalization (VSN) [184, 216] has gradually acted as a powerful normalization method in GC/MS-based metabolic analysis. With the unique capacity of correcting for the systematic bias among both samples and metabolites [217], the VSN method has been widely utilized in pharmacometabonomics for analyzing liver tissues during the progression of liver cancer [218] and has also adopted to process urine ^{1}H NMR spectra with factors such as diseases, drugs and toxins for metabolic profiling [219].

IS-based normalization algorithms

In addition to the sample and/or metabolite-based normalization algorithms discussed above, multiple popular normalization algorithms based on the IS have been extensively utilized in modern pharmacometabonomics studies. The ISs were ideally stable isotopically labelled compounds introduced during sample processing and could be easily distinguished from endogenous metabolites. The aim of the introduction was to correct for uncontrolled sample losses or compound degradation and subsequent sample losses, thus to improve method precision and accuracy [220]. Here, four IS-based normalization algorithms for removing overall unwanted experimental and biological variations were discussed, including the Cross-contribution Compensating Multiple standard Normalization (CCMN) [221], Normalization using Optimal selection of Multiple Internal Standards (NOMIS) [61], Remove Unwanted Variation-random (RUV-random) [196, 222] and 'Remove Unwanted Variation, 2-step' (RUV-2) [223].

In current pharmacometabonomics, the CCMN (based on a supervised statistical model for identifying and removing the overall unwanted variations) and the NOMIS (built on multiple ISs to normalize features and to eliminate unwanted variations) were reported to facilitate the development of Thai traditional medicine [224]. The RUV-random has been adopted to remove systematic noise in blood metabolomics data from maintenance hemodialysis patients with chronic kidney disease [225]. Moreover, the RUV-2 has been applied to detect and adjust for unwanted variation in drug metabolomics data [226].

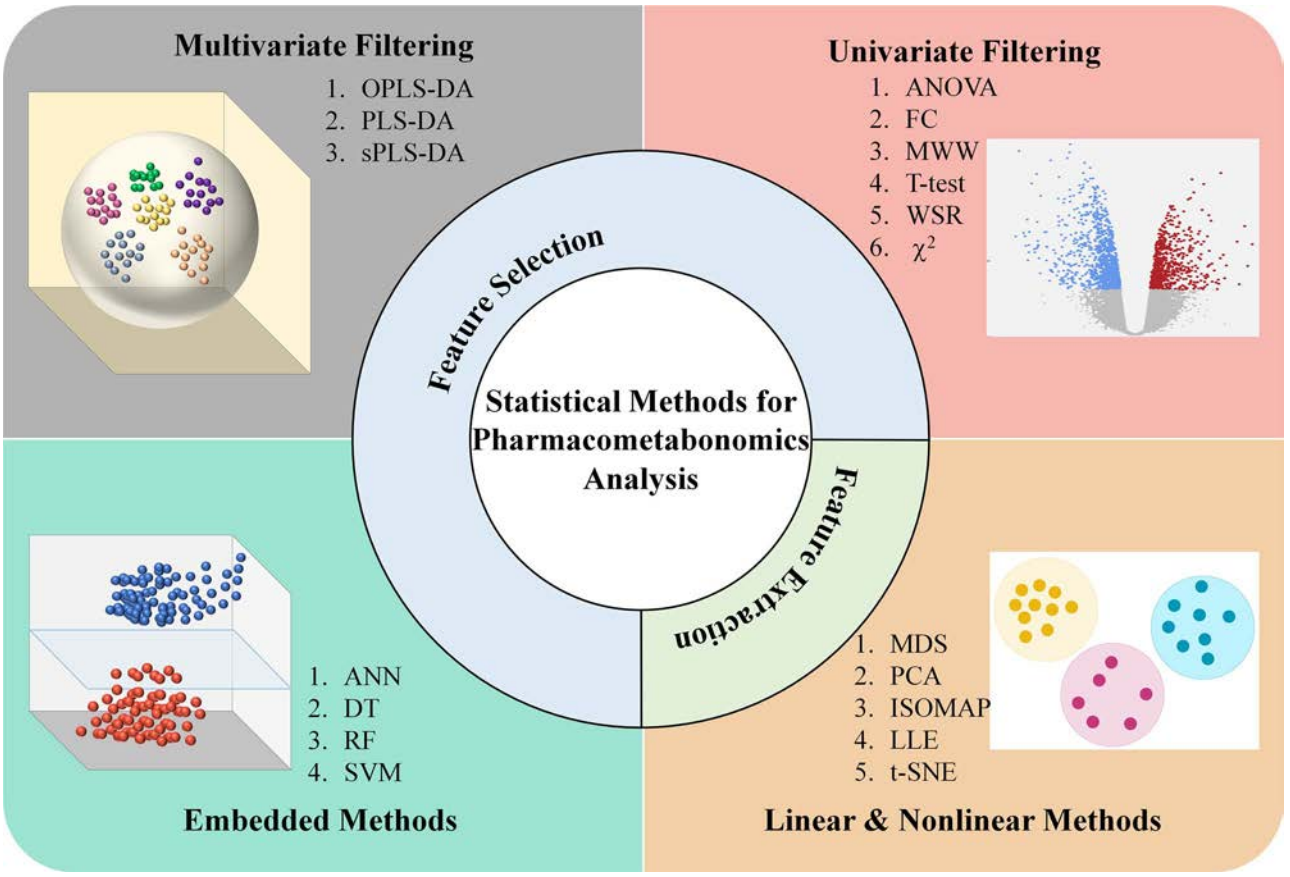

Figure 3. Statistical methods of feature selection and extraction strategies for dimensionality reduction in pharmacometabonomics analysis.

Statistical methods for biomarker identification and sample classification in pharmacometabonomics analysis

The feature selection and feature extraction strategies are greatly required for dimensionality reduction [67] and are very popular in the study of pharmacometabonomics for biomarker identification and sample classification. An appropriate feature selection method can facilitate the identification of optimal differential metabolic features and thus more accurately predict how patients will respond to certain medications. And the feature extraction methods can facilitate the classification of samples and help to identify the most significantly altered features [227]. The feature extraction strategies applied in pharmacometabonomics fall into two groups, including linear and nonlinear methods [228]. The classification of methods in the category of feature selection and feature extraction strategies used in pharmacometabonomics for dimensionality reduction was shown in Figure 3. And the descriptions of various commonly used feature selection and feature extraction methods for pharmacometabonomics studies were shown below and summarized in Table 3.

Univariate feature selection algorithms for biomarker identification in pharmacometabonomics

The univariate filtering strategy treats each feature individually and independently by evaluating and ranking each feature according to the certain criteria. Methods belonging to this category include ANOVA, FC, MWW, T-test, WSR and χ^2.

The Analysis of Variance (ANOVA) method uses linear statistical hypothesis testing with or without parameters. It focuses on comparing the dissimilitude between various groups variance or average with respect to a specific metabolite. ANOVA has been used for pharmacometabonomics study of patients treated with L-carnitine and placebo for discovering the effect of L-carnitine on different metabolic phenotypes [90].

The simple statistical method Fold Change (FC) is often utilized with many other parametric or non-parametric algorithms for assessing the change of absolute value between two sample groups. It calculates the original or the log value of the ratio and reports the result as significant when the FC value exceeds the predefined threshold. FC was used to discover biological predictors of the clinical response of cytosine arabinoside plus anthracycline treatment for acute myeloid leukemia [135].

Mann–Whitney–Wilcoxon test (MWW) is a non-parametric algorithm that sometimes referred as the Wilcoxon rank-sum test, which can be alternative to two-sample *t*-test for identifying the differences between two groups (unpaired samples) [229]. MWW are often used when the data do not meet the assumptions of the *t*-test. It is the null hypothesis that compares two means of the same sample and then tests whether these two means are equally distributed. The MWW was utilized to identify the biomarker for predicting alcohol-dependent treatment outcomes of acamprosate [117].

T-test is used to estimate whether there is a significant difference between the mean value of two datasets. It is one of the most prevalent tests used in the medical field and the most powerful unbiased test when the processed data fit a normal

Table 3. Summary of popular feature selection and feature extraction strategies and algorithms that are widely used in pharmacometabonomics studies

Algorithm (abbr.)		R package (function)	Applications in pharmacometabonomics study
Univariate filtering	ANOVA	ANOVA.TFNs (fanova)	Predicting different metabolic phenotypes of the L-carnitine [90].
	FC	metabolomics (FoldChange)	Discovering biological predictors of the clinical response of cytosine arabinoside plus anthracycline treatment for acute myeloid leukemia [135].
	MWW	stats (wilcox.test)	Identifying the biomarker for predicting alcohol-dependent treatment outcomes of acamprosate [117].
	T-test	stats (t.test)	Predicting capecitabine-induced toxicity in patients with inoperable colorectal cancer [95].
	WSR	stats (wilcox.test)	Comparing changes of salivary cortisol and IL-6 before and after decompression therapy in breast cancer survivors [231].
	χ^2	stats (chisq.test)	Identifying significant categorical clinical variables for predicting response of lisinopril in treating hypertension [233].
Multivariate filtering	OPLS-DA	ropls (opls)	Predicting metabolism characteristics of losartan in healthy volunteers [235].
	PLS-DA	caret (plsda)	Predicting therapeutic effects of trastuzumab-paclitaxel in HER-2 positive breast cancer patients [236].
	sPLS-DA	mixOmics (splsda)	Predicting response to disease modifying treatment in patients with multiple sclerosis [237].
Embedded feature selection	ANN	neuralnet (neuralnet)	Applied for the study of pharmacometabonomics with the aim of precise drug treatment for Alzheimer's disease [238].
	DT	dtree (pca)	Successfully predicting the development of IFNβ antibodies in patients with multiple sclerosis [239].
	RF	randomForest (randomFores)	Excavating the relationship between repeated meloxicam administration and the damage to kidneys in cats [240].
	SVM	e1071 (svm)	Being trained based on the pharmacometabonomics data to predict hepatotoxicity of six drugs including L-carnitine [241].
Linear feature extraction	PCA	ropls (pca)	Predicting metabolic phenotypes and pharmacokinetic parameters of atorvastatin in healthy volunteers [110].
	MDS	stats (cmdscale)	Assessing the effects of sample classification in breast cancer metabolic profiling by using Spearman's correlation as similarity measure [243].
Nonlinear feature extraction	ISOMAP	RDRToolbox (Isomap)	Assisting the discovery of underlying therapeutic effects and functional patterns of Radix Paeoniae Alba administration [245].
	LLE	RDRToolbox (LLE)	Helping to discover the underlying therapeutic effects and functional patterns of Radix Paeoniae Alba administration [245].
	t-SNE	Rtsne (Rtsne)	Identification of blood diagnostic biomarker of concussion in adolescent male hockey players based on metabolic profiling [248].

Notes: The specific realization of algorithms in R packages were listed. Abbreviation (abbr.) was assigned to each method and was described in section of the 'Statistical methods for biomarker identification and sample classification in pharmacometabonomics analysis'.

distribution. It was used to analyze the relationship between the toxicity after being exposed to capecitabine and the baseline metabolic profiles in a ^{1}H NMR pharmacometabonomics study [95].

The Wilcoxon signed-rank test (WSR) is a common non-parametric test for comparison of paired samples on the basis of independent units of analysis, which is a non-parametric substitute for the paired sample *t*-test [230]. It was used to measure twice for comparing changes of blood pressure in patients before and after drug exposure [229]. The WSR has also been applied to compare changes of salivary cortisol and interleukin-6 (IL-6) before and after decompression therapy in breast cancer survivors [231].

Chi-square (χ^2) is a popular non-parametric statistical test which can be used to assess the independence of two events. It relies on degrees of the sample size and the freedom, which makes its performance unreliable in very few cases [232]. Chi-square was applied to discover significantly altered categorical clinical features, and thus to predict the outcomes of lisinopril in patients with hypertension [233] and predicting SSRI therapeutic response in adults with major depressive disorder [234].

Multivariate feature selection algorithms for biomarker identification in pharmacometabonomics

Each metabolite does not act independently in the body, while they interact with each other and collectively influence the metabolomic phenotype, which makes the multivariate filtering strategy crucial for feature selection in the field of pharmacometabonomics.

OPLS-DA is a complicated multivariate algorithm that is applicable for analysis containing single or multiple groups especially for dataset with multi-collinear variables. It is a robust statistical method that is similar to the standard PLS-DA, which is capable of investigating as well as predicting qualitative structures of the dataset. This method has widely utilized in pharmacometabonomics studies for predicting metabolism characteristics of losartan in healthy volunteers [235].

As a most commonly used feature selection method, the PLS-DA is classified as a linear binary classifier. It is a supervised multivariate statistical approach that maximizes the interval between predefined classes. It has been applied to discover biomarkers that is significantly related with pathologic complete outcome of trastuzumab plus paclitaxel as neoadjuvant therapy among patients with HER-2-positive breast cancers [236].

Compared with the model of PLS-DA, the Sparse Partial Least Squares Discriminant Analysis (sPLS-DA) reduces more data by utilizing a lasso penalization combined with SVD. Therefore, it tends to neglect features that only differentiate between few samples. In this method, model construction and feature selection can be done at the same time and it uses valuable graphical output to modify interpretability. The sPLS-DA method has been used to predict the outcome of disease-modifying treatments in multiple sclerosis patients [237].

Embedded feature selection algorithms for biomarker identification in pharmacometabonomics

In addition to the feature selection strategy based on the filtering algorithms, another strategy commonly used for selecting the optimal feature subset in the field of pharmacometabonomics is the embedded strategy. The basic principles and applications in pharmacometabonomics of methods contained to this category were discussed below.

As a supervised embedding algorithm, the Artificial Neural Network (ANN) can mimic the structure and function of biological neural networks, and the basic unit of which is artificial neurons representative of mathematical functions. ANN processes the data by adjusting weight of large number neuron connections. It is suitable for investigating complicated non-linear association of dependent and independent features. ANN has been used in the study of pharmacometabonomics with the aim of precise drug treatment for Alzheimer's disease [238].

As a popular decision support tool with the structure like a tree, leaf nodes in the Decision Tree (DT) denote class labels, while the non-leaf nodes denote tests on characteristics, and branches denote the test result. Relying on the training dataset, the DT selects an attribute to split the given set of examples. It needs less pre-knowledge and can be verified by testing data. Moreover, the induction and classification steps are easy and quick. Coupled with other five algorithms, this method has been utilized to successfully predict the development of IFNβ antibodies in patients with multiple sclerosis [239].

Random forest (RF) is a popular supervised machine-learning algorithm that integrates multiple DTs. It is excellent in accuracy and has a better performance in classification tasks comparing with SVM, which makes it a common method in biomarker selection and clinical phenotypic discrimination. Together with PLS-DA, RF was used to excavate the relationship between repeated meloxicam administration and the damage to kidneys in cats [240].

As a supervised machine learning algorithm, the Support Vector Machine (SVM) allows classification and regression against the pharmacometabonomics dataset, with the aim of finding the segmentation surface that successfully classify data points into different categories. SVM can analyze complex large-scale and time-consuming data and perform well with very few samples with high dimensionality. An SVM model was trained based on the metabolite data to predict hepatotoxicity of six drugs including L-carnitine [241].

Linear feature extraction algorithms for sample classification in pharmacometabonomics

Linear feature extraction method supposes that the data are located on a lower dimensional linear subspace and it uses matrix decomposition strategy to project the original data on this subspace. Two methods belonging to this classification discussed here are PCA and MDS.

Relying on analyzing the linear association between different metabolites, the PCA transforms the dataset and then bases on the obtained variances to obtain the most important principal components, which explain the decreasing amount of the dataset variance. Along with the PLS regression, PCA was used to predict and differentiate various treatment metabotypes and pharmacokinetic properties of atorvastatin among healthy subjects [110].

Multidimensional Scaling (MDS) can project high-dimensional data into a low-dimensional space, which represents the similarity between preserved data points. The dimensionality reduction facilitates the discovery of hidden true structures in the data and reduces the complexity of information retrieval in large-scale datasets [242]. The MDS method using Spearman's correlation as similarity measure was utilized to evaluate the effects of sample classification in breast cancer metabolic profiling [243].

Nonlinear feature extraction algorithms for sample classification in pharmacometabonomics

Non-linear feature extraction method can lower the dimensionality of data by mapping features on a low-dimensional surface into a high-dimensional space through a lifting function, which allows to find non-linear relationships between features. Methods belonging to this category include ISOMAP, LLE and t-SNE.

Isometric Mapping (ISOMAP) analyzes and operationalizes the intrinsic nonlinear degrees of freedom of high-dimensional observations by finding meaningful low-dimensional structures in high-dimensional observations [244]. In the field of pharmacometabonomics, the ISOMAP could help to discover the underlying therapeutic effects and functional patterns of Radix Paeoniae Alba administration [245].

Locally Linear Embedding (LLE) is an unsupervised learning algorithm that projects high-dimensional data into a low-dimensional embedding space while conserving object adjacencies in the original high-dimensional feature space. Based on the local symmetry of linear reconstruction, LLE can study the overall structure of non-linear manifolds [246]. The LLE algorithm has also been used to discover the underlying therapeutic effects and functional patterns of Radix Paeoniae Alba administration [245].

Similarities between low-dimensional and high-dimensional data can be maintained during dimensionality reduction when using the t-Distributed Stochastic Neighbor embedding (t-SNE), which thus ensures the majority integrity of the structural information [247]. The t-SNE is a dimensionality reduction method with minimal loss of structural information and has been widely applied in pharmacometabonomics for metabolic profiling of concussion in adolescent male hockey players [248].

Metabolic databases for pharmacometabonomics research

Accurately identifying and determining a set of metabolites (or specific metabolites) of analyzed samples based on metabolic

databases are crucial issues involved in metabolomics, which is a prerequisite for subsequent biological interpretations [249, 250]. After the putative metabolites searching followed by the validation of them, metabolic pathway analysis based on metabolic databases is required, which attempts to discover the association of altered metabolites to corresponding metabolic pathways. Nowadays, numerous commercial or free databases have been constructed for pharmacometabonomics researches, which mainly fall into two categories according to the intended purpose (for metabolite identification and metabolic pathway analysis) [251]. Here, the summary of various widely used data repositories for pharmacometabonomics studies was provided in Table 4.

Databases for metabolite identification

Metabolite identification is only required for untargeted metabolomics studies, whereas the metabolite or metabolic class of interest in targeted metabolomics studies is already known [249]. A host of databases developed with the purpose of overcoming the challenge of illuminating the dark matter in metabolomics [252] have emerged recently, seven of which were described as follows.

As a free online resource containing comprehensive information of metabolites, interactional enzymes and transport proteins [253] as well as their chemical properties, biological roles, disease-related properties, metabolic pathways and reference spectrograms [75], The Human Metabolome Database (HMDB) is regarded as the greatest and most exhaustive metabolomics database all over the world. The latest version of HMDB contained pharmacometabonomics data involving the analysis of changes in metabolite levels in tissues, cells or biofluids after drug administration, which contributed to the achievement of precision drug delivery [254].

PubChem is the chemical information resource of the National Center for Biotechnology Information for many areas of biomedical research, such as the cheminformatics, chemical biology, medicinal chemistry and drug discovery [255]. PubChem consists of three interrelated databases of Substances, Compounds and Bioassays. The BioAssay database contains descriptions and test results of bioassay experiments [256, 257]. In a case of untargeted pharmacometabonomics study based on liquid chromatography coupled with electrospray ionization tandem mass spectrometry (LC-ESI-MS/MS), researchers matched the measured experimental features with compounds mined from PubChem and thus obtained candidate structures of unknown metabolites [258].

MassBank is the initial publicly available database of small molecular compounds in the life sciences [259]. It contains high-resolution MS information of metabolites and is an online spectral search tool and repository [260]. By specifying one or more experimental conditions, users can obtain access to the whole or part of the data in MassBank [259]. The application of pharmacometabonomics in translational and clinical research has been improved by using MassBank to identify high confidence metabolites, elucidate unknown metabolites, enable biological interpretation of complex systems, and build reliable predictive metabolic models [261].

Metabolite Link (METLIN) is an open-access, cloud-based metabolite database providing a comprehensive set of more than 1 million molecules, such as amino acids, small peptides, lipids, carbohydrates and natural products [262]. The database is capable of characterizing and identifying hundreds or thousands of naturally existing metabolites in analyzed samples, automating the identification of metabolites as well as overcoming the drawbacks of traditional analytical methods for more efficient identification of metabolites [77]. METLIN has been utilized for untargeted pharmacometabonomics analysis of plasma samples following psychoactive stimulants administration, which provided a basis for further deepening targeted metabolomic studies of pharmacological effects and finding biomarkers of drug use [263].

The LIPID Metabolites and Pathway Strategy (LIPID MAPS) is a freely available online database of lipid structural resources [264]. Lipids are important metabolites that affect the physiological and pathophysiological conditions of the human body. The LIPID MAPS plays an important global role in advancing technologies and resources [265]. The latest version of LIPID MAPS adds the software tool LipidFinder, which eliminates artifactual signals from processed data obtained from MS with high resolution. The combination of long time-course chromatographic analysis and high-resolution MS allows the identification of specific candidate lipids from large-scale lipidomic data. The application of LIPID MAPS in pharmacometabonomics studies facilitates the discovery of bioactive species that can be used as biomarkers for diseases or new therapeutic targets in the field of precision medicine [266].

The Chemical Entities of Biological Interest (ChEBI) is a manually annotated database of molecular entities with a focus on small molecule compounds, providing a wide range of data entries including chemical nomenclature, ontology and chemical structures [267]. The latest version of ChEBI extended our collection of endogenous metabolites from human, mouse, *Escherichia coli* and yeast, allowing for wide applications in a variety of scientific settings for different types of users [268]. Ontology-based enrichment analysis helps to interpret and understand large-scale biological data. Moreover, the BiNChE is a ChEBI ontology-based small molecule enrichment analysis tool that enables automated metabolite identification, facilitates understanding of pharmacometabonomics and promotes the development of metabolomics and systems biology [269].

Therapeutic Target Database (TTD) is an open and comprehensive database containing information of target modulators such as target-interacting proteins, target-regulated microRNAs and transcription factors, patent granted drugs and their targets [270]. By providing exhaustive information of the therapeutic effects of drugs and the underlying functional patterns in metabolism, TTD significantly facilitated the biological interpretations of pharmacometabolomics data [271]. It provided numerous disease targets for scanning diagnostic biomarkers in traditional Chinese medicine (TCM), which in turn could clarify the complexity metabolic effects on human bodies and pharmacological mechanisms of TCM [272].

As a global leading database containing chemical compounds mentioned in many literatures, the CAS Registry covers chemicals which could date back to more than 150 years ago. CAS also stores data on the framework of each registered substance that meets specific criteria [273]. All new molecular entities (NMEs) are included in CAS, allowing insights into the origin of NMEs by identifying compounds that are similar to them, which helps the implementation of structural methods for assessing the innovativeness of new drugs and advancing innovative drug discovery [274].

Databases for metabolic pathway analysis

Both targeted and untargeted metabolomics studies include the critical step of biological interpretation. Once the putative

Table 4. Summary of various available data repositories that are commonly utilized in pharmacometabonomics studies

Database (abbr.)		Access[a]	URL	Application in pharmocometabonomics studies
Metabolite identification	HMDB	d, w	www.hmdb.ca	Incorporating pharmacometabonomics data to analyze changes in metabolite levels in tissues, cells or biofluids after drug administration [254].
	PubChem	d, w	https://pubchem.ncbi.nlm.nih.gov	Matching the measured experimental features with compounds mined from PubChem, and thus obtaining candidate structures of unknown metabolites in an untargeted pharmacometabonomics study based on ESI-LC–MS/MS [258].
	MassBank	d, w	http://www.massbank.jp	Identifying high confidence metabolites, elucidating unknown metabolites, interpreting complex systems from a biological perspective, building reliable predictive metabolic models, and improving the application of pharmacometabonomics in clinical research [261].
	METLIN	w	http://metlin.scripps.edu/	Applied to untargeted pharmacometabonomics analysis of plasma samples after psychoactive stimulant administration to find biomarkers of drug use [263].
	LIPID MAPS	d, w	http://www.lipidmaps.org/	Addressing pharmacometabonomics by facilitating the discovery of bioactives that can be used as disease biomarkers and new therapeutic targets in the field of precision medicine [266].
	ChEBI	d, w	http://www.ebi.ac.uk/chebi	Automatically identifying metabolites, facilitating metabolomics and systems biology, and deepening understanding of pharmacometabonomics [269].
	TTD	d, w	http://db.idrblab.net/ttd	Helping to clarify the complexity metabolic effects on human bodies and pharmacological mechanisms of TCM by providing numerous disease targets for scanning diagnostic biomarkers [272].
	CAS Registry	c	https://www.cas.org/	Facilitating the implementation of a structural approach to assess the innovativeness of new drugs by identifying compounds that are similar to NMEs [274].
Metabolic pathway analysis	KEGG	d, w	http://www.genome.jp/kegg/	Providing an integrated view on biological mechanisms of breast cancer data based on BRCA-Pathway [281].
	Reactome	d, w	https://reactome.org	Providing bioinformatics tools in pharmacometabonomics studies and assisting in visualization, interpretation and analysis of metabolic pathways [285].
	WikiPathways	d, w	http://www.wikipathways.org	Providing biotransformation pathway maps that directly visualize expression changes associated with drug metabolism [288].
	MetaCyc	d, w	https://MetaCyc.org	Providing a solid basis for predicting metabolic pathways in other organisms and a reliable resource for researchers in metabolic engineering, drug discovery and many other disciplines [291].
	SMPDB	d, w	http://www.smpdb.ca	Used in conjunction with other tools, SMPDB could obtain information about miRNAs and their target genes, clarify relevant pathways and elucidate miRNA regulatory mechanisms [294].

Notes: Details regarding accessibility and application covered by these databases discussed in this paper were summarized. Abbreviation (abbr.) was assigned to each database and was described in section of the 'Metabolic databases for pharmacometabonomics research'.
[a]Accessibility with respect to each database, c, d and w represent commercial, downloadable and online access, respectively.

metabolites are listed and their identification is confirmed, the next step is to search for the corresponding metabolic pathways [249, 275, 276]. Several databases have been established for interpretating complicated interconnections between metabolic pathways [277].

As an exhaustive database covering biological interpretations of genome sequences as well as many other high-throughput data [278], the Kyoto Encyclopedia of Genes and Genomes (KEGG) consists of three generic databases of systematic, genomic and chemical information for giving molecular-level and higher level

functional information of genes and genomes [279, 280]. Breast cancer-associated-pathway (BRCA-Pathway) supplies the KEGG with capacity of exploring and visualizing the breast cancer data interactively. Moreover, it provides researchers with a comprehensive and novel view of discovering the mechanisms of breast cancer [281].

The Reactome Knowledgebase (Reactome) is a free, peer-reviewed and human-managed database of metabolic pathways [282], which provides the molecular details of cellular process. Reactome is a comprehensive version of the typical metabolic graph, providing an organized network of molecular conversions in a singular consistent data model [283]. The Reactome allows the discovery of functional relationships between data and the storage of biological processes [284]. In the context of pharmacometabonomics studies, Reactome seeks to offer bioinformatics tools for visualizing, interpreting as well as analyzing pathway information to advance physiological and biomedical research [285].

WikiPathways is a free collaborative repository containing biological pathway patterns for data visualization and analysis [286]. Recent works have focused on adding more metabolite information to the database and provided more detailed knowledge of interactions to improve the identification of metabolites and corresponding pathways, which make WikiPathways to be a powerful and popular database on metabolic pathways [287]. Although we have gained quite deepness of understanding of drug metabolizing enzymes, the number of available biotransformation pathway maps is limited and difficult for researchers to use for visualization of multi-omics data. WikiPathways was thus constructed to provide us with the ability of directly visualizing biotransformation pathway maps of expression changes associated with drug metabolism [288].

MetaCyc is an open-access integrated database of metabolic enzymes and pathways, with the majority of MetaCyc pathways being experimentally identified small molecule metabolic pathways [289]. It contains 2749 pathways from more than 60 000 literatures and is the most comprehensive set of metabolic pathways [290]. The convergence of genome sequencing and bioinformatics has produced many metabolic databases for describing known and predicted metabolism in a variety of organisms, with a focus on the MetaCyc family of metabolic databases containing experimentally elucidated pathways. Therefore, the MetaCyc provides users with both reliable basis for the prediction of metabolic pathways and crucial resource in the fields of toxicology, metabolic profiling, drug discovery [291].

The Small Molecule Pathway Database (SMPDB) is an exhaustive, highly interactive and entirely searchable database with the capacity of visualizing metabolic pathways of signals, diseases as well as drugs [292]. Containing more than 600 pathways, it provides exhaustive metabolic information within the human body, such as information about tissues, organelles and subcellular compartments. Transporter proteins and much biological information about target organs, tissues and reaction compartments are involved in the latest version of SMPDB [293]. Used in conjunction with other tools, SMPDB could obtain information about miRNAs and their target genes, clarify relevant pathways and elucidate miRNA regulatory mechanisms [294].

Conclusions

As the research and development (R&D) of novel drugs require huge investments of time and funds, it is non-trivial for the pharmaceutical industry to enhance the speed and success rate of this process by using advanced technologies. Given the potential of 'omics' showed in the new drug R&D, such as the discovery of promising targets and the prediction of drug effets, pharmacometabonomics which focused on pharmacokinetic and pharmacodynamic properties, was thus proposed and developed in recent years. It is worth noting that the exsitence of various strategies and methods for processing pharmacometabonomics data posed challenges for proper selection and application on the specific dataset analyzed [295]. Therefore, this review provided guidance for researchers engaged in pharmacometabonomics and metabolomics, and it would promote the wide application of metabolomics in drug research and personalized medicine.

Key Points

- Recent progress on pharmacometabonomics shed lights on the individualized treatment and precision medicine from the sight of PK and PD.
- Rapid development of analytical techniques reinforced the urgent need to review their strengths and weaknesses, as well as corresponding applications in pharmacometabonomics.
- As a crucial prerequisite for complicated analysis of pharmacometabonomics data, appropriate utilization of data processing methods with different underlying theories requires extensive considerations for distinct dataset analyzed.
- The emergence of various novel statistical analysis strategies and algorithms as well as numerous metabolomic-related databases posed challenges for proper selection and application of them with respect to specific high-throughput pharmacometabonomics data.

Author Contributions

Author contributions: J.F., Y.Z.: conceptualized, searched and reviewed literature, created the tables and drafted the manuscript. J.L., X.L. and J.T.: searched literature. F.Z.: conceptualized and critically reviewed the paper.

Funding

National Natural Science Foundation of China (81872798, U1909208); Natural Science Foundation of Zhejiang Province (LR21H300001); National Key R&D Program of China (2018YFC0910500); Leading Talent of the 'Ten Thousand Plan'—National High-Level Talents Special Support Plan of China; Fundamental Research Fund for Central Universities (2018QNA7023); Key R&D Program of Zhejiang Province (2020C03010). Supported by Information Technology Center, Zhejiang University.

References

1. Otte C, Gold SM, Penninx BW, *et al*. Major depressive disorder. *Nat Rev Dis Primers* 2016;**2**:16065.
2. Dean J, Keshavan M. The neurobiology of depression: an integrated view. *Asian J Psychiatr* 2017;**27**:101–11.
3. Srivastava N, Mishra BN, Srivastava P. In-silico identification of drug lead molecule against pesticide

exposed-neurodevelopmental disorders through network-based computational model approach. *Curr Bioinform* 2019; **14**:460–7.
4. Libby P, Buring JE, Badimon L, *et al*. Atherosclerosis. *Nat Rev Dis Primers* 2019;**5**:56.
5. Forbes JM, Cooper ME. Mechanisms of diabetic complications. *Physiol Rev* 2013;**93**:137–88.
6. Fu TT, Zheng GX, Tu G, *et al*. Exploring the binding mechanism of metabotropic glutamate receptor 5 negative allosteric modulators in clinical trials by molecular dynamics simulations. *ACS Chem Neurosci* 2018;**9**:1492–502.
7. Jiang P, Du W, Wu M. Regulation of the pentose phosphate pathway in cancer. *Protein Cell* 2014;**5**:592–602.
8. Li FC, Zhou Y, Zhang XY, *et al*. SSizer: determining the sample sufficiency for comparative biological study. *J Mol Biol* 2020;**432**:3411–21.
9. Tang W, Wan S, Yang Z, *et al*. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* 2018;**34**:398–406.
10. Ji J, Tang J, Xia K, *et al*. LncRNA in tumorigenesis microenvironment. *Curr Bioinform* 2019;**14**:640–1.
11. Nadia RJ. The human oncobiome database: a database of cancer microbiome datasets. *Curr Bioinform* 2020;**15**:472–7.
12. Wang M, Kuchiba A, Ogino S. A meta-regression method for studying etiological heterogeneity across disease subtypes classified by multiple biomarkers. *Am J Epidemiol* 2015;**182**:263–70.
13. Fereshtehnejad SM, Postuma RB. Subtypes of parkinson's disease: what do they tell us about disease progression? *Curr Neurol Neurosci Rep* 2017;**17**:34.
14. Zhao X, Chen L, Guo Z, *et al*. Predicting drug side effects with compact integration of heterogeneous networks. *Curr Bioinform* 2019;**14**:709–20.
15. Li B, He X, Jia W, *et al*. Novel applications of metabolomics in personalized medicine: a mini-review. *Molecules* 2017;**22**:1173.
16. Doestzada M, Vila AV, Zhernakova A, *et al*. Pharmacomicrobiomics: a novel route towards personalized medicine? *Protein Cell* 2018;**9**:432–45.
17. Su R, Liu X, Wei L, *et al*. Deep-Resp-Forest: a deep forest model to predict anti-cancer drug response. *Methods* 2019;**166**:91–102.
18. Yang QX, Li B, Chen SJ, *et al*. MMEASE: online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis. *J Proteomics* 2021;**104023**:232.
19. Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat Rev Mol Cell Biol* 2016;**17**:451–9.
20. Lee D, Choi YH, Seo J, *et al*. Discovery of new epigenomics-based biomarkers and the early diagnosis of neurodegenerative diseases. *Ageing Res Rev* 2020;**61**:101069.
21. Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat Rev Drug Discov* 2016;**15**:473–84.
22. Latini A, Borgiani P, Novelli G, *et al*. miRNAs in drug response variability: potential utility as biomarkers for personalized medicine. *Pharmacogenomics* 2019;**20**:1049–59.
23. Mayer B, Heinzel A, Lukas A, *et al*. Predictive biomarkers for linking disease pathology and drug effect. *Curr Pharm Des* 2017;**23**:29–54.
24. Ru X, Wang L, Li L, *et al*. Exploration of the correlation between GPCRs and drugs based on a learning to rank algorithm. *Comput Biol Med* 2020;**119**:103660.
25. Fu TT, Tu G, Ping M, *et al*. Subtype-selective mechanisms of negative allosteric modulators binding to group I metabotropic glutamate receptors. *Acta Pharmacol Sin* 2020;**0**:1–14.
26. McColl ER, Asthana R, Paine MF, *et al*. The age of omics-driven precision medicine. *Clin Pharmacol Ther* 2019;**106**:477–81.
27. Irshad O, Khan MUG. Integration and querying of heterogeneous omics semantic annotations for biomedical and biomolecular knowledge discovery. *Curr Bioinform* 2020;**15**:41–58.
28. Hong JJ, Luo YC, Mou MJ, *et al*. Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery. *Brief Bioinform* 2020;**21**:1825–36.
29. Evans DA, Clarke CA. Pharmacogenetics. *Br Med Bull* 1961;**17**:234–40.
30. Wang L. Pharmacogenomics: a systems approach. *Wiley Interdiscip Rev Syst Biol Med* 2010;**2**:3–22.
31. Han Z, Xue W, Tao L, *et al*. Identification of novel immune-relevant drug target genes for alzheimer's disease by combining ontology inference with network analysis. *CNS Neurosci Ther* 2018;**24**:1253–63.
32. Clayton TA, Lindon JC, Cloarec O, *et al*. Pharmacometabonomic phenotyping and personalized drug treatment. *Nature* 2006;**440**:1073–7.
33. Schmidt CW. Metabolomics: what's happening downstream of DNA. *Environ Health Perspect* 2004;**112**:A410–5.
34. Kaddurah-Daouk R, Kristal BS, Weinshilboum RM. Metabolomics: a global biochemical approach to drug response and disease. *Annu Rev Pharmacol Toxicol* 2008;**48**: 653–83.
35. Kaddurah-Daouk R, Weinshilboum RM. Pharmacometabolomics: implications for clinical pharmacology and systems pharmacology. *Clin Pharmacol Ther* 2014;**95**: 154–67.
36. Kaddurah-Daouk R, Baillie RA, Zhu H, *et al*. Lipidomic analysis of variation in response to simvastatin in the cholesterol and pharmacogenetics study. *Metabolomics* 2010;**6**:191–201.
37. Kaddurah-Daouk R, Baillie RA, Zhu H, *et al*. Enteric microbiome metabolites correlate with response to simvastatin treatment. *PLoS One* 2011;**6**:e25482.
38. Trupp M, Zhu H, Wikoff WR, *et al*. Metabolomics reveals amino acids contribute to variation in response to simvastatin treatment. *PLoS One* 2012;**7**:e38386.
39. Li YH, Li XX, Hong JJ, *et al*. Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs. *Brief Bioinform* 2020;**21**:649–62.
40. Nandal S, Burt T. Integrating pharmacoproteomics into early-phase clinical development: state-of-the-art, challenges, and recommendations. *Int J Mol Sci* 2017;**18**: 448.
41. Xia J, Wishart DS. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nat Protoc* 2011;**6**:743–60.
42. Tugizimana F, Steenkamp PA, Piater LA, *et al*. A conversation on data mining strategies in LC–MS untargeted metabolomics: pre-processing and pre-treatment steps. *Metabolites* 2016;**6**:40.
43. Yang QX, Wang YX, Li FC, *et al*. Identification of the gene signature reflecting schizophrenia's etiology by constructing artificial intelligence-based method of enhanced reproducibility. *CNS Neurosci Ther* 2019;**25**:1054–63.

44. Wanichthanarak K, Jeamsripong S, Pornputtapong N, *et al*. Accounting for biological variation with linear mixed-effects modelling improves the quality of clinical metabolomics data. *Comput Struct Biotechnol J* 2019;**17**:611–8.
45. Dunn WB, Broadhurst D, Begley P, *et al*. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat Protoc* 2011;**6**:1060–83.
46. Giacomoni F, Le Corguillé G, Monsoor M, *et al*. Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. *Bioinformatics* 2015;**31**:1493–5.
47. Martínez-Arranz I, Mayo R, Pérez-Cormenzana M, *et al*. Enhancing metabolomics research through data mining. *J Proteomics* 2015;**127**:275–88.
48. Lu W, Su X, Klein MS, *et al*. Metabolite measurement: pitfalls to avoid and practices to follow. *Annu Rev Biochem* 2017;**86**:277–304.
49. Yin JY, Sun W, Li FC, *et al*. VARIDT 1.0: variability of drug transporter database. *Nucleic Acids Res* 2020;**48**:D1042–50.
50. Idle JR, Gonzalez FJ. Metabolomics. *Cell Metab* 2007;**6**:348–51.
51. Bawadikji AA, Teh CH, Sheikh Abdul Kader MAB, *et al*. Plasma metabolites as predictors of warfarin outcome in atrial fibrillation. *Am J Cardiovasc Drugs* 2020;**20**:169–77.
52. Liu D, An ZL, Li PF, *et al*. A targeted neurotransmitter quantification and nontargeted metabolic profiling method for pharmacometabolomics analysis of olanzapine by using UPLC-HRMS. *RSC Adv* 2020;**10**:18305–14.
53. Pedersen HK, Forslund SK, Gudmundsdottir V, *et al*. A computational framework to integrate high-throughput '-omics' datasets for the identification of potential mechanistic links. *Nat Protoc* 2018;**13**:2781–800.
54. Mirza B, Wang W, Wang J, *et al*. Machine learning and integrative analysis of biomedical big data. *Gen* 2019;**10**:87.
55. Tang J, Wang YX, Fu JB, *et al*. A critical assessment of the feature selection methods used for biomarker discovery in current metaproteomics studies. *Brief Bioinform* 2020;**21**:1378–90.
56. Hoffmann N, Rein J, Sachsenberg T, *et al*. mzTab-M: a data standard for sharing quantitative results in mass spectrometry metabolomics. *Anal Chem* 2019;**91**:3302–10.
57. Schiffman C, Petrick L, Perttula K, *et al*. Filtering procedures for untargeted LC–MS metabolomics data. *BMC Bioinformatics* 2019;**20**:334.
58. Taylor SL, Ruhaak LR, Kelly K, *et al*. Effects of imputation on correlation: implications for analysis of mass spectrometry data from multiple biological matrices. *Brief Bioinform* 2017;**18**:312–20.
59. Han ZJ, Xue WW, Tao L, *et al*. Genome-wide identification and analysis of the eQTL lncRNAs in multiple sclerosis based on RNA-seq data. *Brief Bioinform* 2020;**21**:1023–37.
60. Jauhiainen A, Madhu B, Narita M, *et al*. Normalization of metabolomics data with applications to correlation maps. *Bioinformatics* 2014;**30**:2155–61.
61. Yang QX, Li B, Tang J, *et al*. Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data. *Brief Bioinform* 2020;**21**:1058–68.
62. Huan T, Forsberg EM, Rinehart D, *et al*. Systems biology guided by XCMS online metabolomics. *Nat Methods* 2017;**14**:461–2.
63. Katajamaa M, Miettinen J, Oresic M. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* 2006;**22**:634–6.
64. Li B, Tang J, Yang Q, *et al*. NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Res* 2017;**45**:W162–70.
65. Yang Q, Wang Y, Zhang Y, *et al*. NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data. *Nucleic Acids Res* 2020;**48**:W436–48.
66. Chong J, Soufan O, Li C, *et al*. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res* 2018;**46**:W486–94.
67. Tang J, Wang Y, Luo Y, *et al*. Computational advances of tumor marker selection and sample classification in cancer proteomics. *Comput Struct Biotechnol J* 2020;**18**:2012–25.
68. Ji Y, Hebbring S, Zhu H, *et al*. Glycine and a glycine dehydrogenase (GLDC) SNP as citalopram/escitalopram response biomarkers in depression: pharmacometabolomics-informed pharmacogenomics. *Clin Pharmacol Ther* 2011;**89**:97–104.
69. Laaksonen R, Katajamaa M, Päivä H, *et al*. A systems biology strategy reveals biological pathways and plasma biomarker candidates for potentially toxic statin-induced changes in muscle. *PLoS One* 2006;**1**:e97.
70. Kaddurah-Daouk R, Boyle SH, Matson W, *et al*. Pretreatment metabotype as a predictor of response to sertraline or placebo in depressed outpatients: a proof of concept. *Transl Psychiatry* 2011;**1**:e26.
71. Gromski PS, Muhamadali H, Ellis DI, *et al*. A tutorial review: metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding. *Anal Chim Acta* 2015;**879**:10–23.
72. Gromski PS, Xu Y, Correa E, *et al*. A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data. *Anal Chim Acta* 2014;**829**:1–8.
73. Mendez KM, Reinke SN. Broadhurst DI. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics* 2019;**15**:150.
74. Trainor PJ, DeFilippis AP, Rai SN. Evaluation of classifier performance for multiclass phenotype discrimination in untargeted metabolomics. *Metabolites* 2017;**7**:30.
75. Wishart DS, Feunang YD, Marcu A, *et al*. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* 2018;**46**:D608–17.
76. Kanehisa M, Goto S, Hattori M, *et al*. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006;**34**(D1):D354–7.
77. Tautenhahn R, Cho K, Uritboonthai W, *et al*. An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat Biotechnol* 2012;**30**:826–8.
78. Everett JR, Holmes E, Veselkov KA, *et al*. A unified conceptual framework for metabolic phenotyping in diagnosis and prognosis. *Trends Pharmacol Sci* 2019;**40**:763–73.
79. Du Preez I, Loots DT. Novel insights into the pharmacometabonomics of first-line tuberculosis drugs relating to metabolism, mechanism of action and drug-resistance. *Drug Metab Rev* 2018;**50**:466–81.
80. Everett JR. NMR-based pharmacometabonomics: a new paradigm for personalised or precision medicine. *Prog Nucl Magn Reson Spectrosc* 2017;**102–103**:1–14.
81. Lindon JC, Holmes E, Nicholson JK. Metabonomics techniques and applications to pharmaceutical research & development. *Pharm Res* 2006;**23**:1075–88.
82. Xue WW, Yang FY, Wang PP, *et al*. What contributes to serotonin-norepinephrine reuptake inhibitors'

dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation. *ACS Chem Nerosci* 2018;**9**:1128–40.
83. Bao X, Wu J, Kim S, *et al*. Pharmacometabolomics reveals irinotecan mechanism of action in cancer patients. *J Clin Pharmacol* 2019;**59**:20–34.
84. Lewis MR, Pearce JT, Spagou K, *et al*. Development and application of ultra-performance liquid chromatography-TOF MS for precision large scale urinary metabolic phenotyping. *Anal Chem* 2016;**88**:9004–13.
85. Dona AC, Jiménez B, Schäfer H, *et al*. Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Anal Chem* 2014;**86**:9887–94.
86. Bales JR, Bell JD, Nicholson JK, *et al*. 1H NMR studies of urine during fasting: excretion of ketone bodies and acetylcarnitine. *Magn Reson Med* 1986;**3**:849–56.
87. Marion D. An introduction to biological NMR spectroscopy. *Mol Cell Proteomics* 2013;**12**:3006–25.
88. Clayton TA, Baker D, Lindon JC, *et al*. Pharmacometabonomic identification of a significant host-microbiome metabolic interaction affecting human drug metabolism. *Proc Natl Acad Sci USA* 2009;**106**:14728–33.
89. Kapoor SR, Filer A, Fitzpatrick MA, *et al*. Metabolic profiling predicts response to anti-tumor necrosis factor α therapy in patients with rheumatoid arthritis. *Arthritis Rheum* 2013;**65**:1448–56.
90. Puskarich MA, Finkel MA, Karnovsky A, *et al*. Pharmacometabolomics of L-carnitine treatment response phenotypes in patients with septic shock. *Ann Am Thorac Soc* 2015;**12**:46–56.
91. Hao D, Sarfaraz MO, Farshidfar F, *et al*. Temporal characterization of serum metabolite signatures in lung cancer patients undergoing treatment. *Metabolomics* 2016; **12**:58.
92. Keun HC, Sidhu J, Pchejetski D, *et al*. Serum molecular signatures of weight change during early breast cancer chemotherapy. *Clin Cancer Res* 2009;**15**:6716–23.
93. Hong JJ, Luo YC, Zhang Y, *et al*. Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning. *Brief Bioinform* 2020;**21**:1437–47.
94. Winnike JH, Li Z, Wright FA, *et al*. Use of pharmacometabonomics for early prediction of acetaminophen-induced hepatotoxicity in humans. *Clin Pharmacol Ther* 2010; **88**:45–51.
95. Backshall A, Sharma R, Clarke SJ, *et al*. Pharmacometabonomic profiling as a predictor of toxicity in patients with inoperable colorectal cancer treated with capecitabine. *Clin Cancer Res* 2011;**17**:3019–28.
96. Wang YX, Li FC, Zhang Y, *et al*. Databases for the targeted COVID-19 therapeutics. *Br J Pharmacol* 2020;**177**: 4999–5001.
97. Cunningham K, Claus SP, Lindon JC, *et al*. Pharmacometabonomic characterization of xenobiotic and endogenous metabolic phenotypes that account for inter-individual variation in isoniazid-induced toxicological response. *J Proteome Res* 2012;**11**:4630–42.
98. Coen M, Goldfain-Blanc F, Rolland-Valognes G, *et al*. Pharmacometabonomic investigation of dynamic metabolic phenotypes associated with variability in response to galactosamine hepatotoxicity. *J Proteome Res* 2012;**11**: 2427–40.
99. Ho CS, Lam CW, Chan MH, *et al*. Electrospray ionisation mass spectrometry: principles and clinical applications. *Clin Biochem Rev* 2003;**24**:3–12.
100. Baumann A, Karst U. Online electrochemistry/mass spectrometry in drug metabolism studies: principles and applications. *Expert Opin Drug Metab Toxicol* 2010;**6**:715–31.
101. Pitt JJ. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *Clin Biochem Rev* 2009;**30**:19–34.
102. Khamis MM, Adamko DJ, El-Aneed A. Mass spectrometric based approaches in urine metabolomics and biomarker discovery. *Mass Spectrom Rev* 2017;**36**:115–34.
103. Xue WW, Wang PP, Tu G, *et al*. Computational identification of the binding mechanism of a triple reuptake inhibitor amitifadine for the treatment of major depressive disorder. *Phys Chem Chem Phys* 2018;**20**:6606–16.
104. Beale DJ, Pinu FR, Kouremenos KA, *et al*. Review of recent developments in GC-MS approaches to metabolomics-based research. *Metabolomics* 2018;**14**:152.
105. Beccaria M, Franchina FA, Nasir M, *et al*. Investigation of mycobacteria fatty acid profile using different ionization energies in GC-MS. *Anal Bioanal Chem* 2018;**410**:7987–96.
106. Umebachi R, Saito T, Aoki H, *et al*. How does chirality determine the selective inhibition of histone deacetylase 6? A lesson from Trichostatin A enantiomers based on molecular dynamics. *ACS Chem Nerosci* 2019;**10**:2467–80.
107. Toyo'oka T. LC–MS determination of bioactive molecules based upon stable isotope-coded derivatization method. *J Pharm Biomed Anal* 2012;**69**:174–84.
108. Wang CH, Su H, Chou JH, *et al*. Solid phase microextraction combined with thermal-desorption electrospray ionization mass spectrometry for high-throughput pharmacokinetics assays. *Anal Chim Acta* 2018;**1021**:60–8.
109. Phapale PB, Kim SD, Lee HW, *et al*. An integrative approach for identifying a metabolic phenotype predictive of individualized pharmacokinetics of tacrolimus. *Clin Pharmacol Ther* 2010;**87**:426–36.
110. Huang Q, Aa J, Jia H, *et al*. A pharmacometabonomic approach to predicting metabolic phenotypes and pharmacokinetic parameters of atorvastatin in healthy volunteers. *J Proteome Res* 2015;**14**:3970–81.
111. Zhang S, Zhou Y, Wang YN, *et al*. The mechanistic, diagnostic and therapeutic novel nucleic acids for hepatocellular carcinoma emerging in past score years. *Brief Bioinform* 2021;**22**:1860–83.
112. Liu L, Cao B, Aa J, *et al*. Prediction of the pharmacokinetic parameters of triptolide in rats based on endogenous molecules in pre-dose baseline serum. *PLoS One* 2012;**7**:e43389.
113. Shin KH, Choi MH, Lim KS, *et al*. Evaluation of endogenous metabolic markers of hepatic CYP3A activity using metabolic profiling and midazolam clearance. *Clin Pharmacol Ther* 2013;**94**:601–9.
114. Lewis JP, Yerges-Armstrong LM, Ellero-Simatos S, *et al*. Integration of pharmacometabolomic and pharmacogenomic approaches reveals novel insights into antiplatelet therapy. *Clin Pharmacol Ther* 2013;**94**:570–3.
115. Ellero-Simatos S, Lewis JP, Georgiades A, *et al*. Pharmacometabolomics reveals that serotonin is implicated in aspirin response variability. *CPT Pharmacometrics Syst Pharmacol* 2014;**3**:e125.
116. Karas-Kuželički N, Šmid A, Tamm R, *et al*. From pharmacogenetics to pharmacometabolomics: SAM modulates TPMT activity. *Pharmacogenomics* 2014;**15**:1437–49.

117. Nam HW, Karpyak VM, Hinton DJ, *et al*. Elevated baseline serum glutamate as a pharmacometabolomic biomarker for acamprosate treatment outcome in alcohol-dependent subjects. *Transl Psychiatry* 2015;**5**:e621.
118. Weng L, Gong Y, Culver J, *et al*. Presence of arachidonoylcarnitine is associated with adverse cardiometabolic responses in hypertensive patients treated with atenolol. *Metabolomics* 2016;**12**:160.
119. Li H, Ni Y, Su M, *et al*. Pharmacometabonomic phenotyping reveals different responses to xenobiotic intervention in rats. *J Proteome Res* 2007;**6**:1364–70.
120. Navarro SL, Randolph TW, Shireman LM, *et al*. Pharmacometabonomic prediction of busulfan clearance in hematopoetic cell transplant recipients. *J Proteome Res* 2016;**15**:2802–11.
121. Muhrez K, Benz-de Bretagne I, Nadal-Desbarats L, *et al*. Endogenous metabolites that are substrates of organic anion transporter's (OATs) predict methotrexate clearance. *Pharmacol Res* 2017;**118**:121–32.
122. Dai D, Tian Y, Guo HM, *et al*. A pharmacometabonomic approach using predose serum metabolite profiles reveals differences in lipid metabolism in survival and non-survival rats treated with lipopolysaccharide. *Metabolomics* 2016;**12**:2–14.
123. Zhang P, Li W, Chen J, *et al*. Branched-chain amino acids as predictors for individual differences of cisplatin nephrotoxicity in rats: a pharmacometabonomics study. *J Proteome Res* 2017;**16**:1753–62.
124. Xia J, Mandal R, Sinelnikov IV, *et al*. MetaboAnalyst 2.0–a comprehensive server for metabolomic data analysis. *Nucleic Acids Res* 2012;**40**:W127–33.
125. Tang J, Mou MJ, Wang YX, *et al*. MetaFS: performance assessment of biomarker discovery in metaproteomics. *Brief Bioinform* 2020;**00**:1–11.
126. Karaman I. Preprocessing and pretreatment of metabolomics data for statistical analysis. *Adv Exp Med Biol* 2017;**965**:145–61.
127. Wu Y, Li L. Sample normalization methods in quantitative metabolomics. *J Chromatogr A* 2016;**1430**:80–95.
128. Taverna F, Goveia J, Karakach TK, *et al*. BIOMEX: an interactive workflow for (single cell) omics data interpretation and visualization. *Nucleic Acids Res* 2020;**48**:W385–94.
129. Southam AD, Weber RJ, Engel J, *et al*. A complete workflow for high-resolution spectral-stitching nanoelectrospray direct-infusion mass-spectrometry-based metabolomics and lipidomics. *Nat Protoc* 2016;**12**:310–28.
130. Hagenbeek FA, Pool R, van Dongen J, *et al*. Heritability estimates for 361 blood metabolites across 40 genome-wide association studies. *Nat Commun* 2020;**11**:39.
131. Chen MX, Wang SY, Kuo CH, *et al*. Metabolome analysis for investigating host-gut microbiota interactions. *J Formos Med Assoc* 2019;**118**:S10–22.
132. Li CY, Song HT, Wang XX, *et al*. Urinary metabolomics reveals the therapeutic effect of HuangQi injections in cisplatin-induced nephrotoxic rats. *Sci Rep* 2017;**7**:3619.
133. Oakes JM, Scadeng M, Breen EC, *et al*. Rat airway morphometry measured from in situ MRI-based geometric models. *J Appl Physiol (1985)* 2012;**112**:1921–31.
134. Dai W, Wei C, Kong H, *et al*. 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* 2011;**56**:86–92.
135. Tan G, Zhao B, Li Y, *et al*. Pharmacometabolomics identifies dodecanamide and leukotriene B4 dimethylamide as a predictor of chemosensitivity for patients with acute myeloid leukemia treated with cytarabine and anthracycline. *Oncotarget* 2017;**8**:88697–707.
136. Wei R, Wang J, Jia E, *et al*. GSimp: a Gibbs sampler based left-censored missing value imputation approach for metabolomics studies. *PLoS Comput Biol* 2018;**14**:e1005973.
137. Reinhold D, Pielke-Lombardo H, Jacobson S, *et al*. Pre-analytic considerations for mass spectrometry-based untargeted metabolomics data. *Methods Mol Biol* 2019;**1978**:323–40.
138. Kokla M, Virtanen J, Kolehmainen M, *et al*. Random forest-based imputation outperforms other methods for imputing LC–MS metabolomics data: a comparative study. *BMC Bioinformatics* 2019;**20**:492.
139. Oba S, Sato MA, Takemasa I, *et al*. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* 2003;**19**:2088–96.
140. Nyamundanda G, Brennan L, Gormley IC. Probabilistic principal component analysis for metabolomic data. *BMC Bioinformatics* 2010;**11**:571.
141. Chai LE, Law CK, Mohamad MS, *et al*. Investigating the effects of imputation methods for modelling gene networks using a dynamic bayesian network from gene expression data. *Malays J Med Sci* 2014;**21**:20–7.
142. Di Guida R, Engel J, Allwood JW, *et al*. Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics* 2016;**12**:93.
143. Erler NS, Rizopoulos D, Rosmalen J, *et al*. Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach. *Stat Med* 2016;**35**:2955–74.
144. Xia J, Psychogios N, Young N, *et al*. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res* 2009;**37**:W652–60.
145. Tran TB, Bergen PJ, Creek DJ, *et al*. Synergistic killing of polymyxin B in combination with the antineoplastic drug mitotane against Polymyxin-susceptible and -resistant Acinetobacter baumannii: a metabolomic study. *Front Pharmacol* 2018;**9**:359.
146. Tang J, Fu J, Wang Y, *et al*. ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Brief Bioinform* 2020;**21**:621–36.
147. Troyanskaya O, Cantor M, Sherlock G, *et al*. Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001;**17**:520–5.
148. Verma P, Devaraj J, Skiles JL, *et al*. A metabolomics approach for early prediction of vincristine-induced peripheral neuropathy. *Sci Rep* 2020;**10**:9659.
149. Rotroff DM, Oki NO, Liang X, *et al*. Pharmacometabolomic assessment of metformin in non-diabetic, African Americans. *Front Pharmacol* 2016;**7**:135.
150. Rotroff DM, Shahin MH, Gurley SB, *et al*. Pharmacometabolomic assessments of atenolol and hydrochlorothiazide treatment reveal novel drug response phenotypes. *CPT Pharmacometrics Syst Pharmacol* 2015;**4**:669–79.
151. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA* 2000;**97**:10101–6.

152. Gan X, Liew AW, Yan H. Microarray missing data imputation based on a set theoretic framework and biological knowledge. *Nucleic Acids Res* 2006;**34**:1608–19.
153. Turi KN, Romick-Rosendale L, Gebretsadik T, *et al*. Using urine metabolomics to understand the pathogenesis of infant respiratory syncytial virus (RSV) infection and its role in childhood wheezing. *Metabolomics* 2018;**14**:135.
154. Kumar N, Hoque MA, Shahjaman M, *et al*. Metabolomic biomarker identification in presence of outliers and missing values. *Biomed Res Int* 2017;**2017**:2437608.
155. Zhang L, Wei TT, Li Y, *et al*. Functional metabolomics characterizes a key role for *N*-acetylneuraminic acid in coronary artery diseases. *Circulation* 2018;**137**:1374–90.
156. Begou O, Gika HG, Theodoridis GA, *et al*. Quality control and validation issues in LC–MS metabolomics. *Methods Mol Biol* 2018;**1738**:15–26.
157. Manier SK, Keller A, Schäper J, *et al*. Untargeted metabolomics by high resolution mass spectrometry coupled to normal and reversed phase liquid chromatography as a tool to study the in vitro biotransformation of new psychoactive substances. *Sci Rep* 2019;**9**:2741.
158. Drotleff B, Lammerhofer M. Guidelines for selection of internal standard-based normalization strategies in untargeted lipidomic profiling by LC-HR-MS/MS. *Anal Chem* 2019;**91**:9836–43.
159. Shen XT, Gong XY, Cai YP, *et al*. Normalization and integration of large-scale metabolomics data using support vector regression. *Metabolomics* 2016;**12**:89–100.
160. Luan H, Ji F, Chen Y, *et al*. statTarget: a streamlined tool for signal drift correction and interpretations of quantitative mass spectrometry-based omics data. *Anal Chim Acta* 2018;**1036**:66–72.
161. Meinicke P, Klanke S, Memisevic R, *et al*. Principal surfaces from unsupervised kernel regression. *IEEE Trans Pattern Anal Mach Intell* 2005;**27**:1379–91.
162. Cervellera C, Macciò D. Local linear regression for function learning: an analysis based on sample discrepancy. *IEEE Trans Neural Netw Learn Syst* 2014;**25**:2086–98.
163. Gamst A, Wolfson T, Parry B. Local polynomial regression modeling of human plasma melatonin levels. *J Biol Rhythms* 2004;**19**:164–74.
164. Zheng WB, Zou Y, Elsheikha HM, *et al*. Serum metabolomic alterations in beagle dogs experimentally infected with Toxocara canis. *Parasit Vectors* 2019;**12**:447.
165. Partha R, Kowalczyk A, Clark NL, *et al*. Robust method for detecting convergent shifts in evolutionary rates. *Mol Biol Evol* 2019;**36**:1817–30.
166. Cooper SM, Baker JS, Tong RJ, *et al*. The repeatability and criterion related validity of the 20 m multistage fitness test as a predictor of maximal oxygen uptake in active young men. *Br J Sports Med* 2005;**39**(4):e19.
167. Schou M, Gustafsson F, Kjaer A, *et al*. Long-term clinical variation of NT-proBNP in stable chronic heart failure patients. *Eur Heart J* 2007;**28**(2):177–82.
168. Altman DG, Bland JM. Detecting skewness from summary information. *BMJ* 1996;**313**(7066):1200.
169. De Livera AM, Dias DA, De Souza D, *et al*. Normalizing and integrating metabolomics data. *Anal Chem* 2012;**84**(24):10768–76.
170. Tang J, Fu J, Wang Y, *et al*. Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains. *Mol Cell Proteomics* 2019;**18**:1683–99.
171. Box GEP, Cox DR. An analysis of transformations. *J R Stat Soc Series B Stat Methodol* 1964;**26**(2):211–52.
172. Manikandan S. Data transformation. *J Pharmacol Pharmacother* 2010;**1**(2):126–7.
173. van den Berg RA, Hoefsloot HC, Westerhuis JA, *et al*. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* 2006;**7**:142.
174. Combrink M, du Preez I, Ronacher K, *et al*. Time-dependent changes in urinary metabolome before and after intensive phase tuberculosis therapy: a pharmacometabolomics study. *OMICS* 2019;**23**(11):560–72.
175. Raji Reddy C, Rani Valleti R, Dilipkumar U. One-pot sequential propargylation/cycloisomerization: a facile [4+2]-benzannulation approach to carbazoles. *Chemistry* 2016;**22**(7):2501–6.
176. Zheng H, Cai A, Zhou Q, *et al*. Optimal preprocessing of serum and urine metabolomic data fusion for staging prostate cancer through design of experiment. *Anal Chim Acta* 2017;**991**:68–75.
177. Sun Y, Kim JH, Vangipuram K, *et al*. Pharmacometabolomics reveals a role for histidine, phenylalanine, and threonine in the development of paclitaxel-induced peripheral neuropathy. *Breast Cancer Res Treat* 2018;**171**(3):657–66.
178. Sakia RM. The Box-Cox transformation technique—a review. *J R Stat Soc Series D Stat* 1992;**41**(2):169–78.
179. Banales JM, Iñarrairaegui M, Arbelaiz A, *et al*. Serum metabolites as diagnostic biomarkers for cholangiocarcinoma, hepatocellular carcinoma, and primary sclerosing cholangitis. *Hepatology* 2019;**70**(2):547–62.
180. Troisi J, Sarno L, Martinelli P, *et al*. A metabolomics-based approach for non-invasive diagnosis of chromosomal anomalies. *Metabolomics* 2017;**13**(11):140–51.
181. Sugimoto M, Kawakami M, Robert M, *et al*. Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. *Curr Bioinform* 2012;**7**(1):96–108.
182. Yang Q, Hong J, Li Y, *et al*. A novel bioinformatics approach to identify the consistently well-performing normalization strategy for current metabolomic studies. *Brief Bioinform* 2020;**21**:2142–52.
183. Astrand M. Contrast normalization of oligonucleotide arrays. *J Comput Biol* 2003;**10**(1):95–102.
184. Fu JB, Tang J, Wang YX, *et al*. Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification. *Front Pharmacol* 2018;**9**:681.
185. Saccenti E. Correlation patterns in experimental data are affected by normalization procedures: consequences for data analysis and network inference. *J Proteome Res* 2017;**16**(2):619–34.
186. Ejigu BA, Valkenborg D, Baggerman G, *et al*. Evaluation of normalization methods to pave the way towards large-scale LC–MS-based metabolomics profiling experiments. *OMICS* 2013;**17**(9):473–85.
187. Karpievitch YV, Nikolic SB, Wilson R, *et al*. Metabolomics data normalization with EigenMS. *PLoS One* 2014;**9**(12):e116221.
188. Karpievitch YV, Taverner T, Adkins JN, *et al*. Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition. *Bioinformatics* 2009;**25**(19):2573–80.
189. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* 2007;**3**(9):1724–35.

190. Andjelkovic V, Thompson R. Changes in gene expression in maize kernel in response to water and salt stress. *Plant Cell Rep* 2006;**25**(1):71–9.
191. Wang W, Zhou H, Lin H, *et al*. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal Chem* 2003;**75**(18):4818–26.
192. Crawford LR, Morrison JD. Computer methods in analytical mass spectrometry—identification of an unknown compound in a catalog. *Anal Chem* 1968;**40**(10):1464–74.
193. Warrack BM, Hnatyshyn S, Ott KH, *et al*. Normalization strategies for metabonomic analysis of urine samples. *J Chromatogr B Analyt Technol Biomed Life Sci* 2009;**877**(5–6):547–52.
194. Emwas AH, Saccenti E, Gao X, *et al*. Recommended strategies for spectral processing and post-processing of 1D (1)H-NMR data of biofluids with a particular focus on urine. *Metabolomics* 2018;**14**(3):31.
195. Bolstad BM, Irizarry RA, Astrand M, *et al*. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003;**19**(2):185–93.
196. De Livera AM, Sysi-Aho M, Jacob L, *et al*. Statistical methods for handling unwanted variation in metabolomics data. *Anal Chem* 2015;**87**(7):3606–15.
197. Li B, Tang J, Yang Q, *et al*. Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis. *Sci Rep* 2016;**6**:38881.
198. Puchades-Carrasco L, Palomino-Schätzlein M, Pérez-Rambla C, *et al*. Bioinformatics tools for the analysis of NMR metabolomics studies focused on the identification of clinically relevant biomarkers. *Brief Bioinform* 2016;**17**(3):541–52.
199. González-Domínguez R, García-Barrera T, Vitorica J, *et al*. Region-specific metabolic alterations in the brain of the APP/PS1 transgenic mice of Alzheimer's disease. *Biochim Biophys Acta* 2014;**1842**(12 Pt A):2395–402.
200. Chen X, de Seymour JV, Han TL, *et al*. Metabolomic biomarkers and novel dietary factors associated with gestational diabetes in China. *Metabolomics* 2018;**14**(11):149.
201. Martinez-Lozano Sinues P, Kohler M, Zenobi R. Human breath analysis may support the existence of individual metabolic phenotypes. *PLoS One* 2013;**8**(4):e59909.
202. Cui X, Yang Q, Li B, *et al*. Assessing the effectiveness of direct data merging strategy in long-term and large-scale pharmacometabonomics. *Front Pharmacol* 2019;**10**:127.
203. Dieterle F, Ross A, Schlotterbeck G, *et al*. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal Chem* 2006;**78**(13):4281–90.
204. Liang YJ, Lin YT, Chen CW, *et al*. SMART: statistical metabolomics analysis—an R tool. *Anal Chem* 2016;**88**(12):6334–41.
205. Jiang L, Lee SC, Ng TC. Pharmacometabonomics analysis reveals serum formate and acetate potentially associated with varying response to gemcitabine-carboplatin chemotherapy in metastatic breast cancer patients. *J Proteome Res* 2018;**17**(3):1248–57.
206. Hu CX, Xu GW. Mass-spectrometry-based metabolomics analysis for foodomics. *Trends Anal Chem* 2013;**52**:36–46.
207. Contrepois K, Jiang L, Snyder M. Optimized analytical procedures for the untargeted metabolomic profiling of human urine and plasma by combining hydrophilic interaction (HILIC) and reverse-phase liquid chromatography (RPLC)-mass spectrometry. *Mol Cell Proteomics* 2015;**14**(6):1684–95.
208. Eriksson L, Antti H, Gottfries J, *et al*. Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Anal Bioanal Chem* 2004;**380**(3):419–29.
209. Smilde AK, van der Werf MJ, Bijlsma S, *et al*. Fusion of mass spectrometry-based metabolomics data. *Anal Chem* 2005;**77**(20):6729–36.
210. Keun HC, Ebbels TMD, Antti H, *et al*. Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling. *Anal Chim Acta* 2003;**490**(1–2):265–76.
211. Struck W, Siluk D, Yumba-Mpanga A, *et al*. Liquid chromatography tandem mass spectrometry study of urinary nucleosides as potential cancer markers. *J Chromatogr A* 2013;**1283**:122–31.
212. Masson P, Spagou K, Nicholson JK, *et al*. Technical and biological variation in UPLC-MS-based untargeted metabolic profiling of liver extracts: application in an experimental toxicity study on galactosamine. *Anal Chem* 2011;**83**(3):1116–23.
213. Leichtle AB, Nuoffer JM, Ceglarek U, *et al*. Serum amino acid profiles and their alterations in colorectal cancer. *Metabolomics* 2012;**8**(4):643–53.
214. Gromski PS, Xu Y, Hollywood KA, *et al*. The influence of scaling metabolomics data on model classification accuracy. *Metabolomics* 2015;**11**(3):684–95.
215. Kultima K, Nilsson A, Scholz B, *et al*. Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides. *Mol Cell Proteomics* 2009;**8**(10):2285–95.
216. Huber W, von Heydebreck A, Sültmann H, *et al*. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 2002;**18**(Suppl 1):S96–104.
217. Hochrein J, Zacharias HU, Taruttis F, *et al*. Data normalization of (1)H NMR metabolite fingerprinting data sets in the presence of unbalanced metabolite regulation. *J Proteome Res* 2015;**14**(8):3217–28.
218. Ibarra R, Dazard JE, Sandlers Y, *et al*. Metabolomic analysis of liver tissue from the VX2 rabbit model of secondary liver tumors. *HPB Surg* 2014;**2014**:310372.
219. Zhang S, Zheng C, Lanza IR, *et al*. Interdependence of signal processing and analysis of urine 1H NMR spectra for metabolic profiling. *Anal Chem* 2009;**81**(15):6080–8.
220. Trezzi JP, Jager C, Galozzi S, *et al*. Metabolic profiling of body fluids and multivariate data analysis. *MethodsX* 2017;**4**:95–103.
221. Redestig H, Fukushima A, Stenlund H, *et al*. Compensation for systematic cross-contribution improves normalization of mass spectrometry based metabolomics data. *Anal Chem* 2009;**81**(19):7974–80.
222. Jacob L, Gagnon-Bartsch JA, Speed TP. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics* 2016;**17**(1):16–28.
223. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* 2012;**13**(3):539–52.
224. Khoomrung S, Wanichthanarak K, Nookaew I, *et al*. Metabolomics and integrative omics for the development of Thai traditional medicine. *Front Pharmacol* 2017;**8**:474.

225. Wang S, Chen X, Dan D, *et al*. MetaboGroup S: a group entropy-based web platform for evaluating normalization methods in blood metabolomics data from maintenance hemodialysis patients. *Anal Chem* 2018;**90**(18): 11124–30.
226. McKennan C, Ober C, Estimation ND. Inference in metabolomics with nonrandom missing data and latent factors. *Ann Appl Stat* 2020;**14**(2):789–808.
227. Liu X, Locasale JW. Metabolomics: a primer. *Trends Biochem Sci* 2017;**42**(4):274–84.
228. Hira ZM, Gillies DFA. Review of feature selection and feature extraction methods applied on microarray data. *Adv Bioinformatics* 2015;**2015**:198363.
229. Divine G, Norton HJ, Hunt R, *et al*. Statistical grand rounds: a review of analysis and sample size calculation considerations for Wilcoxon tests. *Anesth Analg* 2013;**117**(3): 699–710.
230. Rosner B, Glynn RJ, Lee ML. The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics* 2006;**62**(1):185–92.
231. Lengacher CA, Reich RR, Paterson CL, *et al*. A large randomized trial: effects of mindfulness-based stress reduction (MBSR) for breast cancer (BC) survivors on salivary cortisol and IL-6. *Biol Res Nurs* 2019;**21**(1):39–49.
232. McHugh ML. The chi-square test of independence. *Biochem Med* 2013;**23**(2):143–9.
233. Sonn BJ, Saben JL, McWilliams G, *et al*. Predicting response to lisinopril in treating hypertension: a pilot study. *Metabolomics* 2019;**15**(10):133.
234. Athreya A, Iyer R, Neavin D, *et al*. Augmentation of physician assessments with multi-omics enhances predictability of drug response: a case study of major depressive disorder. *IEEE Comput Intell Mag* 2018;**13**(3):20–31.
235. He C, Liu Y, Wang Y, *et al*. SLive_RefAppend (1)H NMR based pharmacometabolomics analysis of metabolic phenotype on predicting metabolism characteristics of losartan in healthy volunteers. *J Chromatogr B Analyt Technol Biomed Life Sci* 2018;**1095**:15–23.
236. Miolo G, Muraro E, Caruso D, *et al*. Pharmacometabolomics study identifies circulating spermidine and tryptophan as potential biomarkers associated with the complete pathological response to trastuzumab-paclitaxel neoadjuvant therapy in HER-2 positive breast cancer. *Oncotarget* 2016;**7**(26):39809–22.
237. Río J, Comabella M, Montalban X. Predicting responders to therapies for multiple sclerosis. *Nat Rev Neurol* 2009;**5**(10):553–60.
238. Hampel H, Vergallo A, Perry G, *et al*. The Alzheimer precision medicine initiative. *J Alzheimers Dis* 2019;**68**(1):1–24.
239. Waddington KE, Papadaki A, Coelewij L, *et al*. Using serum metabolomics to predict development of anti-drug antibodies in multiple sclerosis patients treated with IFNβ. *Front Immunol* 2020;**11**:1527.
240. Broughton-Neiswanger LE, Rivera-Velez SM, Suarez MA, *et al*. Pharmacometabolomics with a combination of PLS-DA and random forest algorithm analyses reveal meloxicam alters feline plasma metabolite profiles. *J Vet Pharmacol Ther* 2020;**43**(6):591–601.
241. Li Y, Wang L, Ju L, *et al*. A systematic strategy for screening and application of specific biomarkers in hepatotoxicity using metabolomics combined with ROC curves and SVMs. *Toxicol Sci* 2016;**150**(2):390–9.
242. Tzeng J, Lu HH, Li WH. Multidimensional scaling for large genomic data sets. *BMC Bioinformatics* 2008;**9**:179.
243. Borgan E, Sitter B, Lingjaerde OC, *et al*. Merging transcriptomics and metabolomics—advances in breast cancer profiling. *BMC Cancer* 2010;**10**:628.
244. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. *Science* 2000;**290**(5500):2319–23.
245. Wang S, Chen H, Zheng Y, *et al*. Transcriptomics- and metabolomics-based integration analyses revealed the potential pharmacological effects and functional pattern of in vivo radix Paeoniae Alba administration. *Chinas Med* 2020;**15**:52.
246. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. *Science* 2000;**290**(5500):2323–6.
247. Zhou B, Jin W. Visualization of single cell RNA-Seq data using t-SNE in R. *Methods Mol Biol* 2020;**2117**:159–67.
248. Daley M, Dekaban G, Bartha R, *et al*. Metabolomics profiling of concussion in adolescent male hockey players: a novel diagnostic method. *Metabolomics* 2016;**12**(12):185–93.
249. Klassen A, Faccio AT, Canuto GA, *et al*. Metabolomics: definitions and significance in systems biology. *Adv Exp Med Biol* 2017;**965**:3–17.
250. Witting M, Böcker S. Current status of retention time prediction in metabolite identification. *J Sep Sci* 2020;**43**(9–10):1746–54.
251. Donatti A, Canto AM, Godoi AB, *et al*. Circulating metabolites as potential biomarkers for neurological disorders-metabolites in neurological disorders. *Metabolites* 2020;**10**(10):389.
252. da Silva RR, Dorrestein PC, Quinn RA. Illuminating the dark matter in metabolomics. *Proc Natl Acad Sci USA* 2015;**112**(41):12549–50.
253. Wishart DS, Jewison T, Guo AC, *et al*. HMDB 3.0–the human metabolome database in 2013. *Nucleic Acids Res* 2013;**41**(Database issue):D801–7.
254. Kaddurah-Daouk R, Weinshilboum R. Metabolomic signatures for drug response phenotypes: pharmacometabolomics enables precision medicine. *Clin Pharmacol Ther* 2015;**98**(1):71–5.
255. Kim S, Chen J, Cheng T, *et al*. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* 2019;**47**(D1):D1102–9.
256. Kim S, Thiessen PA, Bolton EE, *et al*. PubChem substance and compound databases. *Nucleic Acids Res* 2016;**44**(D1):D1202–13.
257. Wang Y, Bryant SH, Cheng T, *et al*. PubChem BioAssay: 2017 update. *Nucleic Acids Res* 2017;**45**(D1):D955–63.
258. Samaraweera MA, Hall LM, Hill DW, *et al*. Evaluation of an artificial neural network retention index model for chemical structure identification in nontargeted metabolomics. *Anal Chem* 2018;**90**(21):12752–60.
259. Horai H, Arita M, Kanaya S, *et al*. MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* 2010;**45**(7):703–14.
260. Cuthbertson DJ, Johnson SR, Piljac-Žegarac J, *et al*. Accurate mass-time tag library for LC/MS-based metabolite profiling of medicinal plants. *Phytochemistry* 2013;**91**:187–97.
261. Blaženović I, Kind T, Ji J, *et al*. Software tools and approaches for compound identification of LC–MS/MS data in metabolomics. *Metabolites* 2018;**8**(2):31.
262. Guijas C, Montenegro-Burke JR, Domingo-Almenara X, *et al*. METLIN: a technology platform for identifying knowns and unknowns. *Anal Chem* 2018;**90**(5):3156–64.
263. Steuer AE, Kaelin D, Boxler MI, *et al*. Comparative untargeted metabolomics analysis of the psychostimulants

3,4-methylenedioxy-methamphetamine (MDMA), amphetamine, and the novel psychoactive substance mephedrone after controlled drug administration to humans. *Metabolites* 2020;**10**(8):306.
264. Sud M, Fahy E, Cotter D, *et al*. LIPID MAPS-nature lipidomics gateway: an online resource for students and educators interested in lipids. *J Chem Educ* 2012;**89**(2):291–2.
265. Wang Y, Shi F, Cao L, *et al*. Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images. *Curr Bioinform* 2019;**14**(4):282–94.
266. Fahy E, Alvarez-Jarreta J, Brasher CJ, *et al*. LipidFinder on LIPID MAPS: peak filtering, MS searching and statistical analysis for lipidomics. *Bioinformatics* 2019;**35**(4):685–7.
267. Degtyarenko K, de Matos P, Ennis M, *et al*. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* 2008;**36**(Database issue):D344–50.
268. Hastings J, Owen G, Dekker A, *et al*. ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res* 2016;**44**(D1):D1214–9.
269. Moreno P, Beisken S, Harsha B, *et al*. BiNChE: a web tool and library for chemical enrichment analysis based on the ChEBI ontology. *BMC Bioinformatics* 2015;**16**(1):56.
270. Wang Y, Zhang S, Li F, *et al*. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res* 2020;**48**:D1031–41.
271. Li YH, Yu CY, Li XX, *et al*. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res* 2018;**46**:D1121–7.
272. Zhang W, Chen Y, Jiang H, *et al*. Integrated strategy for accurately screening biomarkers based on metabolomics coupled with network pharmacology. *Talanta* 2020;**211**:120710.
273. Lipkus AH, Watkins SP, Gengras K, *et al*. Recent changes in the scaffold diversity of organic chemistry as seen in the CAS registry. *J Org Chem* 2019;**84**(21):13948–56.
274. Wills TJ, Lipkus AH. Structural approach to assessing the innovativeness of new drugs finds accelerating rate of innovation. *ACS Med Chem Lett* 2020;**11**(11):2114–9.
275. Yin J, Li F, Zhou Y, *et al*. INTEDE: interactome of drug-metabolizing enzymes. *Nucleic Acids Res* 2021;**49**:D1233–43.
276. Tang J, Wu X, Mou M, *et al*. GIMICA: host genetic and immune factors shaping human microbiota. *Nucleic Acids Res* 2021;**49**:D715–22.
277. Pearce EL, Pearce EJ. Metabolic pathways in immune cell activation and quiescence. *Immunity* 2013;**38**(4):633–43.
278. Kanehisa M, Sato Y, Kawashima M, *et al*. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016;**44**(D1):D457–62.
279. Kanehisa M, Sato Y, Furumichi M, *et al*. New approach for understanding genome variations in KEGG. *Nucleic Acids Res* 2019;**47**(D1):D590–5.
280. Kanehisa M, Furumichi M, Tanabe M, *et al*. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**(D1):D353–61.
281. Kim I, Choi S, Kim S. BRCA-pathway: a structural integration and visualization system of TCGA breast cancer data on KEGG pathways. *BMC Bioinformatics* 2018;**19**(Suppl 1):42.
282. Sidiropoulos K, Viteri G, Sevilla C, *et al*. Reactome enhanced pathway visualization. *Bioinformatics* 2017;**33**(21):3461–7.
283. Jassal B, Matthews L, Viteri G, *et al*. The reactome pathway knowledgebase. *Nucleic Acids Res* 2020;**48**(D1):D498–503.
284. Fabregat A, Jupe S, Matthews L, *et al*. The Reactome pathway knowledgebase. *Nucleic Acids Res* 2018;**46**(D1):D649–55.
285. Fabregat A, Sidiropoulos K, Viteri G, *et al*. Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* 2017;**18**(1):142.
286. Kutmon M, Riutta A, Nunes N, *et al*. WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res* 2016;**44**(D1):D488–94.
287. Slenter DN, Kutmon M, Hanspers K, *et al*. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res* 2018;**46**(D1):D661–7.
288. Jennen DG, Gaj S, Giesbertz PJ, *et al*. Biotransformation pathway maps in WikiPathways enable direct visualization of drug metabolism related expression changes. *Drug Discov Today* 2010;**15**(19–20):851–8.
289. Caspi R, Billington R, Ferrer L, *et al*. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 2016;**44**(D1):D471–80.
290. Caspi R, Billington R, Keseler IM, *et al*. The MetaCyc database of metabolic pathways and enzymes—a 2019 update. *Nucleic Acids Res* 2020;**48**(D1):D445–53.
291. Karp PD, Caspi R. A survey of metabolic databases emphasizing the MetaCyc family. *Arch Toxicol* 2011;**85**(9):1015–33.
292. Frolkis A, Knox C, Lim E, *et al*. SMPDB: the small molecule pathway database. *Nucleic Acids Res* 2010;**38**(Database issue):D480–7.
293. Jewison T, Su Y, Disfany FM, *et al*. SMPDB 2.0: big improvements to the small molecule pathway database. *Nucleic Acids Res* 2014;**42**(Database issue):D478–84.
294. Backes C, Kehl T, Stöckel D, *et al*. miRPathDB: a new dictionary on microRNAs and target pathways. *Nucleic Acids Res* 2017;**45**(D1):D90–6.
295. Chawade A, Alexandersson E, Levander F. Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. *J Proteome Res* 2014;**13**(6):3114–20.
296. Castanar L, Parella T. Broadband 1H homodecoupled NMR experiments: recent developments, methods and applications. *Magn Reson Chem* 2015;**53**(6):399–426.
297. Lutz NW, Beraud E, Cozzone PJ. Metabolomic analysis of rat brain by high resolution nuclear magnetic resonance spectroscopy of tissue extracts. *J Vis Exp* 2014;**91**:51829.
298. Kim HK, Choi YH, Verpoorte R. NMR-based plant metabolomics: where do we stand, where do we go? *Trends Biotechnol* 2011;**29**(6):267–75.
299. Wishart DS. Advances in metabolite identification. *Bioanalysis* 2011;**3**(15):1769–82.
300. Markley JL, Bruschweiler R, Edison AS, *et al*. The future of NMR-based metabolomics. *Curr Opin Biotechnol* 2017;**43**:34–40.
301. Emwas AH, Roy R, McKay RT, *et al*. NMR spectroscopy for metabolomics research. *Metabolites* 2019;**9**(7):123.
302. Gika H, Virgiliou C, Theodoridis G, *et al*. Untargeted LC/MS-based metabolic phenotyping (metabonomics/metabolomics): the state of the art. *J Chromatogr B Analyt Technol Biomed Life Sci* 2019;**1117**:136–47.
303. Cui L, Lu H, Lee YH. Challenges and emergent solutions for LC–MS/MS based untargeted metabolomics in diseases. *Mass Spectrom Rev* 2018;**37**(6):772–92.
304. Zhou B, Xiao JF, Tuli L, *et al*. LC–MS-based metabolomics. *Mol Biosyst* 2012;**8**(2):470–81.
305. Fang ZZ, Gonzalez FJ. LC–MS-based metabolomics: an update. *Arch Toxicol* 2014;**88**(8):1491–502.

306. Chaleckis R, Meister I, Zhang P, *et al*. Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. *Curr Opin Biotechnol* 2019;**55**: 44–50.
307. Lubes G, Goodarzi M. GC-MS based metabolomics used for the identification of cancer volatile organic compounds as biomarkers. *J Pharm Biomed Anal* 2018;**147**:313–22.
308. Begou O, Gika HG, Wilson ID, *et al*. Hyphenated MS-based targeted approaches in metabolomics. *Analyst* 2017;**142**(17):3079–100.
309. Luan H, Wang X, Cai Z. Mass spectrometry-based metabolomics: targeting the crosstalk between gut microbiota and brain in neurodegenerative disorders. *Mass Spectrom Rev* 2019;**38**(1):22–33.