Received: 22 July 2020 | Revised: 27 October 2020 | Accepted: 29 October 2020

DOI: 10.1002/mas.21672

REVIEW ARTICLE

WILEY

Evaluating and minimizing batch effects in metabolomics

Wei Han | **Liang Li**

Department of Chemistry, University of Alberta, Edmonton, Alberta, Canada

Correspondence
Liang Li, Department of Chemistry, University of Alberta, Chemistry Centre W3-39C, Edmonton, Alberta T6G 2G2, Canada.
Email: liang.li@ualberta.ca

Abstract

Determining metabolomic differences among samples of different phenotypes is a critical component of metabolomics research. With the rapid advances in analytical tools such as ultrahigh-resolution chromatography and mass spectrometry, an increasing number of metabolites can now be profiled with high quantification accuracy. The increased detectability and accuracy raise the level of stringiness required to reduce or control any experimental artifacts that can interfere with the measurement of phenotype-related metabolome changes. One of the artifacts is the batch effect that can be caused by multiple sources. In this review, we discuss the origins of batch effects, approaches to detect interbatch variations, and methods to correct unwanted data variability due to batch effects. We recognize that minimizing batch effects is currently an active research area, yet a very challenging task from both experimental and data processing perspectives. Thus, we try to be critical in describing the performance of a reported method with the hope of stimulating further studies for improving existing methods or developing new methods.

KEYWORDS

batch effect, mass spectrometry, metabolome analysis, metabolomics, NMR

1 | INTRODUCTION

Metabolomics, focusing on the comprehensive and systematic analysis of small molecules in a biological system, has become a rapidly growing field in many application areas, including biomarker discovery (Xia et al., 2013), drug development (Kell, 2006), and precision medicine (Wishart, 2016). Quantitative metabolomics uses analytical techniques, such as nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS), to perform absolute or relative quantification of metabolites in comparative samples with the main objective of investigating the metabolic changes associated with phenotypes (Y. Wu & Li, 2016).

Because of the need to use a large sample size for achieving the desired statistical power of comparative analysis (Button et al., 2013), metabolomics often involves the analysis of hundreds or thousands of biological samples, for which the analytical process itself may take a long time. In addition, samples may be collected in multiple batches from the same or different laboratories (e.g., longitudinal studies and multicenter validation studies). Although continuous analysis of all samples on a dedicated platform is preferable, many studies have been split into multiple batches due to limitations in instrumental availability or the timeline of sample collection (Thonusin et al., 2017). Here, a batch is defined as

Abbreviations: ANOVA, analysis of variance; CIL, chemical isotope labeling; ComBat, combating batch effects; HCA, hierarchical clustering analysis; ICA, independent component analysis; IS, internal standard; LOESS, locally estimated scatterplot smoothing; LS, least-squares; NOMIS, normalization using optimal selection of multiple internal standards; PBS, phosphate-buffered saline; PC, principal component; PCA, principal component analysis; QC, quality control; QCM, quality control metabolites; QC-RFSC, quality control-based random forest signal correction; QC-RLSC, quality control-based robust LOESS signal correction; QC-RSC, quality control-based robust spline correction; QC-SVRC, quality control-based support vector regression correction; RSD, relative standard deviation; SVR, support vector regression; UMS, universal metabolome standard.

a set of samples processed and analyzed by the same experimental procedure (i.e., same operator and instrument) in an uninterrupted manner (Wehrens et al., 2016). Moreover, because of the high workload, some large-scale studies are inevitably conducted by multiple operators or even several collaborating laboratories (Goh et al., 2017).

To process the data of multiple batches, some researchers have successfully adopted a meta-analysis approach (Goveia et al., 2016; Patti et al., 2012), which analyzes the datasets separately and then finds the common significant metabolites in a second-order comparison. However, in a meta-analysis, the statistical power of each subset is limited by its relatively small sample size (Goh et al., 2017). As increasing sample size can generally improve the statistical performance, merging data of multiple batches remains highly desirable for in-depth metabolomics. For example, Salerno et al. (2017) reported that doubling the sample size by combining two batches increased the statistical power of their analysis, with area-under-the-curve values of the receiver-operating characteristic curves improved.

A major hurdle to merging multiple batches is the existence of batch effects, which refer to the situation that the quantitative results of different batches significantly differ due to irrelevant factors (Leek et al., 2010). In addition to biological variations associated with the phenotypes being studied and other biological factors, there are also analytical variations arising from the experimental process or instrumental analysis. Analyzed independently, each batch may have its unique analytical variations. When multiple batches are directly merged without proper treatments, the overall analytical variability, a major confounding factor for revealing the metabolic changes of interest, may significantly increase, thereby hindering the statistical performance. For instance, several studies have shown cases that interbatch variations were more significant than interphenotype variations, dominating in the statistical analysis and preventing the true metabolic changes to be highlighted (Boccard et al., 2019; Deng et al., 2019; Zhao et al., 2016). Moreover, when the batch-group design is unbalanced (i.e., sample numbers of a specific phenotype are not evenly distributed among batches), batch-related variations can be confused with interphenotype differences, and the corresponding findings may become misleading (Baggerly et al., 2004). Therefore, unidentified interbatch variability is a substantial challenge to the validity and reproducibility of quantitative metabolomics discoveries.

In recent years, with rapid advances in sample preparation methods and analytical techniques for accurate quantification of an increasing number of metabolites, metabolomics has become a popular high-throughput tool for studying large numbers of samples (Dunn et al., 2015; Fuhrer & Zamboni, 2015; Soininen et al., 2015). To generate any meaningful results from large-scale metabolomics, the impacts of interbatch variations must be minimized by experimental or data-processing strategies. In the literature, several experimental methods, such as the use of internal metabolite standards (Bijlsma et al., 2006; Fei et al., 2014), have been proposed to correct batch effects in metabolomics. In addition, genomics and proteomics, which were established earlier than metabolomics, have also been facing the problem of batch effects, and researchers have developed various computational approaches to remove or minimize batch effects during data processing (Gregori et al., 2012; Haghverdi et al., 2018; Karpievitch et al., 2009; Tung et al., 2017). As all the omics data have the common form of measuring multiple variables across a large number of samples, some of these computational corrections can potentially be applied to metabolomics. For example, the combating batch effects (ComBat) function, which was originally developed for microarray data, has also been used to adjust metabolomics data (Reisetter et al., 2017; Sánchez-Illana et al., 2018).

This review covers the topic of minimizing interbatch variations in metabolomics. The origins of batch effects are discussed first, followed by a summary of frequently used methods for identifying and evaluating interbatch variations. Next, existing strategies for batch effect correction are presented. Some techniques require additional experimental steps, whereas others directly adjust the results using univariate or multivariate algorithms. Finally, as there is no unified standard for dealing with batch effects in metabolomics, we summarize the advantages and limitations of each strategy and emphasize that the user may choose a combination of methods that are most suitable to the data being studied or most convenient for interstudy comparisons (Tables 1 and 2).

In the literature, there are publications covering the topic of ensuring the overall quality of clinical metabolomics (Long et al., 2020) or using computational techniques to improve the quality of metabolomics data (Q. Yang et al., 2020). Still, this review focuses explicitly on the roles of experimental precautions and computational adjustments in dealing with batch effects. With experience drawn from years of metabolomics work, we aim to provide a more comprehensive and in-depth review of the topic, including detailed introductions to various methods, and to propose a standardized protocol to overcome batch effects in quantitative metabolomics.

We note that as the selectivity, sensitivity and metabolite identification ability of untargeted analyses have been greatly improved, there is an increasing number of

TABLE 1 Summary of representative batch effect correction methods

Experimental requirement	Correction method	Reference	Most suitable situation	Main pitfalls
None	Median normalization	Atwal et al. (2015)	All samples have similar metabolite concentration distributions	Not suitable for dealing with the situation that metabolites vary in different patterns
	Unit-norm	Scholz et al. (2004)		
	Quantile normalization	Brodsky et al. (2010)		
	Probabilistic quotient normalization	Di Guida et al. (2016)		
	Combatting batch effects	Johnson et al. (2007)	The study has a well-balanced batch-group design	Data must have balanced batch-group design
	Two-way analysis of variance	Boccard et al. (2019)		
	EigenMS	Karpievitch et al. (2014)		
	WaveICA	Deng et al. (2019)		
Isotope-labeled internal standards (ISs)	Single-IS normalization	Bromke et al. (2015)	All metabolites and the IS have the same response to the batch effects	Single IS cannot represent all metabolites
	Multiple-IS normalization	S. Yang et al. (2010)	ISs are available for most metabolites or the available ISs can comprehensively reflect the batch effects	High cost of preparing many ISs; No IS for unidentified metabolites in untargeted analysis
		Zukunft et al. (2013)		
	Normalization using optimal selection of multiple internal standards	Sysi-Aho et al. (2007)		
	Cross-contribution-compensating multiple-standard normalization	Redestig et al. (2009)		
	Best-matched internal standard normalization	Boysen et al. (2018)	QC samples are available and cover all the metabolites	Only normalizes metabolites existing in the QC samples
Quality control (QC) samples	Batch normalizer	Wang et al. (2012)	A large number of QC samples can be analyzed throughout the study; The batch effects are mainly induced after sample collection	High time and cost of collecting enough QC data to achieve high-quality correction; Difficulty in making a large amount of QC by collecting aliquots from limited amounts of individual samples; Risk of over-fitting; Cannot reflect inter-batch differences induced during sample collection
	QC-least-squares (LS) linear regression	Wang et al. (2012)		
		Wehrens et al. (2016)		
	QC-quadratic regression	Thonusin et al. (2017)		
	QC-based robust locally estimated scatterplot smoothing signal correction	Dunn et al. (2011)		
	QC-based robust spline correction	Kirwan et al. (2013)		
		Brunius et al. (2016)		

(Continues)

TABLE 1 (Continued)

Experimental requirement	Correction method	Reference	Most suitable situation	Main pitfalls
	QC-based support vector regression correction	Kuligowski et al. (2015) Sanchez-Illana et al. (2018)		
	QC-based random forest signal correction	Luan et al. (2018)		
	NormAE	Rong et al. (2020)		
Quality control metabolites (QCMs)	QCM-Remove unwanted variations	de Livera et al. (2012) Livera et al. (2015)	Ideal QCMs are known	Not easy to choose the best QCMs
Chemical isotope labeling (CIL)	CIL-liquid chromatography-mass spectrometry	Guo & Li (2009) Peng et al. (2014)	The batch effects are mainly induced after the individual samples are mixed with the reference sample	Cannot cover batch effects arising before mixing with the reference sample

attempts to merge targeted and untargeted metabolomics (Cajka & Fiehn, 2015; Y. Li et al., 2012). Despite the fact that in targeted metabolomics, normalization based on isotope-labeled internal standards (IS), which are commonly available in targeted studies, is the gold standard for irrelevant variability removal (Boysen et al., 2018), many of the correction methods for untargeted metabolomics can also be used for targeted analyses when required. Hereinafter, we will not discuss targeted and untargeted approaches separately.

We also note that, though metabolome analysis may include the detection of some lipids, lipidome analysis often involves the use of a different workflow, including dedicated lipid extraction protocols and optimized separation conditions for lipids. As the extent of batch effects is dependent on the analytical techniques and methods used, it would not be too surprising that batch effects on lipidome analysis might be very different from those for metabolome analysis. This review focuses on batch effects on metabolome analysis. Some of the methods reviewed in this paper might be applicable for overcoming batch effects in lipidome analysis; however, more studies on batch effects exclusively focused on lipidome analysis are still needed.

2 | ORIGIN OF INTERBATCH VARIATIONS

Figure 1 illustrates the general experimental workflow of quantitative metabolomics (Y. Wu & Li, 2016). Briefly, biological samples are collected from the subjects with proper pretreatments, such as filtration of urine, clotting of blood, or metabolic quenching of cell cultures. After collection, samples are temporarily stored under specific conditions (e.g., in a −80°C freezer) until the analytical work begins. The sample processing step further treats the biological samples by extracting the metabolites and making the final samples suitable for instrumental analysis. Then the raw data reflecting metabolite concentrations are acquired by analytical instruments, whose performance determines the accuracy, precision, and metabolome coverage of the measurement.

NMR and MS are the two leading instruments in quantitative metabolomics. With nondestructive measurement over a wide dynamic range, NMR-based metabolomics generates highly reproducible quantification results (Song et al., 2011). However, the relatively poor sensitivity has limited its application in untargeted metabolomics, which favors high metabolome coverage to reveal more biological information (Issaq et al., 2009; Zhang et al., 2012). On the contrary, MS-based analysis has been coupled with chromatography or other

TABLE 2 Performance comparison of five strategies for batch effect correction

Criteria	Method			
	Sample-data-driven	Internal standards (ISs)	Quality control	Chemical isotope labeling
Remove batch effects induced during sample collection and storage	Depends	Partially (best for metabolites with ISs)	No	No
Cover all metabolites detected in the samples	Yes	Partially	No	Yes
Reference for each metabolite	No	Yes	Yes	Yes
Extra time or cost	No	high if many ISs are used	Acceptable	Acceptable

separation techniques to achieve high sensitivity and metabolome coverage (Lu et al., 2008). The most widely used MS-based platforms are gas chromatography-mass spectrometry (GC-MS) (Jonsson et al., 2004; Kanani et al., 2008) and liquid chromatography-mass spectrometry (LC-MS) (Theodoridis et al., 2008; B. Zhou et al., 2012).

After data acquisition, the raw data are aligned to form a matrix containing concentrations of metabolites (i.e., variables) across the samples (i.e., observations). At last, statistical analysis reveals the metabolic differences between the phenotypes of interest. Additionally, in some metabolomics studies, a preacquisition normalization step is added to balance the total amount of metabolites in each sample before instrumental analysis (Y. Wu & Li, 2012, 2016), whereas other studies implement a postacquisition treatment in data processing to computationally normalize the distribution of metabolite concentrations in each sample (Peralbo-Molina et al., 2015; Veselkov et al., 2011). Both normalizations can help reduce the analyte-irrelevant variations, and their roles in minimizing interbatch variations will be discussed later.

During each step throughout the workflow of quantitative metabolomics, analytical variations could be introduced. In Figure 2, we summarize the major sources of unwanted variations in quantitative metabolomics, as well as the typical strategies to minimize

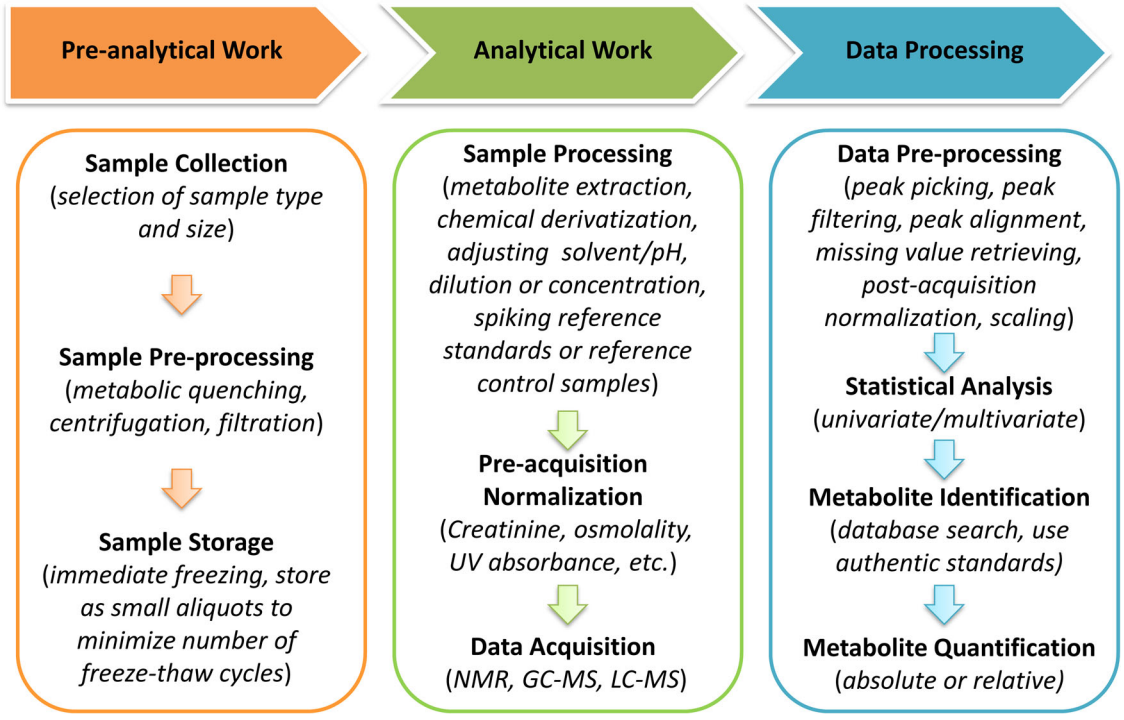


FIGURE 1 Workflow for quantitative metabolomics (experimental part) (Adapted with permission from Y. Wu & Li, 2016) [Color figure can be viewed at wileyonlinelibrary.com]

unwanted variations. First, a well-balanced study design can minimize the unwanted variations rising from irrelevant biological factors. Second, during the preanalytical stage, collection, initial preparation, and storage of samples may induce unwanted variations. Specifically, for each of these operations in a large-scale study, batch-to-batch variations due to differences in operators, collection containers, reagents, equipment, and others may be introduced. Last, the analytical process can be a major source of technical variations. Similar to that in the preanalytical stage, the factors leading to analyte-irrelevant variability include, but not limited to, operators conducting the experiment, reagents and containers for processing the samples and running the instruments, conditions of sample handling during the sample workup, storage of processed samples and standards, and the stability of analytical instruments. In some cases, using different parameters to process the raw spectra might be normal for multibatch instrumental analysis, but there is also a potential risk of increasing interbatch variations. To minimize and control the technical variations during the preanalytical and analytical stages, there is a need for developing and implementing a stringent analytical workflow with standard operating procedures (SOPs) in each step. Overall, after the data collection is done, data normalization and correction become the major strategy to overcome any unwanted variations.

It has been noticed that operator bias may cause extra variations in experimental research (Griffiths & Rosenfeld, 1954; Saenz et al., 1999). For example, when samples are transferred by manual pipetting, there is a risk of operator-wise variability potentially due to differences in pipetting techniques, as revealed in the study by Pandya et al. (2010). When different batches of samples are processed by different operators independently, the cumulative differences in sample volumes may become nonnegligible.

Moreover, chemical reagents are often added to samples during sample collection or preparation. A commonly encountered example is that plasma samples are collected into tubes coated with anticoagulants. It has been reported that different types or concentrations of anticoagulants can affect the results of metabolic profiling (Barri & Dragsted, 2013; Chen et al., 2017). In a large-scale metabolomics study of blood samples, if different types or batches of collection tubes are used, there could be interbatch variability arising from matrix effects. This situation is likely to happen when samples are collected in multiple hospitals or from multiple blood banks to increase the sample size. Similarly, phosphate-buffered saline (PBS), which is widely used as the diluent or washing solution in tissue and cell metabolomics (Gonzalez-Riano et al., 2016; Ser et al., 2015), also causes concentration-dependent matrix effects to MS detection (Han & Li, 2015). When the use of PBS is not strictly

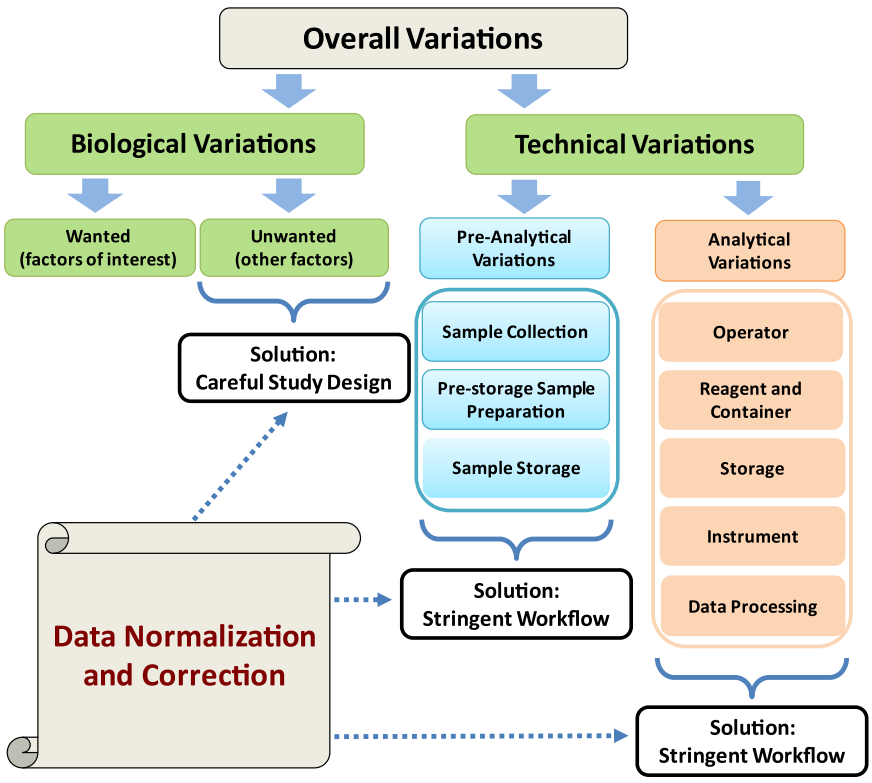


FIGURE 2 Graphical representation of sources of variations in quantitative metabolomics and typical strategies to minimize the unwanted variations [Color figure can be viewed at wileyonlinelibrary.com]

standardized, samples processed in different batches may experience batch-wise biases.

Sample storage condition, although sometimes overlooked, is another important contributor to interbatch variations, as conversion or degradation of metabolites may slowly happen (Álvarez-Sánchez et al., 2010; Teahan et al., 2006). Samples stored under nonidentical conditions (e.g., different temperatures or different time lengths) may have systematic differences in metabolome profile (Maher et al., 2007; H. Zhou et al., 2006). Furthermore, freeze-thaw cycles (FTCs) have an even stronger influence on sample integrity (Hirayama et al., 2015; Sykes, 2007). Unless all batches have experienced the same number of FTCs, even small changes in metabolite concentrations from each cycle can lead to a considerable number of false-positive discoveries (Chen et al., 2020a).

A major part of interbatch variations comes from the instrumental analysis step. In MS detection, sensitivity drift over time and across batches is a significant source of signal variability (Fernández-Albert et al., 2014; Shen et al., 2016). As metabolite quantification relies on the intensities of the MS peaks, such a drift will lead to varying measurement results even though the true values are the same in different batches.

Also, when chromatography is coupled to MS, both GC-MS and LC-MS face the problems of sample carryover and contamination building up, which may differ across batches (Burton et al., 2008). It is interesting to note that batch variations and correction methods used are likely different for GC-MS and LC-MS. For example, in a study examining various methods for correcting GC-MS batch-effect on mouse serum analysis, it was found that the use of total-signal-intensity for signal normalization is better than the isotope-standard method (Zaitsu et al., 2019). In LC-MS, the isotope-standard method would perform better in general as it corrects for matrix effect. In GC-MS, the matrix effect on metabolome analysis is less than that in LC-MS. It is not surprising that the isotope-standard method did not effectively correct the batch effect.

Possibly because NMR-based metabolomics requires minimal pretreatment of biological samples (Markley et al., 2017), which means it is less susceptible to analytical variations, in the literature, there are much fewer discussions about batch effects in NMR-based metabolomics than that of MS-based metabolomics. We note that if the instrumental variability is significant in NMR measurement and thus corrections are needed, many of the computational approaches that remove batch effects from MS data should also work for NMR data, as the formats of results are similar (i.e., concentrations of a set of metabolites across multiple samples). Hence, this review will focus on MS-based metabolomics hereinafter.

The instrumental changes mentioned above, especially the intensity drift, can also happen gradually within the data acquisition of a single batch of samples, causing within-batch variations. When caused by the same technical factors, interbatch and within-batch effects may have no clear boundaries in between. The major difference can be that within-batch variations often follow a more continuous and monotonic pattern than interbatch differences (Deng et al., 2019). Within-batch variations may also affect the statistical analysis, and it is often mixed with interbatch differences. Although this review aims to summarize reported approaches for minimizing interbatch effects, some of the methods should also be applicable to perform within-batch corrections.

3 | DETECTION AND EVALUATION OF BATCH EFFECTS

In general, a three-step workflow is used to handle unwanted variations: (1) identifying the unwanted variations, (2) removing or accommodating the unwanted variations in statistical analysis, and (3) evaluating the performance of batch effect removal (Livera et al., 2015). To handle batch effects, the first step is particularly important, because some batch effect removal methods should only be conducted when necessary (Sánchez-Illana et al., 2018), and choosing the best correction approach depends on the sources and patterns of the interbatch differences.

Many existing batch effect correction approaches are developed to generate a batch-effect-free dataset by excluding affected metabolites or adjusting the measured values of each metabolite (Salerno et al., 2017; Wang et al., 2012). As these adjustments extensively modify the original data, batch effect removal may lead to unpredictable outcomes to the results (e.g., underestimated or overestimated biological differences) and should be conducted with caution (Nygaard et al., 2016). Particularly, when there is no significant batch effect that would interfere with the biological variations, batch effect removal, which takes extra effort and time, becomes unnecessary. Hence, instead of applying batch effect removal to all datasets indistinguishably, we should first assess the relative severity of batch effects compared to the strength of biological effects.

In some studies where isotope-labeled internal standards are used, batch-wise variations can be straightforwardly visualized by plotting the intensities of internal standards against the injection order or batch number (Kuligowski et al., 2014). However, for more complicated untargeted metabolomic profiling, a more in-depth

analysis of the dataset is required. Here, we review some more commonly used computational strategies for detecting and evaluating batch effects. For most of them, information on the injection order or batch label is needed.

3.1 | Principal component analysis (PCA)

PCA is an unbiased dimensionality reduction method that has been widely used in metabolomics (Worley & Powers, 2013). It finds the principal components (PCs) through a linear combination of a set of variables, with the resulting PCs ranked by the percentage of total variance that each can explain. The PCA score plot projects all the samples onto the surface of PC 1 and PC 2, providing a simple and graphical overview of the data without any extra assumptions.

Considering that the consequences of batch effects are often complicated and even mixed with within-batch variations, we prepared a simulated dataset based on real serum metabolomics data, to demonstrate and explain the batch effects in a simple way. Figure 3A shows the score plot of this simulated dataset without batch effects, including samples of phenotype A (in orange), samples of phenotype B (in blue), and quality control (QC) samples (in green). Because QCs are experimental replicates of the same sample, the average distance between them represents the analytical variability. Phenotype A is clearly separated from phenotype B, indicating that there are significant metabolic differences between the two phenotypes. The distance between the two groups represents the interphenotype variability, which we are most interested in.

Figure 3B shows the analysis of the same set of samples, but during the second half of instrumental analysis, the measurement experienced a 20% intensity drop. Strong batch effects make the between-batch variability predominant on the score plot, as the two batches are clearly separated on PC 1. In this example, the PCA score plot directly displays the clustering of batches to prove the existence of batch effects.

We note that in real-world data, the newly emerging nonlinear regression models may better fit the complicated changing patterns of batch effects mixed from multiple sources. However, unlike searching for biomarkers, we will not care too much about the details of interbatch variations when their contribution to the overall variation is minor. As an easy-to-use unsupervised analysis, PCA can best serve the needs of evaluating the severity of batch effects over biological variations without any bias.

3.2 | Guided PCA

Usually, the PCA approach only examines the first few PCs. When the interbatch variability is not one of the largest sources of variance, PCA is not able to identify the batch effects (Benito et al., 2004). Although the batch effects are relatively insignificant when compared with the overall variance, it may still be strong enough to interfere with the interphenotype variability, especially when the latter is not very large (Chen et al., 2020a). Guide PCA is designed as an extension of the traditional PCA to deal with this situation (Reese et al., 2013).

Briefly, batch information is included in the PCA modeling so that the resulting top PCs will prefer explaining interbatch variability. Afterward, a δ value, which is the ratio of PC 1 from guided PCA to that from PCA, is calculated to evaluate the severity of batch variations. As a numeric value between 0 and 1, the larger δ is, the stronger the batch effects are. A permutation test is performed to determine the statistical significance of δ, and an empirical p-value is given to the users. For the data set in Figure 3, when there is no significant batch effect, δ is only 0.244 ($p < 0.718$), but when batch effects occur, δ dramatically increases to 0.900 ($p < 0.001$), indicating the existence of batch effects.

3.3 | Hierarchical clustering analysis (HCA)

HCA partitions the samples into homogenous groups according to their similarity (Johnson, 1967), and visualizes the hierarchy of samples by a dendrogram. If batch effects exist, samples will tend to cluster by batch label (Leek et al., 2010). For observing batch effects, HCA is less quantitative and more susceptible to other variances than the other methods.

3.4 | QC-based analysis

The use of QC samples is a powerful means of assuring high-quality data and has become a routine practice in metabolomics (Godzien et al., 2015). In most cases, QCs are prepared by either spiking metabolites with known concentrations into blank matrix samples or pooling aliquots from some or all of the individual samples (Sangster et al., 2006). During GC-MS or LC-MS analysis, QCs are injected and analyzed at intermittent points throughout the entire sequence of running samples. Theoretically, all QCs should have the same measured concentration of every single metabolite, as the true concentration and sample matrix are identical among

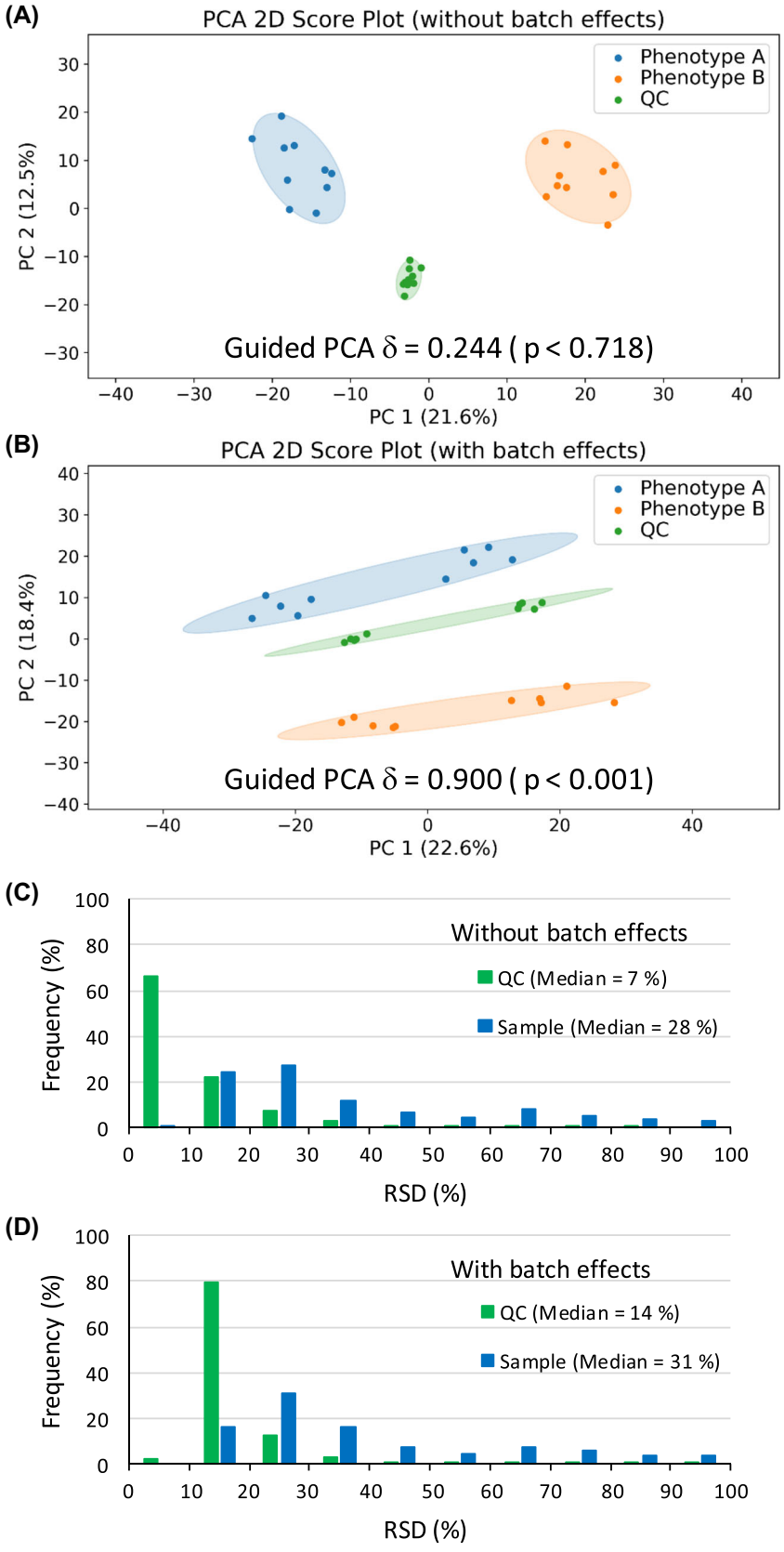


FIGURE 3 (A) Principal component analysis (PCA) score plot showing the separation between two phenotype groups in a data set without significant batch effects. The δ value from guided PCA is also provided. (B) PCA score plot and guided PCA result showing the separation between the same biological samples in (A) when strong batch effects exist. (C) Histogram showing the distributions of metabolite quantification relative standard deviations (RSDs) among quality controls (QCs) (green) and samples (blue) for the batch-effect-free data set in (A). (D) Histogram showing the distributions of metabolite quantification RSDs among QCs (green) and samples (blue) for the batch-effect-affected data set in (B). The data set is simulated by adding a simple and clear batch effect pattern to real data of serum metabolomics [Color figure can be viewed at wileyonlinelibrary.com]

QCs. The differences among QCs, which are usually minimal, reflect the analytical variability induced during sample processing or data acquisition (Burton et al., 2008). When interbatch variability becomes considerable, the inter-QC variations also increase.

The relative standard deviation (RSD) of each metabolite among the QCs is a commonly used indicator of inter-QC variations. For absolute quantification, the United States Food and Drug Administration has issued a Bioanalytical Method Validation Guidance for Industry, requiring the RSD among replicates in bioanalytical methods to be below 15%. When the analyte's concentration is at the limit of quantification, the guidelines allow variations at 20%, instead of 15%. In untargeted metabolomics analysis, as it is challenging to achieve complete absolute quantification of all metabolites, there are no regulatory guidelines governing accuracy and precision. A typically accepted RSD threshold in biomarker discovery studies is 30% (Wang et al., 2012; Zhao et al., 2016). Because untargeted analysis detects many metabolites at the signals of above the detection limit and below the quantification limit, relaxing RSD values to 30% seems to be reasonable. Moreover, in untargeted metabolomics, it is a common practice to exclude a metabolite when the metabolite's RSD in QCs is larger than a specific threshold (e.g., 20%) (Vinaixa et al., 2012). During this process, batch effects can lead to a problem that many informative metabolites, which could become biomarker candidates, are considered to be quantitatively unreliable and then wrongly removed.

Figures 3C and 3D are the distributions of RSD values of all metabolites in the two datasets corresponding to Figures 3A and 3B. The green histogram represents the QCs and the blue histogram is for the samples. With no batch effects, the median RSD among QCs is 7%, showing excellent analytical repeatability. The median RSD of samples is 28%, which means biological variability is clearly greater than analytical variability. When there are batch effects, the median QC RSD significantly increases to 14%, suggesting much larger analytical variations. With the median RSD of samples only slightly increased to 31%, the analytical variability becomes more significant compared with biological variability.

Because large inter-QC differences do not directly indicate the existence of batch effects, we may first group the QCs into batches and calculate the median RSD for each batch, then examine if the median RSD remarkably goes up when the batches are combined (Kirwan et al., 2013). Also, as RSD is susceptible to outliers, sometimes the ratio of interquartile range to the median is used as a more outlier-resistant alternative (Myers et al., 2012).

With known batch labels, analysis of variance (ANOVA) can also be used to test if the mean value of a QC metabolite in one batch is statistically different from those in other batches (Gregori et al., 2012). ANOVA tells how many QC metabolites are potentially affected by interbatch variations and generates a list of them. The main limitation of QC-based assessment is that a relatively large number of QCs are required in each batch to accurately reflect the batch effects. Although QCs are commonly used in metabolomics, for some small-scale studies, a user may choose to run no more than five QCs in each batch.

4 | CORRECTION OF BATCH EFFECTS

When the existence of strong batch effects is confirmed, actions are required to avoid confounding effects in data analysis. Various strategies have been proposed to remove or minimize batch effects. Some of them adjust each metabolite independently, whereas others implement multivariate procedures. Although an adequate number of ISs or QCs can facilitate many advanced correction strategies, methods solely relying on the sample data may also generate satisfying performance. Here, we summarize these experimental or computational strategies that can be used to improve the data quality for quantitative metabolomics. (Table 1). Table 2 highlights the performance comparison of these methods.

4.1 | Sample data-based normalization

Sample data-based normalization is the most extensively used data correction strategy in metabolomics. As the data-driven methods do not require any additional experimental elements, such as ISs or QCs, they help with lowering the cost and complexity of the overall study. In particular, when ISs and QCs are unavailable or in limited numbers, these methods are indispensable. They are also complementary with the IS- or QC-based approaches.

Postacquisition sample normalization has been a routine practice in metabolomics. Although the main purpose is often to offset the differences in total metabolite concentration among samples (Y. Wu & Li, 2016), many statistical normalizations can also correct the batch effects, especially those caused by sample pipetting or extraction.

The simplest approach is the median normalization, which adjusts the data to make all samples have the same

median concentration (represented by MS peak intensity or peak area), which is normalized to 1 in most cases (Atwal et al., 2015). Similarly, Unit-norm scales the data of each sample to make the sum of all metabolite concentrations equal to 1 (Scholz et al., 2004). Probabilistic quotient normalization, on the contrary, assumes that the detected intensities of most metabolites are affected by dilution only, and normalizes samples to a reference sample by probable quotients (Di Guida et al., 2016). Similarly, quantile normalization is another sample-wise adjustment making all the samples have the same distribution (Brodsky et al., 2010). The statistical normalizations lie in the assumption that the distributions of metabolite intensities are similar among samples from multiple batches (Deng et al., 2019). Although this assumption is not always true, these simple normalizations can still be used as an additional correction during data processing.

Matrix factorization methods, on the contrary, decompose the data matrix into mutually orthogonal submatrices associated with different effects and then remove the batch-dependent components. Two-way ANOVA has been a popular choice to decompose the data matrix to submatrices according to phenotypes and batch labels so that the interphenotype variability and interbatch variability are separated (Boccard et al., 2019). Similarly, the EigenMS tool first preserves the phenotype-related variations using ANOVA, and then applies singular value decomposition to identify and remove batch effects from the residual (Karpievitch et al., 2014). These methods may face difficulties in handling data with missing values or missing samples, and a large sample size is preferred for optimal performance.

To deal with the limitations of matrix factorization and lower the risk of removing real biological changes, ComBat (Johnson et al., 2007), which is widely used in genomics, has also been implemented to adjust metabolomics data. ComBat corrects the data based on an empirical Bayesian framework, and it is good at adjusting batches with small sample sizes. Although no extra experiments are required, a well-balanced batch-group design is crucial for both ANOVA and Combat. Nygaard et al. (2016) have thoroughly discussed this issue and pointed out that using ComBat or ANOVA to adjust an unbalanced dataset may deflate or inflate the differences between phenotypes.

WaveICA, a more recently developed method, adopts wavelet transform and independent component analysis (ICA) to decompose the data matrix (Deng et al., 2019). In a well-randomized analysis sequence, the switching between phenotypes through the analysis time is very frequent, making it possible to use time-scale frequency to isolate the slower-changing interbatch differences from interphenotype variations. Figure 4 shows the intensity of a selected metabolite against injection order and batch in the original data and in those processed with four different batch effect correction methods. Although ComBat effectively adjusted the interbatch differences, it is not designed to handle the within-batch drift. Despite the requirement of an ideal block design during the LC-MS analysis, Deng et al. (2019) demonstrated that WaveICA had better correction performance than ComBat (Table 1).

4.2 | IS-based methods

Isotope-labeled ISs, which are widely used in targeted analysis, can effectively cope with analytical variations during the analysis. Spiked into the samples right after sample collection, ISs and analytes experience the same sample handling and instrumental analysis steps. After data acquisition, the concentrations of each metabolite are normalized by taking the ratio of its intensity to that of the corresponding IS.

In the work of Bromke et al. (2015), a ^{13}C-labeled IS was used to normalize all the detected metabolites. Theoretically, when the MS peak intensity of a metabolite is enhanced or reduced by batch-related matrix effects or instrumental drift, the same effects happen to the IS. Consequently, the absolute intensities of the metabolite and the IS may vary, but the relative concentration to the IS remains constant. As the ISs are spiked into the samples, they experience all following technical variations together with the analytes, so all unwanted effects induced after adding the ISs can be monitored and corrected. Unlike the QCs, which cannot gauge the variability during sample collection, the ISs are added to the samples at the very beginning of the study. Thus, all interbatch variations can be captured. Furthermore, the IS approach has the flexibility of evaluating and normalizing the variations of each experimental step separately. In the microbial metabolome analysis platform proposed by van der Werf et al., ISs were added at each experimental stage to correct the analytical variations arising from that stage (van der Werf et al., 2007).

We note that although adding ISs right after sample collection is the ideal approach to monitor variations induced during sample handling, it is practically difficult as biological samples are often collected by health professionals rather than researchers. Hence, we believe spiking in ISs after the first freeze-thaw cycle, which would be more achievable, is the second-best approach and should be recommended in metabolomics. The downside of this second-best strategy is that any unwanted variations caused by the first FTC cannot be

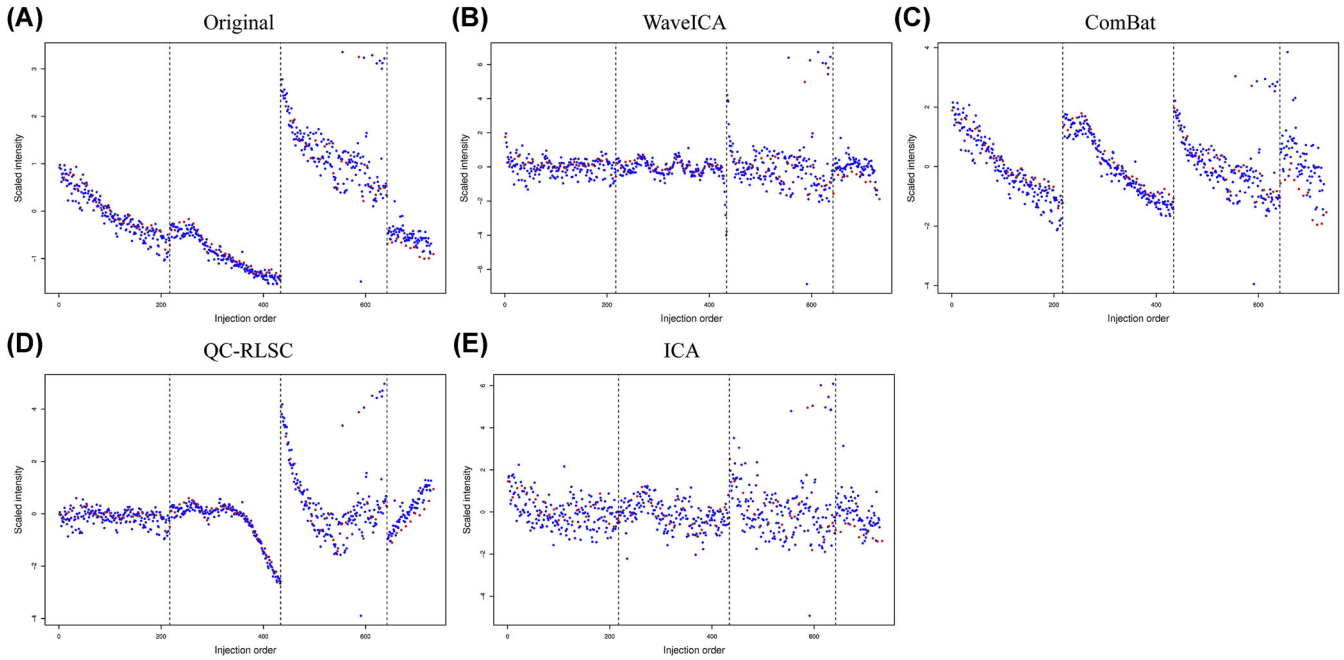


FIGURE 4 Intensity of a selected metabolite changing with injection order and batch, (A) before and after applying (B) WaveICA, (C) combatting batch effects (ComBat), (D) quality control-based robust locally estimated scatterplots moothing signal correction (QC-RLSC), and (E) independent component analysis (ICA). Red points represent QCs and blue points represent samples (Adapted with permission from Deng et al., 2019) [Color figure can be viewed at wileyonlinelibrary.com]

corrected. Studies to understand the FTC effects with a hope to minimize or eliminate the FTC effects are still ongoing. On the basis of our current knowledge, if all samples have experienced only one FTC, the impact of FTC effects on biomarker discovery will be limited (Chen et al., 2020a).

Using a single IS is a simple and economical approach, but it relies on the dubious assumption that all metabolites would respond to batch-related interferences in the same way as the IS does. With the highly diverse physical and chemical properties of metabolites, this assumption is not very likely to be true. To overcome this challenge, more studies have included multiple ISs, with one or several standards representing a class of metabolites (S. Yang et al., 2010; Zukunft et al., 2013). Regardless, having multiple ISs is still not a perfect solution. First, as it cannot be guaranteed that all metabolites in the same class behave similarly during the analysis, the number of ISs should be as large as possible. Ideally, every single metabolite should have its isotope-labeled IS. However, an untargeted metabolomics study typically can detect more than a thousand metabolites, with many unidentified. Hence, it is not practically possible to have ISs for all metabolites. For most laboratories, the cost of acquiring or making even hundreds of isotope-labeled standards is beyond the reach (Ejigu et al., 2013). Second, even if a metabolite has an isotope-labeled counterpart being analyzed together in LC-MS, the two forms do not encounter the same instrumental variability when they do not coelute, which is very common for deuterium-labeled standards (Stokvis et al., 2005).

To deal with the challenge that representative ISs cannot perfectly reflect the behaviors of all metabolites, Sysi-Aho et al. (2007) developed a computational method called normalization using optimal selection of multiple internal standards (NOMIS). For each metabolite, the changing patterns of multiple ISs are examined and used together to calculate an optimal normalization factor. The authors reported that NOMIS outperformed direct correction using ISs for three metabolites. The limitation of this strategy is that when the mathematical assumptions are violated in a real dataset, some biological variations of interest may also be removed (de Livera et al., 2012). Redestig et al. (2009) further considered the influences of analytes on ISs and proposed a cross-contribution-compensating multiple-standard normalization approach to correct the cross contributions.

More recently, Boysen et al. (2018) developed the best-matched internal standard normalization approach. In this method, QC samples containing a collection of isotope-labeled ISs are repeatedly injected into the LC-MS system throughout the analysis. When the QC RSD of a metabolite is larger than 10%, the metabolite is normalized by each IS candidate and the RSDs after normalization are calculated. Finally, the IS generating the smallest RSD is chosen for the normalization of the

specific metabolite. With a carefully chosen set of ISs, this strategy may overcome the challenges of IS-based normalization. The remaining challenge is that the correction only works for the metabolites detected in QC samples.

4.3 | QC-based correction

With QC samples analyzed intermittently throughout the instrumental analysis, it becomes easier to monitor the change of instrumental performance. For example, QCs can reflect the gradual change of instrument sensitivity, which is extremely useful for correcting within-batch variations. Furthermore, QCs provide a quantitative criterion for assessing the performance of batch effect removal, that is, the QC RSDs should become smaller.

Similar to the statistical normalizations, a simple and straightforward way of QC-based batch effect correction is to build a regression model between the total signal of each QC and the injection order or batch number (Wang et al., 2012). Regardless, considering that different metabolites may have different responses to batch effects, a more popular strategy is to study the QC metabolites separately. In other words, for each metabolite existing in QCs, we can find a mathematical pattern describing its concentration change as a function of injection order or batch number. Because samples are "bracketed" in between of QCs, we assume that the changing patterns of QCs also apply to the samples. When the batch information of samples is substituted into the model, we can acquire the predicted batch variations. Finally, the predicted batch variability is subtracted from the original data to generate a batch-effect-free dataset for statistical analysis.

In fact, a large variety of regression-based methods have been established to correct batch effects. A qualified QC-based correction should not only accurately fit the complex inter- and within-batch variations, but also be resistant to overfitting. To achieve this goal, although linear least-squares (LS)-regression-based correction can significantly improve the quality of data with better performance than other traditional data-based normalization methods (Wang et al., 2012; Wehrens et al., 2016), nonlinear or nonparametric models are more frequently adopted due to the complexity of batch-related changing patterns.

A LOESS (locally estimated scatterplot smoothing) algorithm, which is also known as moving regression, is often used for curve smoothing, and the corresponding correction method is named QC-based robust LOESS signal correction (QC-RLSC) (Dunn et al., 2011). Alternatively, cubic smoothing spline algorithms are also popular choices of curve fitting for batch effect removal, with the method termed QC-based robust spline correction (QC-RSC) (Brunius et al., 2016; Kirwan et al., 2013). Thonusin et al. (2017) reported that QC-RLSC, QC-RSC, and a QC-based LS quadratic regression performed equally well to correct their dataset.

In addition to curve fitting, various nonparametric approaches have been developed to fit the interbatch changes. On the basis of the support vector regression (SVR), which projects the data to a higher-dimensional space and then builds a linear model there, QC-SVR correction tools (QC-SVRC) have been developed (Kuligowski et al., 2015; Sanchez-Illana et al., 2018). Shen et al. (2016) compared QC-SVRC with QC-RLSC, and reported that although both methods significantly reduced the QC RSDs, the prediction ability of QC-RLSC was relatively poorer and there was a high risk of overfitting. This observation is consistent with Figure 4 where the performance of QC-RLSC was not as good as WaveICA for correcting interbatch or intrabatch variations. On the contrary, Rong et al. (2020) reported that their NormAE algorithm, which was built on deep neural networks, demonstrated stronger correction power than WaveICA. Furthermore, Luan et al. (2018) processed a dataset using QC-RLSC, QC-SVRC, and a random forest-based method (QC-RFSC), and the results showed that QC-RFSC outperformed the other two corrections.

Notably, the correction of batch effects is not limited to only one method at a time. Rodríguez-Coira et al. (2019) implemented a combination of two algorithms, QC-SVRC followed by normalizing the shift of QC median, to achieve the optimal correction performance for their serum metabolomics data. Furthermore, as we suggested earlier, a data-driven correction can be used together with a QC-based method to maximize the performance.

QC-based methods provide an efficient and low-cost way to remove interbatch variations when compared with IS-based approaches. As metabolomics studies usually include running QC samples, using QCs should not take much extra time and effort. Having plenty of QCs is crucial for maximizing the performance of the correction and lowering the risk of overfitting (Rong et al., 2020). However, injecting too many QCs will significantly extend the analysis time, which means more noticeable instrumental drift is likely to happen (Kuligowski et al., 2015; Wehrens et al., 2016). Also, using a pooled sample as QC is highly preferred for these correction methods to cover all metabolites detected among the samples, but when the available amount of each individual sample is limited, it becomes difficult to make a large amount of QC. At last, as stated previously, QC-based normalization cannot cover the interbatch

differences before QC samples are made (e.g., the variations during sample collection).

4.4 | Selected QC metabolite (QCM)-based correction

QCMs are the metabolites existing in biological samples but assumed to be unrelated to the phenotypes according to the known biological background (de Livera et al., 2012; Livera et al., 2015). The algorithm of the QCM-based correction methods assumes that QCMs are not changing across samples, which means their variations are all irrelevant. Therefore, QCM-based correction identifies not only batch effects but also other sources of unwanted variations. The limitation of this approach is that the performance is highly dependent on the selection of QCMs. In a real-world metabolomics study where the target metabolome is not fully understood yet, the selection of these negative controls may not be very easy (Salerno et al., 2017).

4.5 | Chemical isotope labeling-based correction

As discussed earlier, although QC-based correction methods have been greatly enriched and improved, identifying batch effects in each individual sample is mainly based on prediction, which is less accurate than a true experimental reference. Despite high costs and limited availability of standards, IS-based strategies have their irreplaceable advantages. Researchers have spent lots of effort to expand the availability of ISs. For example, using isotope-enriched substrates, we can culture a fully isotope-labeled cell extract, which contains a complete isotope-labeled metabolome that can be used as the reference (Weindl et al., 2015; L. Wu et al., 2005). However, this approach is not practically possible for analyzing samples that cannot be cultured in isotope-enriched media, such as biofluids, human tissues, and many other types of samples.

Alternatively, a reference sample, such as a pooled sample, can be isotope-labeled after sample collection. The chemical isotope labeling (CIL) method (Guo & Li, 2009), which is mainly focusing on LC-MS applications, realizes relative quantification through a chemical derivatization reaction with isotopic labeling reagents. More specifically, a tag group is attached to every single metabolite in an individual sample, and in the meanwhile, the isotopic counterpart of the labeling group (e.g., a ^{13}C-labeled reagent) is attached to the same metabolite in a reference sample. After the labeled samples are mixed, the metabolite from the reference sample serves as the internal standard.

The CIL method is very different from the IS method, although they share the same keyword, "isotope." The major advantage of CIL LC-MS is the significant expansion of the metabolome coverage through improved LC separation and enhanced MS detection by using rationally designed reagents for labeling metabolites. At the same time, every single metabolite, detected in the QCs or not, has its isotope-labeled IS to overcome analytical variations. After the individual sample is mixed with the heavy-isotope-labeled reference sample, the pair of labeled metabolites will experience the same analytical variations, and thus the relative quantification will not be affected anymore. Figure 5 shows an example of a differentially chemical-isotope-labeled metabolite detected in two batches (some of the runs in batch 1 had encountered a small leak in LC) (Chen et al., 2020b). The absolute intensities of peaks shown in the mass spectra were changed, whereas the peak ratio remained to be similar. Unlike the computational batch correction strategies suffering from artificial biases or overfitting, CIL LC-MS provides a complete set of ISs so that the quality of adjustment is guaranteed.

Traditionally, the reference sample in CIL LC-MS is prepared by pooling all the individual samples of a study. However, when the sample collection and analysis are done in batches, it may be inconvenient to generate the pooled sample from individual samples in all batches. In some cases, the available amount of each individual sample is very limited and thus taking an aliquot from each sample for pooling may not be feasible. To facilitate the application of CIL LC-MS to large-scale metabolomics, Peng et al. (2014) proposed a universal metabolome standard (UMS) method. Available in large amounts, the UMS is a pooled sample of a specific sample type. The UMS can serve as the reference sample for multiple batches and even multiple studies. Using the UMS-based CIL LC-MS approach, the authors analyzed multiple batches of urine samples and achieved excellent reproducibility.

It should be noted that the CIL LC-MS platform alone can only offset analytical variability induced after the labeled individual samples are mixed with the labeled reference sample. Therefore, a stringent experimental workflow is needed to avoid batch effects during sample handling. CIL-based preacquisition normalization methods have been developed to deal with systematic variations during sample collection and treatment (e.g., dilution effects and pipetting errors) (Y. Wu & Li, 2012). Still, the user should make sure that there are no strong matrices or contaminations introduced to some of the batches (Han & Li, 2015).

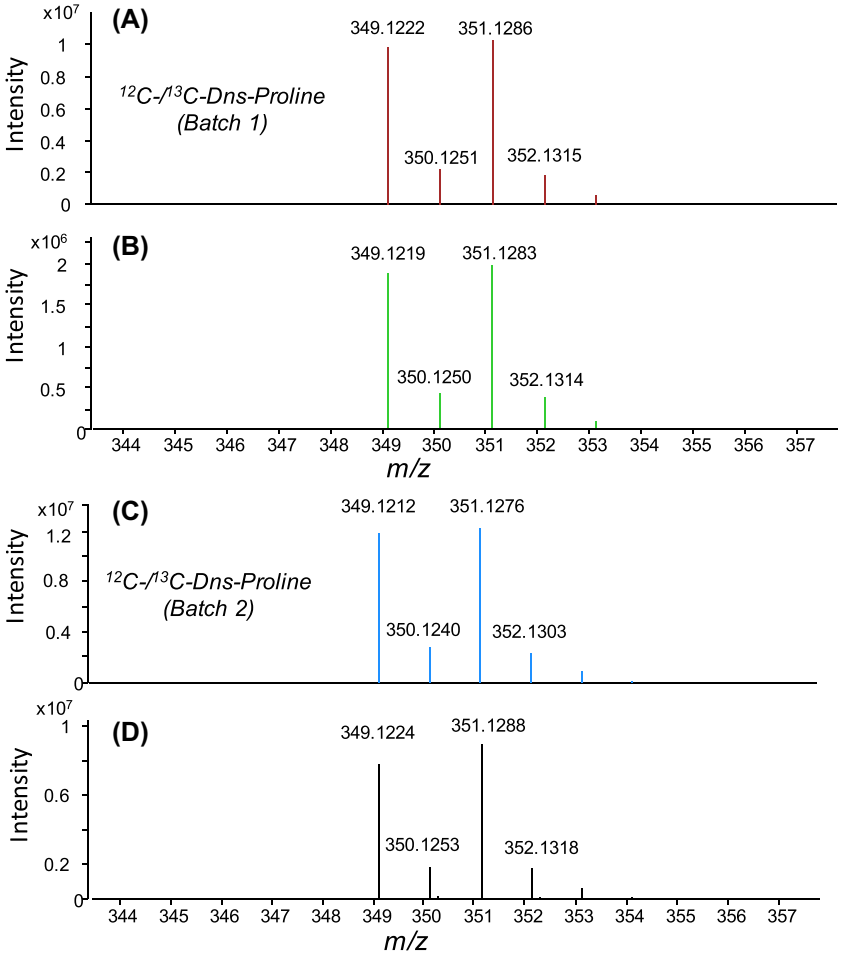


FIGURE 5 Expanded mass spectra of dansyl labeled proline detected from (A) a quality control (QC) rat plasma sample in Batch 1, (B) a different QC sample in Batch 1, (C) a QC sample in Batch 2, and (D) a different QC sample in Batch 2. The QC sample was prepared by mixing an equal mole amount of the ^{12}C-labeled and ^{13}C-labeled pooled samples and injected into liquid chromatography-mass spectrometry after every 10 sample runs. A total of 468 rat plasma samples were analyzed over a period of 15 days in three batches (Adapted with permission from Chen et al. 2020b) [Color figure can be viewed at wileyonlinelibrary.com]

5 | EVALUATION OF BATCH EFFECT REMOVAL

The last step of the batch effect handling workflow is to evaluate the performance of batch effect removal. Theoretically, all the batch effect detection methods discussed in the previous section can be applied for evaluation. For example, PCA plots before and after correction are often compared to visually demonstrate the improvement of clustering (Deng et al., 2019; Rong et al., 2020). However, visually improved clustering may not be quantitative enough to assess the difference in an objective manner (Čuklina et al., 2020). To quantitatively evaluate the PCA result, Goodpaster and Kennedy (2011) proposed a stick plot showing the Euclidean distance between each sample point and the centroid of the phenotype group that it belongs to.

Figures 6A and 6B show this type of stick plots for the datasets without and with batch effects, as discussed in Figure 3. In these plots, the number below the group label is the mean distance between samples in that group and its centroid, and the horizontal line represents the distance between the group's centroid to that of the baseline group (in this case, the QCs). Additionally, each vertical stick, centered at the horizontal line, reflects the distance between an individual sample and the centroid of its phenotype group. In this way, we can quantitatively observe that batch effects can remarkably increase the within-group variance, compared to the intergroup variance. Especially when QCs are not available, PCA has been proved to be a robust method for evaluating the effectiveness of batch effect removal, and as discussed previously, guided PCA can serve as another powerful tool to quantitatively measure the PCA results.

If there is a large number of QCs available throughout the study, QC-based information is often examined to check the elimination of batch effects. Specifically, in PCA analysis, the mean distance (Shaham et al., 2017) or median distance (Shen et al., 2016) between QCs is frequently used to evaluate analytical variability. As shown in Figures 6A and 6B, compared to the dataset with significant batch effects, whose QC mean distance is 13.69, the batch-effect-free dataset has a much better clustering of QCs with a mean distance of 1.81. Similarly, the Silhouette plot, which comprehensively assesses the clustering in a quantitative manner, is also a useful tool to evaluate the correction (Sánchez-Illana et al., 2018).

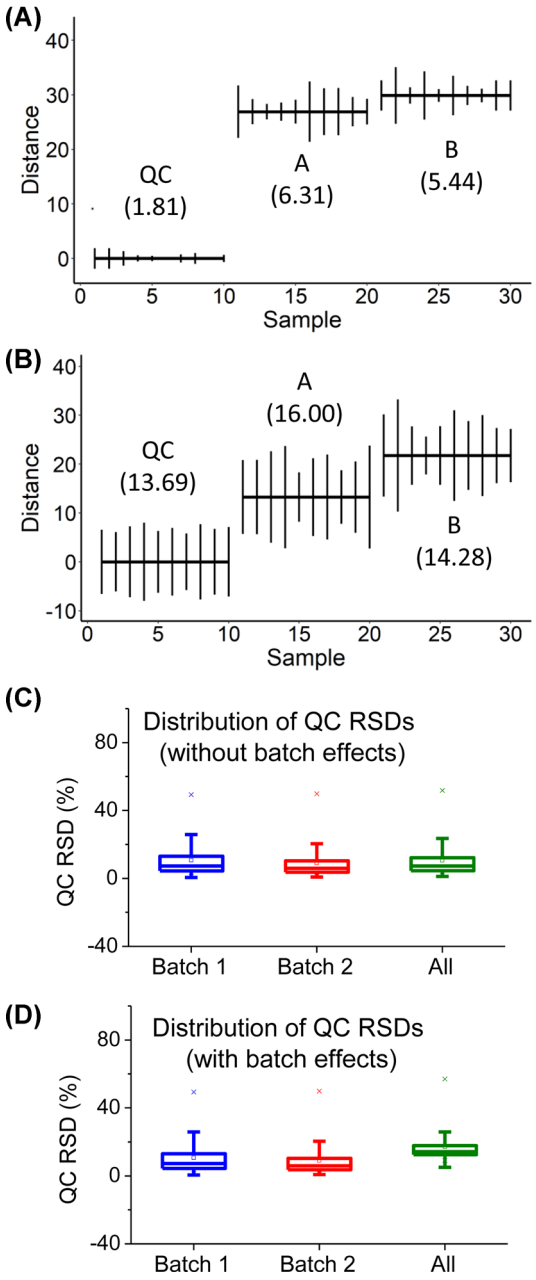


FIGURE 6 (A) Stick plot showing the intergroup and within-group distances for the principal component analysis (PCA) plot in Figure 3A. (B) Stick plot showing the intergroup and within-group distances for the PCA plot in Figure 3B. (C) Box plot showing the distribution of quality control-relative standard deviations (QC RSDs) in Batches 1 and 2 for the batch-effect-free data set in Figure 3A. (D) Box plot showing the distribution of QC RSDs in Batches 1 and 2 for the batch-effect-affected data set in Figure 3B [Color figure can be viewed at wileyonlinelibrary.com]

Additionally, as discussed in the batch effect detection section, we may group the QCs by batches and examine the RSDs. Figures 6C and 6D are the box plots showing the distributions of QC RSDs in the two batches, as well as among all QCs, without and with batch effects, respectively. In a dataset with strong batch effects, as shown in Figure 6D, the median of QC RSDs among all QCs is significantly larger than that of a single batch. However, in Figure 6C, the two batches have similar QC RSD distributions, and there is no noticeable change when we examine them together.

A remaining problem with the QC-based evaluation is that when overfitting happens, the correction may work well with QCs but perform poorly on other samples. To overcome this problem, Fan et al. (2019) proposed a fivefold cross-validated QC RSD (cvRSD) approach by randomly splitting QCs into training sets and testing sets.

In addition to assessing the removal of batch effects, some studies also mentioned the importance of evaluating the retention of biological information by examining the number of detected metabolite biomarkers and their discriminative powers (B. Li et al., 2017; Rong et al., 2020). Although this is an important aspect, there are practical difficulties for biomarker discovery studies where the true biomarkers are unknown and false positives may come from both batch effects and batch effect removal. Overall, PCA-based analyses, together with trends demonstrated by QCs, are sufficient to evaluate batch effect removal in most cases.

6 | MINIMIZING BATCH EFFECTS IN LARGE-SCALE METABOLOMICS

With more advanced analytical platforms and the need to study subtle metabolic perturbations, the typically required sample size of metabolomics studies has been increasing. Generally, the sample size used for a clinical study is dependent on the nature of the study and the complexity of the system. For example, the sample size requirement for discovering biomarkers of dementia would be very different from that for brain tumors using human tissues (e.g., brain tissues). Referring to large-scale metabolomics, we mean the sample size is sufficiently large for discovering and validating biomarkers for a clinical study. In real-world applications, the sample size is far larger than those used in the initial discovery projects, requiring analysis in batches and proper data processing. As discussed earlier, the discovery and validation of biomarkers involving multiple sites can be advantageous over a single site analysis. Moreover, with quantitative metabolome profiling becoming a mature analytical platform, data sharing among the metabolomics community is viewed as the next important step to accelerate the development of the field (Haug et al., 2012; Sud et al., 2015).

Ideally, the performance of a robust biomarker for an application such as diagnosis of a disease will not be significantly affected by batch effects, including variations in instruments and operators. However, small differences between data from batches are unavoidable in such large-scale projects. In the past, wrong conclusions from studies misled by analytical variations raised serious concerns about the application of omics techniques in clinical testing (Ransohoff, 2005).

Nonetheless, the discovery of a biomarker also depends on the extent of changes caused by the phenotype. If the changes are much greater than those from unwanted factors, one can always detect some significant changes when comparing the groups with and without this phenotype. Hence, when we analyze the result at each site independently and then merge the biomarker lists, which is also known as the meta-analysis, only the biomarker candidates with the most significant changes would be consistently recognized. Any metabolites with changes that can be influenced significantly by unwanted variations are not likely to be commonly detected at different sites. Therefore, with the use of proper data analysis workflow, the biomarkers commonly found in multisite studies would be more reliable.

Still, to discover a moderately robust biomarker or to monitor a specific metabolite of interest in a large-scale study, we need to minimize the extent of unwanted variations including batch effects. In addition to meta-analysis, further standardization of metabolomics protocols will make it possible to merge data of multiple studies on the same biological effects, which may lead to more profound findings. Therefore, minimizing the impacts of unwanted variations, including batch effects, is crucial for the future applications of quantitative metabolomics.

Given the various potential sources of batch effects, it is obvious that a careful study design, along with standardization of the experimental workflow, will help minimize interbatch variations. Despite that study design is a broad topic for revealing the true biological meanings effectively, here we emphasize planning for the consequences when samples are split into batches. For example, if samples are collected from blood banks, they should come from the same batch to avoid inconsistencies in sample collection practice (Long et al., 2020). Also, as mentioned in the previous section, all biological factors that could potentially be evaluated in the following analysis should be balanced among the batches. Furthermore, if operators at different sites follow the same stringent experimental protocol to handle samples and to operate instruments, the systematic differences will be minimized. Researchers have also promoted automated systems for sample handling to reduce unwanted experimental variations to the minimum (Long et al., 2020).

Instrumental drift over time and across batches may be more challenging to control. We can set up SOPs to lower the chance of the unexpected drift. For example, before the analysis begins, we can prepare sufficient volumes of the LC mobile phases, clean the LC and ionization source, and condition the computer (Rodríguez-Coira et al., 2019). Moreover, we recommend frequently monitoring the intensity of the ISs or QCs throughout the analysis so that the instrumental variability can be noticed in a timely manner. Even if considerable interbatch variations still exist after taking these precautions, the ISs or QCs can very well capture the pattern of instrumental variations and therefore, the computational correction methods are available for reducing the batch effects.

It should be noted that the data-based normalizations and QC-based regression methods have limitations, especially on the real-world data with complicated compositions and not well-balanced experimental designs. As Nygaard et al. (2016) suggested, rather than using these corrections to generate a batch-effect-free data set, it is better to take batch effects into account in statistical analysis. For example, Salerno et al. (2017) have applied an RRmix (random main effect and random compound-specific error variance with a mixture model) method to achieve simultaneous main effect detection and batch effect removal. RRmix excludes batch effects without prior knowledge about the nature of batch effects and detects the significant metabolites with improved sensitivity and specificity.

Despite the limitations and potential risks of batch effect removal, there is still a pressing need for having batch-effect-free data, as it becomes easier to incorporate other novel data analysis techniques, such as machine learning (Čuklina et al., 2020). Regardless, careful study design and stringent experimental protocols are the most important measures to minimize batch effects. Ideally, if the batch effects are well controlled such that they are much less significant than the main phenotype effects, there is no need for the correction. Furthermore, careful consideration of the batch effects during the experimental protocol design can at least simplify the sources and patterns of batch effects, which is highly preferred by the correction methods when batch effect removal is unavoidable.

The CIL LC-MS method, as an alternative experimental workflow of quantitative metabolomics, has the advantage of high-quality batch effect adjustment for all labeled metabolites. However, this method cannot cope with interbatch variations arising from sample collection or CIL reaction. To the best of the current knowledge, we believe that a combination of CIL and IS is the optimal

way to deal with the batch effect issue. IS is for controlling batch effects before the CIL labeling, and CIL is for minimizing batch effects in the following steps. If we cannot add in the ISs right after sample collection, the performance of the IS correction will be weakened. Still, it can monitor the extra FTCs and other sample handling steps. The limitation of CIL could be that in some field works where equipment is limited and quick results are wanted, performing CIL might be impractical or too costly.

Finally, computational implementation is an important step for the combination of IS and CIL, as well as other data-driven corrections. Fortunately, a number of open-source data normalization packages have been developed, and user-friendly graphical interfaces are also emerging (B. Li et al., 2017; Luan et al., 2018). We note that several MS vendors' software and data output formats have integrated data correction; however, if the source codes are not open, it may not be easy to implement other correction strategies. It would be difficult for laboratories with no strong IT support to choose the best option for batch effect correction. We hope, in the future, as the research community publishes more refined methods for batch effect correction, the vendors will adopt these methods in their software package to allow the usage of more and better options in dealing with different types of batch effects.

7 | CONCLUDING REMARKS

Minimizing interbatch variations is crucial in quantitative metabolomics that involves the analysis of many samples that are either collected in different batches or analyzed in batches or a combination of two. In this review, we discuss the origins and impacts of batch effects on metabolome analysis. We provide a summary of a number of batch effect removal methods mainly for MS-based metabolomics. Each method of dealing with interbatch variations has its own merits and limitations. When doing quantitative metabolomics, we should plan in advance to minimize batch variations. On the basis of the knowledge of potential sources of batch effects, multiple experimental steps, such as sample collection, storage, and thawing, should be carefully controlled. Analytical variations during sample preparation leading to instrumental analysis should be minimized. SOPs for running samples and analyzing the resultant data need to be strictly followed. Batch variations or batch effects should be evaluated after metabolome data acquisition and initial analysis. When the evaluation step reveals that batch effects are minor, the best strategy is not doing correction, but treating statistical results with caution. If batch effect removal is necessary, a combination of two or more batch removal methods may be used to achieve optimal and consistent results.

ACKNOWLEDGMENTS

This study was supported by the Natural Sciences and Engineering Research Council of Canada, Canada Research Chairs, Canada Foundation for Innovation, Genome Canada, and Alberta Innovates.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

REFERENCES

Álvarez-Sánchez B, Priego-Capote F, De Castro ML. 2010. Metabolomics analysis I. Selection of biological samples and practical aspects preceding sample preparation. Trends in Analytical Chemistry 29:111-119.

Atwal PS, Donti TR, Cardon AL, Bacino C, Sun Q, Emrick L, Sutton VR, Elsea SH. 2015. Aromatic L-amino acid decarboxylase deficiency diagnosed by clinical metabolomic profiling of plasma. Molecular Genetics and Metabolism 115: 91-94.

Baggerly KA, Edmonson SR, Morris JS, Coombes KR. 2004. High-resolution serum proteomic patterns for ovarian cancer detection. Endocrine-Related Cancer 11:583-584.

Barri T, Dragsted LO. 2013. UPLC-ESI-QTOF/MS and multivariate data analysis for blood plasma and serum metabolomics: effect of experimental artefacts and anticoagulant. Analytica Chimica Acta 768:118-128.

Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, Marron JS. 2004. Adjustment of systematic microarray data biases. Bioinformatics 20:105-114.

Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, Macdonald IA, Van Ommen B, Smilde AK. 2006. Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. Analytical Chemistry 78:567-574.

Boccard J, Tonoli D, Strajhar P, Jeanneret F, Odermatt A, Rudaz S. 2019. Removal of batch effects using stratified subsampling of metabolomic data for in vitro endocrine disruptors screening. Talanta 195:77-86.

Boysen AK, Heal KR, Carlson LT, Ingalls AE. 2018. Best-matched internal standard normalization in liquid chromatography–mass spectrometry metabolomics applied to environmental samples. Analytical Chemistry 90:1363-1369.

Brodsky L, Moussaieff A, Shahaf N, Aharoni A, Rogachev I. 2010. Evaluation of peak picking quality in LC-MS metabolomics data. Analytical Chemistry 82:9177-9187.

Bromke MA, Sabir JS, Alfassi FA, Hajarah NH, Kabli SA, Al-Malki AL, Ashworth MP, Méret M, Jansen RK, Willmitzer L. 2015. Metabolomic profiling of 13 diatom cultures and their adaptation to nitrate-limited growth conditions. PLOS One 10:e0138965.

Brunius C, Shi L, Landberg R. 2016. Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction. Metabolomics 12:173.

Burton L, Ivosev G, Tate S, Impey G, Wingate J, Bonner R. 2008. Instrumental and experimental effects in LC–MS-based metabolomics. Journal of Chromatography B 871:227-235.

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR. 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience 14:365-376.

Cajka T, Fiehn O. 2015. Toward merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics. Analytical chemistry 88:524-545.

Chen D, Han W, Su X, Li L, Li L. 2017. Overcoming sample matrix effect in quantitative blood metabolomics using chemical isotope labeling liquid chromatography-mass spectrometry. Analytical Chemistry 89:9424-9431.

Chen D, Han W, Tao H, Li L, Li L. 2020a. Effects of freeze-thaw cycles of blood samples on high-coverage quantitative metabolomics. Analytical Chemistry 92:9265-9272.

Chen D, Zhao S, Han W, Lo E, Su X, Li L, Li L. 2020b. High tolerance to instrument drifts by differential chemical isotope labeling LC-MS: A case study of the effect of LC leak in long-term sample runs on quantitative metabolome analysis. Journal of Mass Spectrometry e4589.

Čuklina J, Pedrioli PG, Aebersold R. 2020. Review of batch effects prevention, diagnostics, and correction approaches. *Mass spectrometry data analysis in proteomics*. Springer. pp. 373-387.

de Livera AM, Dias DA, de Souza D, Rupasinghe T, Pyke J, Tull D, Roessner U, McConville M, Speed TP. 2012. Normalizing and integrating metabolomics data. Analytical Chemistry 84: 10768-10776.

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

Di Guida R, Engel J, Allwood JW, Weber RJ, Jones MR, Sommer U, Viant MR, Dunn WB. 2016. Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. Metabolomics 12:93.

Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, Brown M, Knowles JD, Halsall A, Haselden JN. 2011. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nature Protocols 6:1060-1083.

Dunn WB, Lin W, Broadhurst D, Begley P, Brown M, Zelena E, Vaughan AA, Halsall A, Harding N, Knowles JD. 2015. Molecular phenotyping of a UK population: defining the human serum metabolome. Metabolomics 11:9-26.

Ejigu BA, Valkenborg D, Baggerman G, Vanaerschot M, Witters E, Dujardin J-C, Burzykowski T, Berg M. 2013. Evaluation of normalization methods to pave the way towards large-scale LC-MS-based metabolomics profiling experiments. Omics: A Journal of Integrative Biology 17:473-485.

Fan S, Kind T, Cajka T, Hazen SL, Tang WW, Kaddurah-Daouk R, Irvin MR, Arnett DK, Barupal DK, Fiehn O. 2019. Systematic error removal using random forest for normalizing large-scale untargeted lipidomics data. Analytical Chemistry 91: 3590-3596.

Fei F, Bowdish DM, McCarry BE. 2014. Comprehensive and simultaneous coverage of lipid and polar metabolites for endogenous cellular metabolomics using HILIC-TOF-MS. Analytical and Bioanalytical Chemistry 406:3723-3733.

Fernández-Albert F, Llorach R, Garcia-Aloy M, Ziyatdinov A, Andres-Lacueva C, Perera A. 2014. Intensity drift removal in LC/MS metabolomics by common variance compensation. Bioinformatics 30:2899-2905.

Fuhrer T, Zamboni N. 2015. High-throughput discovery metabolomics. Current Opinion in Biotechnology 31:73-78.

Godzien J, Alonso-Herranz V, Barbas C, Armitage EG. 2015. Controlling the quality of metabolomics data: new strategies to get the best out of the QC sample. Metabolomics 11: 518-528.

Goh WWB, Wang W, Wong L. 2017. Why batch effects matter in omics data, and how to avoid them. Trends in Biotechnology 35:498-507.

Gonzalez-Riano C, Garcia A, Barbas C. 2016. Metabolomics studies in brain tissue: a review. Journal of Pharmaceutical and Biomedical Analysis 130:141-168.

Goodpaster AM, Kennedy MA. 2011. Quantification and statistical significance analysis of group separation in NMR-based metabonomics studies. Chemometrics and Intelligent Laboratory Systems 109:162-170.

Goveia J, Pircher A, Conradi LC, Kalucka J, Lagani V, Dewerchin M, Eelen G, DeBerardinis RJ, Wilson ID, Carmeliet P. 2016. Meta-analysis of clinical metabolic profiling studies in cancer: challenges and opportunities. EMBO Molecular Medicine 8:1134-1142.

Gregori J, Villarreal L, Méndez O, Sánchez A, Baselga J, Villanueva J. 2012. Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics. Journal of Proteomics 75: 3938-3951.

Griffiths J, Rosenfeld MA. 1954. Operator variation in experimental research. The Journal of Geology 62:74-91.

Guo K, Li L. 2009. Differential ^{12}C-/^{13}C-isotope dansylation labeling and fast liquid chromatography/mass spectrometry for absolute and relative quantification of the metabolome. Analytical Chemistry 81:3919-3932.

Haghverdi L, Lun AT, Morgan MD, Marioni JC. 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nature Biotechnology 36:421-427.

Han W, Li L. 2015. Matrix effect on chemical isotope labeling and its implication in metabolomic sample preparation for quantitative metabolomics. Metabolomics 11:1733-1742.

Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, Mahendraker T, Williams M, Neumann S, Rocca-Serra P. 2012. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic acids research 41:D781-D786.

Hirayama A, Sugimoto M, Suzuki A, Hatakeyama Y, Enomoto A, Harada S, Soga T, Tomita M, Takebayashi T. 2015. Effects of processing and storage conditions on charged metabolomic profiles in blood. Electrophoresis 36:2148-2155.

Issaq HJ, Van QN, Waybright TJ, Muschik GM, Veenstra TD. 2009. Analytical and statistical approaches to metabolomics research. Journal of Separation Science 32:2183-2199.

Johnson SC. 1967. Hierarchical clustering schemes. Psychometrika 32:241-254.

Johnson WE, Li C, Rabinovic A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8:118-127.

Jonsson P, Gullberg J, Nordström A, Kusano M, Kowalczyk M, Sjöström M, Moritz T. 2004. A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. Analytical Chemistry 76:1738-1745.

Kanani H, Chrysanthopoulos PK, Klapa MI. 2008. Standardizing GC–MS metabolomics. Journal of Chromatography B 871: 191-201.

Karpievitch YV, Nikolic SB, Wilson R, Sharman JE, Edwards LM. 2014. Metabolomics data normalization with EigenMS. PLOS One 9:e116221.

Karpievitch YV, Taverner T, Adkins JN, Callister SJ, Anderson GA, Smith RD, Dabney AR. 2009. Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition. Bioinformatics 25:2573-2580.

Kell DB. 2006. Systems biology, metabolic modelling and metabolomics in drug discovery and development. Drug Discovery Today 11:1085-1092.

Kirwan J, Broadhurst D, Davidson R, Viant M. 2013. Characterising and correcting batch variation in an automated direct infusion mass spectrometry (DIMS) metabolomics workflow. Analytical and Bioanalytical Chemistry 405:5147-5157.

Kuligowski J, Pérez-Guaita D, Lliso I, Escobar J, León Z, Gombau L, Solberg R, Saugstad O, Vento M, Quintás G. 2014. Detection of batch effects in liquid chromatography-mass spectrometry metabolomic data using guided principal component analysis. Talanta 130:442-448.

Kuligowski J, Sánchez-Illana Á, Sanjuán-Herráez D, Vento M, Quintás G. 2015. Intra-batch effect correction in liquid chromatography-mass spectrometry using quality control samples and support vector regression (QC-SVRC). Analyst 140:7810-7817.

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics 11:733-739.

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

Li Y, Ruan Q, Li Y, Ye G, Lu X, Lin X, Xu G. 2012. A novel approach to transforming a non-targeted metabolic profiling method to a pseudo-targeted method using the retention time locking gas chromatography/mass spectrometry-selected ions monitoring. Journal of Chromatography A 1255:228-236.

Livera AMD, Sysi-Aho M, Jacob L, Gagnon-Bartsch JA, Castillo S, Simpson JA, Speed TP. 2015. Statistical methods for handling unwanted variation in metabolomics data. Analytical Chemistry 87:3606-3615.

Long NP, Nghi TD, Kang YP, Anh NH, Kim HM, Park SK, Kwon SW. 2020. Toward a standardized strategy of clinical metabolomics for the advancement of precision medicine. Metabolites 10:51.

Lu W, Bennett BD, Rabinowitz JD. 2008. Analytical strategies for LC–MS-based targeted metabolomics. Journal of chromatography B 871:236-242.

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

Maher AD, Zirah SF, Holmes E, Nicholson JK. 2007. Experimental and analytical variation in human urine in 1H NMR spectroscopy-based metabolic phenotyping studies. Analytical Chemistry 79:5204-5211.

Markley JL, Brüschweiler R, Edison AS, Eghbalnia HR, Powers R, Raftery D, Wishart DS. 2017. The future of NMR-based metabolomics. Current Opinion in Biotechnology 43:34-40.

Myers RP, Pollett A, Kirsch R, Pomier-Layrargues G, Beaton M, Levstik M, Duarte-Rojo A, Wong D, Crotty P, Elkashab M. 2012. Controlled attenuation parameter (CAP): a noninvasive method for the detection of hepatic steatosis based on transient elastography. Liver International 32:902-910.

Nygaard V, Rødland EA, Hovig E. 2016. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics 17:29-39.

Pandya K, Ray CA, Brunner L, Wang J, Lee JW, DeSilva B. 2010. Strategies to minimize variability and bias associated with manual pipetting in ligand binding assays to assure data quality of protein therapeutic quantification. Journal of Pharmaceutical and Biomedical Analysis 53:623-630.

Patti GJ, Tautenhahn R, Siuzdak G. 2012. Meta-analysis of untargeted metabolomic data from multiple profiling experiments. Nature Protocols 7:508-516.

Peng J, Chen Y-T, Chen C-L, Li L. 2014. Development of a universal metabolome-standard method for long-term LC–MS metabolome profiling and its application for bladder cancer urine-metabolite-biomarker discovery. Analytical chemistry 86:6540-6547.

Peralbo-Molina A, Calderón-Santiago M, Priego-Capote F, Jurado-Gámez B, De Castro ML. 2015. Development of a method for metabolomic analysis of human exhaled breath condensate by gas chromatography–mass spectrometry in high resolution mode. Analytica Chimica Acta 887:118-126.

Ransohoff DF. 2005. Lessons from controversy: ovarian cancer screening and serum proteomics. Journal of the National Cancer Institute 97:315-319.

Redestig H, Fukushima A, Stenlund H, Moritz T, Arita M, Saito K, Kusano M. 2009. Compensation for systematic cross-contribution improves normalization of mass spectrometry based metabolomics data. Analytical Chemistry 81:7974-7980.

Reese SE, Archer KJ, Therneau TM, Atkinson EJ, Vachon CM, De Andrade M, Kocher J-PA, Eckel-Passow JE. 2013. A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. Bioinformatics 29:2877-2883.

Reisetter AC, Muehlbauer MJ, Bain JR, Nodzenski M, Stevens RD, Ilkayeva O, Metzger BE, Newgard CB, Lowe WL, Scholtens DM. 2017. Mixture model normalization for non-

targeted gas chromatography/mass spectrometry metabolomics data. BMC Bioinformatics 18:84.

Rodríguez-Coira J, Delgado-Dolset MI, Obeso D, Dolores-Hernández M, Quintás G, Angulo S, Barber D, Carrillo T, Escribese MM, Villaseñor A. 2019. Troubleshooting in large-scale LC-ToF-MS metabolomics analysis: solving complex issues in big cohorts. Metabolites 9:247.

Rong Z, Tan Q, Cao L, Zhang L, Deng K, Huang Y, Zhu Z-J, Li Z, Li K. 2020. NormAE: deep adversarial learning model to remove batch effects in liquid chromatography mass spectrometry-based metabolomics data. Analytical Chemistry 92:5082-5090.

Saenz AJ, Petersen CE, Valentine NB, Gantt SL, Jarman KH, Kingsley MT, Wahl KL. 1999. Reproducibility of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for replicate bacterial culture analysis. Rapid Communications in Mass Spectrometry 13:1580-1585.

Salerno Jr S, Mehrmohamadi M, Liberti MV, Wan M, Wells MT, Booth JG, Locasale JW. 2017. RRmix: a method for simultaneous batch effect correction and analysis of metabolomics data in the absence of internal standards. PLOS One 12:e0179530.

Sanchez-Illana A, Pérez-Guaita D, Cuesta-García D, Sanjuan-Herráez JD, Vento M, Ruiz-Cerdá JL, Quintas G, Kuligowski J. 2018. Model selection for within-batch effect correction in UPLC-MS metabolomics using quality control-support vector regression. Analytica Chimica Acta 1026:62-68.

Sánchez-Illana Á, Piñeiro-Ramos JD, Sanjuan-Herráez JD, Vento M, Quintás G, Kuligowski J. 2018. Evaluation of batch effect elimination using quality control replicates in LC-MS metabolite profiling. Analytica Chimica Acta 1019:38-48.

Sangster T, Major H, Plumb R, Wilson AJ, Wilson ID. 2006. A pragmatic and readily implemented quality control strategy for HPLC-MS and GC-MS-based metabonomic analysis. Analyst 131:1075-1078.

Scholz M, Gatzek S, Sterling A, Fiehn O, Selbig J. 2004. Metabolite fingerprinting: detecting biological features by independent component analysis. Bioinformatics 20:2447-2454.

Ser Z, Liu X, Tang NN, Locasale JW. 2015. Extraction parameters for metabolomics from cultured cells. Analytical Biochemistry 475:22-28.

Shaham U, Stanton KP, Zhao J, Li H, Raddassi K, Montgomery R, Kluger Y. 2017. Removal of batch effects using distribution-matching residual networks. Bioinformatics 33:2539-2546.

Shen X, Gong X, Cai Y, Guo Y, Tu J, Li H, Zhang T, Wang J, Xue F, Zhu Z-J. 2016. Normalization and integration of large-scale metabolomics data using support vector regression. Metabolomics 12:89.

Soininen P, Kangas AJ, Würtz P, Suna T, Ala-Korpela M. 2015. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. Circulation: Cardiovascular Genetics 8:192-206.

Song X, Zhang B-L, Liu H-M, Yu B-Y, Gao X-M, Kang L-Y. 2011. IQMNMR: open source software using time-domain NMR data for automated identification and quantification of metabolites in batches. BMC Bioinformatics 12:337.

Stokvis E, Rosing H, Beijnen JH. 2005. Stable isotopically labeled internal standards in quantitative bioanalysis using liquid chromatography/mass spectrometry: necessity or not? Rapid Communications in Mass Spectrometry: an International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry 19:401-407.

Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, Edison A, Fiehn O, Higashi R, Nair KS. 2015. Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Research 44:D463-D470.

Sykes BD. 2007. Urine stability for metabolomic studies: effects of preparation and storage. Metabolomics 3:19-27.

Sysi-Aho M, Katajamaa M, Yetukuri L, Orešič M. 2007. Normalization method for metabolomics data using optimal selection of multiple internal standards. BMC Bioinformatics 8:93.

Teahan O, Gamble S, Holmes E, Waxman J, Nicholson JK, Bevan C, Keun HC. 2006. Impact of analytical bias in metabonomic studies of human blood serum and plasma. Analytical Chemistry 78:4307-4318.

Theodoridis G, Gika HG, Wilson ID. 2008. LC-MS-based methodology for global metabolite profiling in metabonomics/metabolomics. Trends in Analytical Chemistry 27:251-260.

Thonusin C, IglayReger HB, Soni T, Rothberg AE, Burant CF, Evans CR. 2017. Evaluation of intensity drift correction strategies using MetaboDrift, a normalization tool for multi-batch metabolomics data. Journal of Chromatography A 1523: 265-274.

Tung P-Y, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK, Gilad Y. 2017. Batch effects and the effective design of single-cell gene expression studies. Scientific Reports 7:39921.

van der Werf MJ, Overkamp KM, Muilwijk B, Coulier L, Hankemeier T. 2007. Microbial metabolomics: toward a platform with full metabolome coverage. Analytical Biochemistry 370: 17-25.

Veselkov KA, Vingara LK, Masson P, Robinette SL, Want E, Li JV, Barton RH, Boursier-Neyret C, Walther B, Ebbels TM. 2011. Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. Analytical Chemistry 83:5864-5872.

Vinaixa M, Samino S, Saez I, Duran J, Guinovart JJ, Yanes O. 2012. A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data. Metabolites 2:775-795.

Wang S-Y, Kuo C-H, Tseng YJ. 2012. Batch Normalizer: a fast total abundance regression calibration method to simultaneously adjust batch and injection order effects in liquid chromatography/time-of-flight mass spectrometry-based metabolomics data and comparison with current calibration methods. Analytical Chemistry 85:1037-1046.

Wehrens R, Hageman JA, van Eeuwijk F, Kooke R, Flood PJ, Wijnker E, Keurentjes JJ, Lommen A, van Eekelen HD, Hall RD. 2016. Improved batch correction in untargeted MS-based metabolomics. Metabolomics 12:88.

Weindl D, Wegner A, Jäger C, Hiller K. 2015. Isotopologue ratio normalization for non-targeted metabolomics. Journal of Chromatography A 1389:112-119.

Wishart DS. 2016. Emerging applications of metabolomics in drug discovery and precision medicine. Nature Reviews Drug Discovery 15:473-484.

Worley B, Powers R. 2013. Multivariate analysis in metabolomics. Current Metabolomics 1:92-107.

Wu L, Mashego MR, van Dam JC, Proell AM, Vinke JL, Ras C, van Winden WA, van Gulik WM, Heijnen JJ. 2005. Quantitative analysis of the microbial metabolome by isotope dilution mass spectrometry using uniformly ^{13}C-labeled cell extracts as internal standards. Analytical Biochemistry 336:164-171.

Wu Y, Li L. 2012. Determination of total concentration of chemically labeled metabolites as a means of metabolome sample normalization and sample loading optimization in mass spectrometry-based metabolomics. Analytical Chemistry 84:10723-10731.

Wu Y, Li L. 2016. Sample normalization methods in quantitative metabolomics. Journal of Chromatography A 1430:80-95.

Xia J, Broadhurst DI, Wilson M, Wishart DS. 2013. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. Metabolomics 9:280-299.

Yang Q, Wang Y, Zhang Y, Li F, Xia W, Zhou Y, Qiu Y, Li H, Zhu F. 2020. NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data. Nucleic Acids Research 48:W436-W448.

Yang S, Sadilek M, Lidstrom ME. 2010. Streamlined pentafluorophenylpropyl column liquid chromatography–tandem quadrupole mass spectrometry and global ^{13}C-labeled internal standards improve performance for quantitative metabolomics in bacteria. Journal of Chromatography A 1217:7401-7410.

Zaitsu K, Noda S, Ohara T, Murata T, Funatsu S, Ogata K, Ishii A, Iguchi A. 2019. Optimal inter-batch normalization method for GC/MS/MS-based targeted metabolomics with special attention to centrifugal concentration. Analytical and Bioanalytical Chemistry 411:6983-6994.

Zhang A, Sun H, Wang P, Han Y, Wang X. 2012. Modern analytical techniques in metabolomics analysis. Analyst 137:293-300.

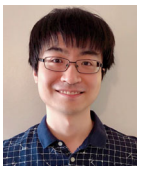
Zhao Y, Hao Z, Zhao C, Zhao J, Zhang J, Li Y, Li L, Huang X, Lin X, Zeng Z. 2016. A novel strategy for large-scale metabolomics study by calibrating gross and systematic errors in gas chromatography–mass spectrometry. Analytical Chemistry 88: 2234-2242.

Zhou B, Xiao JF, Tuli L, Ressom HW. 2012. LC-MS-based metabolomics. Molecular BioSystems 8:470-481.

Zhou H, Yuen PS, Pisitkun T, Gonzales PA, Yasuda H, Dear JW, Gross P, Knepper MA, Star RA. 2006. Collection, storage, preservation, and normalization of human urinary exosomes for biomarker discovery. Kidney International 69: 1471-1476.

Zukunft S, Sorgenfrei M, Prehn C, Möller G, Adamski J. 2013. Targeted metabolomics of dried blood spot extracts. Chromatographia 76:1295-1305.

AUTHOR BIOGRAPHIES



Wei Han received his PhD in analytical chemistry at the University of Alberta in 2017, under the supervision of Prof. Liang Li. His thesis work focused on the development of blood metabolomics for biomarker discovery using high-performance chemical isotope labeling LC-MS. After graduation, he continues to work in Prof. Li's lab to bring metabolomics technologies into mainstream bioscience and clinical labs. He develops data processing, metabolite identification, and statistical strategies for large-scale metabolomics. He also develops and applies novel bioinformatics tools for biomarker discovery of various diseases.

Liang Li received his PhD in Analytical Chemistry at the University of Michigan in 1989 under the supervision of Prof. David Lubman. He joined the University of Alberta in July 1989, where he is Professor of Chemistry and Adjunct Professor of Biochemistry. He is a Co-Director of the Metabolomics Innovation Centre (TMIC) of Canada. He is an elected fellow of the Royal Society of Canada (Academy of Science). Prof. Li was Tier 1 Canada Research Chair in Analytical Chemistry from 2005 to 2019. He served as Director, Alberta Cancer Board Proteomics Resource Laboratory, from 2000 to 2005. He was Chair of Analytical Chemistry Division at the University of Alberta from 2007 to 2019. Prof. Li has received a number of national and international awards and honors. He is an editor of *Analytica Chimica Acta* since 2005. He is also a member of the editorial advisory board in a number of scientific journals.

How to cite this article: Han W, Li L. Evaluating and minimizing batch effects in metabolomics. *Mass Spec Rev*. 2020;1–22.
https://doi.org/10.1002/mas.21672