Analytical and Bioanalytical Chemistry
https://doi.org/10.1007/s00216-018-0947-4

PAPER IN FOREFRONT

MCEE: a data preprocessing approach for metabolic confounding effect elimination

Yitao Li[1] · Mengci Li[1] · Wei Jia[1,2] · Yan Ni[2] · Tianlu Chen[1]

Received: 19 December 2017 / Revised: 26 January 2018 / Accepted: 5 February 2018

Abstract

It is well recognized that physiological and environmental factors such as race, age, gender, and diurnal cycles often have a definite influence on metabolic results that statistically manifests as confounding variables. Currently, removal or controlling of confounding effects relies heavily on experimental design. There are no available data processing techniques focusing on the compensation of their effects. We therefore proposed a new method, Metabolic confounding effect elimination (MCEE), to remove the influence of specified confounding factors and make the data more accurate. The method consists of three steps: metabolites grouping, confounder-related metabolites selection, and metabolites modification. Its effectiveness and advantages were evaluated comprehensively by several simulated models and real datasets, and were compared with two typical methods, the principal component analysis (PCA)- and the direct orthogonal signal correction (DOSC)-based methods. MCEE is simple, effective, and safe, and is independent of sample number, association degree, and missing value. Hence, it may serve as a good complement to existing metabolomics data preprocessing methods and aid in better understanding the metabolic and biological status of interest.

Keywords Metabolomics · Confounding factor · Generalized linear model · Principal component analysis · Direct orthogonal signal correction

Introduction

Confounding factors are those where its presence affects one or more of the variables being studied, leading to an increase or decrease in signal that can mislead final results [1]. Such confounding factors are often independent of experiments, but can affect the results greatly and, hence, usually lead to inaccurate or biased findings and conclusions. In the metabolomics field, more and more studies have reported and verified that physiological factors (e.g., race, BMI, and age) and environmental factors (e.g., diet, lifestyle, diurnal cycles, and geographical diversity) often contribute high levels of variability in characterizing the status and progression of many diseases and thus hinder our understanding of the metabolic and biological status of interest [2–4]. For example, glutamine, glycine, and glycerol phosphatidylcholine 42:0 (PCaa 42:0) serum concentrations were higher in obese compared with lean individuals, whereas PCaa 32:0, PCaa 32:1, and PCaa 40:5 were decreased in obese individuals, indicating that people with different BMI have different metabolic levels of phospholipids [5]. Our group also reported that some serum metabolites of Chinese (n = 211) and American (n = 72) obese populations [6] and some serum bile acids of a healthy Chinese population (n = 502) [7] were gender-dependent. In addition, gender, sleep and meal times were also important confounding factors in drug pharmacokinetic studies of Huangqi decoction, a famous traditional Chinese medicine [8].

For years, researchers have recognized this challenge and have been trying to control or partially remove confounding effects. Currently, three types of approaches exist but have limitations. The first type involves the experimental design or sample collection stage, where previous identification of possible confounding factors is done so as to compile only

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00216-018-0947-4) contains supplementary material, which is available to authorized users.

✉ Tianlu Chen
chentianlu@sjtu.edu.cn

[1] Center for Translational Medicine, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai 200233, China

[2] University of Hawaii Cancer Center, 701 Ilalo Street, Honolulu, HI 96813, USA

Published online: 24 February 2018

well-matched samples for the study. This approach may avoid the influence of specified confounding factors but with definite reduction of sample size and an increased burden involving time and cost. Furthermore, in some clinical studies, certain confounding factors are difficult to identify and quantify. The second type involves the data processing stage, and is done when experimental designs are premature, impractical, or impossible. Studies involving males and females or samples of different ages can be analyzed separately, with less statistical "power" [9, 10]. Metabolites related to confounding factors can be removed before data processing [11] at the cost of losing potential markers. Notably, some statistical analyses methods such as Stratification, logistic/linear regression, analysis of covariance (ANCOVA), and partial least squares (PLS) have also been used in this stage [12, 13]. Unfortunately, their major aims are not directly to remove confounding effects. For example, the objective of Stratification is to isolate partially but not remove confounders. The goal of logistic/linear regression and ANCOVA are group difference identification or variables association evaluation, although confounding effects are adjusted or considered to some extents. The aim of DOSC is to remove components unrelated (orthogonal) to the instructor variable, which often leads to an over-processed dataset. The third type is designed for experimental bias (e.g., non-constant instrument calibration, imperfect sample preparation, run order, and so on) correction [14–18]. A series of quality controls (QCs) and additional experimental work are required. We here reported a new method, MCEE (metabolic confounding effect elimination) for predefined confounding effect elimination. The effectiveness and advantages of MCEE were verified in simulated and real datasets and by comprehensive comparisons with two typical methods. It is actually a data preprocessing method without the requirement of additional experimental work and without the loss of samples or metabolites. It is safe as only specified confounding effects are removed and all other information remains intact. It is insensitive to data sparsity, association degree, and sample numbers. MCEE is beneficial for purer data and hence more accurate results.

Method and materials

Simulation dataset

Basic model

A series of simulated datasets were generated to assess various performances of MCEE and two other methods. First, a basic model with ideal conditions was generated. The basic model consisted of 400 samples (200 controls and 200 cases), one classification variable (*Y*), one confounding variable (*F*), and 45 metabolites (*M1-M45*) obeying normal distribution. The first few metabolites (*M1–M6*) were related to both *F* and *Y* (set randomly within the range of 0.9–1). The following eight metabolites (*M5–M12*) were set as differential ones for *Y* (e.g., case and control).

Different sample ratio model

In a well-designed study, equal numbers of case and control samples is a common requirement. However, in most practical situations, sample numbers of two different groups are often unequal. Consequently, based on the basic model, we designed another model to test the sensitivity of MCEE to a differing sample number ratio between two groups. The ratios of the case versus the control group were 0.5, 0.4, 0.3, 0.2, and 0.1, respectively.

Different correlation coefficient model

The correlation coefficients of the first few metabolites (*M1–M6*) and *F* and *Y* were set within the range of 0.9–1, indicating a close correlation. A series of datasets with correlation coefficients of 0.9, 0.8, 0.7, 0.6, 0.5, and 0.4 were generated in this model, hoping to examine the dependency of MCEE to association degree.

Sparse model

Missing values are very common in a metabolomic dataset and result probably from the fact that (1) some samples do not contain certain specific metabolites; (2) there exist low abundance peaks that are under the detection limit [19]; (3) software errors. We here constructed a model with various proportions of 0 values (0.1, 0.2, 0.3, 0.4, and 0.5) by replacing randomly some of the metabolite intensities by 0 values.

Real data

Human hepatocellular carcinoma (HCC) dataset

The metabolomic dataset of HCC was derived from our previously published paper [20]. The clinical diagnosis and pathological reports for all the patients were obtained from Zhongshan Hospital, Fudan University, Shanghai, China. In the HCC dataset for this report, a total of 48 patients diagnosed as hepatocellular carcinoma (88% patients with hepatitis B) and 17 matched healthy controls (18% with chronic hepatitis B) were enrolled. Any subject with steatohepatitis, inflammatory conditions, or gastrointestinal tract disorders was excluded from the control group. A total of 317 metabolites derived from GC-TOF/MS platform were reserved for the following analysis. HBV-induced HCC occurs in an environment of regeneration and inflammation that results from chronic liver damage, suggesting that the pathogenesis of HCC is immune-mediated [21]. Proteins encoded by HBV change host gene expression and cellular phenotypes that are considered as markers of cancer. These changes contribute to promoting growth factor-independent proliferation, tissue invasion, and metastasis, but most importantly the reprogramming of metabolism [22]. Hepatitis B viral load (HBV-DNA) level not only predicted the risk for cirrhosis independent of serum alanine transaminase level or HBeAg status but also was the strongest risk predictor. Therefore, we identified HBV-DNA level as a confounding factor.

Human arthritis dataset

The arthritis dataset was also based on our previously published paper [23]. The patients were from three hospitals, China-Japan Friendship Hospital in Beijing, Shanghai Guanghua Rheumatic Hospital, and the First Hospital affiliated to Anhui University of Chinese medicine. The arthritis data in this report contained a total of 51 patients, 26 (mean age = 51) diagnosed with gout arthritis (GA) and 25 (mean age = 31) diagnosed with Ankylosing spondylitis (AS). A total of 90 metabolites derived from GC-TOF/MS platform were ready for subsequent analysis. One more dataset from UPLC-QTOF/MS platform was also employed. Among the 53 patients, 27 (mean age = 53) were diagnosed with rheumatoid arthritis (RA) and 26 (mean age = 31) were diagnosed with Ankylosing spondylitis (AS). Three types of arthritis in these real datasets are very different in the age of onset. Ankylosing spondylitis (AS) mainly occurs among young men, and gout (GA) is mainly among middle-aged men [24–27]. Prevalence of rheumatoid arthritis (RA) is highest in women older than 65 years [28]. The metabolism of different ages is not the same [29]. In summary, we chose age as a confounding factor.

Data processing methods

Metabolic confounding effect elimination (MCEE)

The MCEE can be described in following steps.

Requirement: Assume we have a matrix *D=(Y,F,m_1,m_2,…,m_z)*, obeying the requirements of basic model.

Step-1(Grouping): Generalized linear model (GLM) [30] is applied on F and $M(m_1,m_2,\ldots,m_z)$ using four models (gamma, inverse Gaussian, normal, and passion) ($F = A \times M + E_1$). Metabolites with $p < 0.01$ from any one of the four models are screened out as F-related ones and saved in a set named M_a. Repeat this procedure on Y and M ($Y = B \times M + E_2$) and a set named M_b containing Y-related metabolites is generated. Descriptions on GLM are provided in the Electronic Supplementary Material (ESM).

Step-2($\mathbf{M_0}$ selection): Sort M_b in descending order. Top *ss*% (a controlling parameter set by users) metabolites in M_b are taken as M_c. A metabolite set named M_0 ($M_0 = \{M_0| , M_0 \in M_a \textit{ and } M_0 \notin M_c\}$) is generated for modification.

The risk parameter "*ss*" is designed to control how many Y-related metabolites will be excluded from the "metabolites to be modified" list (M_0) in order to avoid over-modification. Only metabolites with close correlations to F but limited (or controlled) correlations to Y will be revised subsequently. The larger the "*ss*" is, the fewer metabolites will be modified and the more conservative MCEE is. We set the "*ss*" value as 0.2 in this study, taking into consideration both group separation performance and biological significance of related metabolites. The guidance and an example of "*ss*" value selection are provided in the ESM (Fig. S1 and Fig. S2, Table S1).

Step-3(Modification): Every metabolite in M_0 awaiting modification can be described as

$$m_i = c_i \times Y + d_i \times F + e_i \quad (i = 1, 2 \ldots n)$$

The F-related term ($d_i \times F$) is removed from every metabolite to get M' and the resulting metabolite is more "pure" with no or decreased confounding effect. If more than one type of confounding effect was supposed to be eliminated, the user should rank all the possible confounding factors (concerning specific study aim and prior knowledge) and remove them orderly by conducting the above procedure several times.

Algorithm: Metabolic confounding effect elimination	Matlab functions
1a. $\boldsymbol{M_a}$←GLM on $\boldsymbol{M}$ and $\boldsymbol{F}$	1a. glmfit glmval
1b. $\boldsymbol{M_b}$←GLM on $\boldsymbol{M}$ and $\boldsymbol{Y}$	1b. glmfit glmval
2a. $\boldsymbol{M_c}$←Sort M_b in descending order and take top ***ss***% metabolites of $\boldsymbol{M_b}$	2a. sort
2b. $\boldsymbol{M_0}$← $\{\boldsymbol{M_0}| , \boldsymbol{M_0} \in \boldsymbol{M_a} \textbf{\textit{ and }} \boldsymbol{M_0} \notin \boldsymbol{M_c}\}$	2b. intersect union
3a. ***Model***← GLMs on every metabolite of $\boldsymbol{M_0}$ and ***F,Y***	3a. glmfit glmval
3b. $\boldsymbol{M'}$←Remove ***F***-related term from $\boldsymbol{M_0}$	

Principal component analysis (PCA)- based method

PCA is commonly used for dimension reduction. In this study, we involved PCA to remove confounding effects and compared its performance with that of MCEE as (1) it is a classic method and is widely used in metabolomics data processing, (2) the confounding effect was first noticed from PCA scores plots, and it is a good way to show the effect. Brief steps for the PCA based confounding effect elimination are as follows. (1) Process the original data by PCA and get some principal components (PCs). (2) Apply correlation analysis in PCs, F and M. PCs correlated to F ($p < 0.05$) but not Y ($p >$ =0.05) were screened out and removed directly. (3) Reconstruct the data using

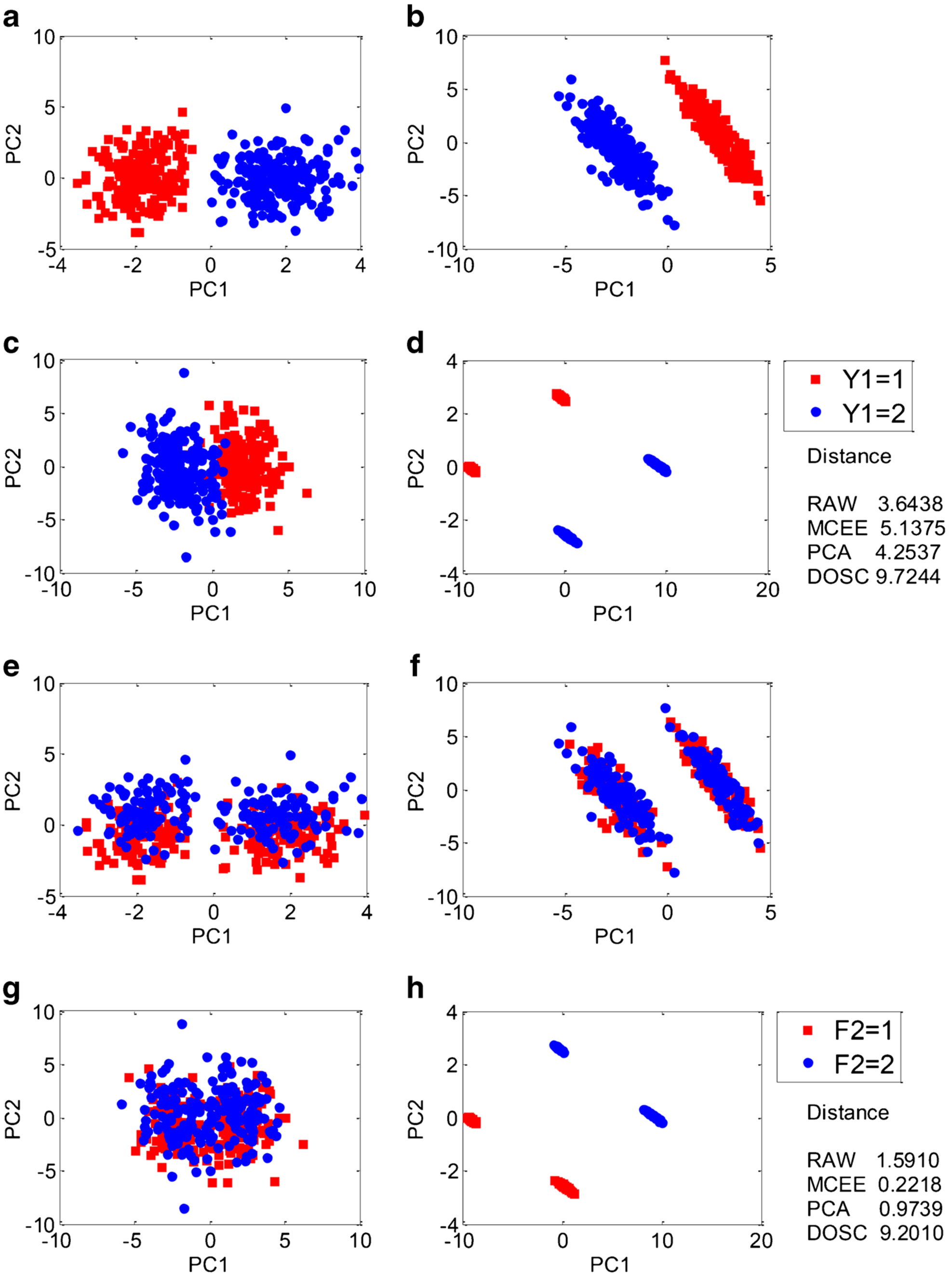


Fig. 1 PCA scores plots of the basic model. **(a)** and **(e)** are based on the raw dataset and are colored by *Y* and *F* groups, respectively, and the group center distance is 3.6438 and 1.5910, respectively; **(b)** and **(f)** are derived from MCEE processed dataset and are colored by *Y* and *F* groups, respectively, and the center distance is 5.1375 and 0.2218, respectively; **(c)** and **(g)** are derived from dataset processed by PCA-based method and are colored by *Y* and *F* groups, respectively, and the center distance is 4.2537 and 0.9739, respectively; **(d)** and **(h)** are derived from dataset processed by DOSC-based method and are colored by *Y* and *F* groups, respectively, and the center distance is 9.7244 and 9.2010, respectively

the remaining PCs and the mixing matrix. PCA was implemented by using the Matlab function "princomp".

Direct orthogonal signal correction (DOSC)-based method

The orthogonal signal correction (OSC) is a popular pretreatment method and its basic principle is to remove the parts unrelated (orthogonal) to the response matrix *Y* from spectral matrix *X* [31]. In metabolomics, OSC is often used together with partial least squares for group separation and detection of potential biomarkers. DOSC, the improvement version of OSC, is based on least squares steps and without such problems as the orthogonality towards *Y*, non-optimal amount of variance removed from *X*, and a non-attainable solution. Consistent with PCA-based method, we removed components unrelated to *Y* from *M* to obtain a "corrected" *M*. The DOSC was implemented by using a DOSC package developed by Biosystems Data Analysis Group of the Universiteit van Amsterdam [32].

Data processing and evaluation

The overall performances of three methods, MCEE (GLM-based), PCA-based, and DOSC-based, were compared. The

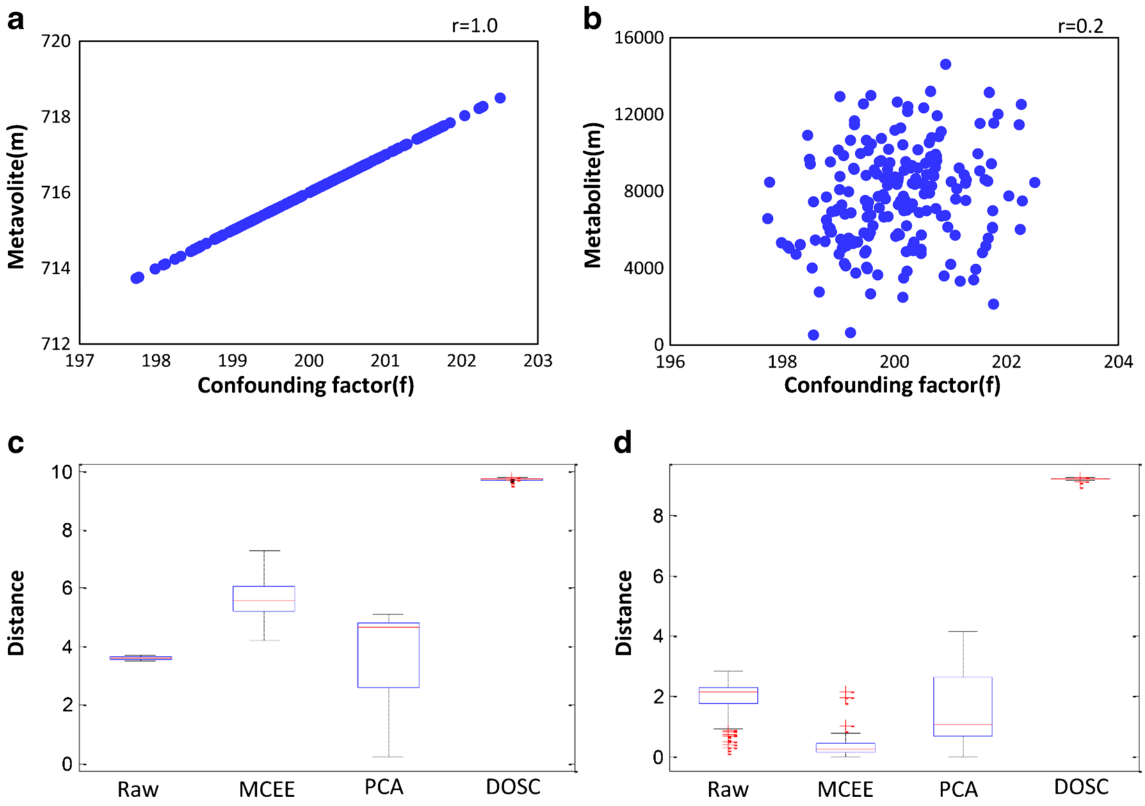


Fig. 2 Scatter plots and correlation coefficients of a metabolite (y axis) and *F* (x axis) before **(a)**, and after **(b)** MCEE processing. **(c)** Bar plots of the distance of the group centers, group by *Y* **(c)** and *F* **(d)**

PCA scores plot based on raw and corrected data were used to illustrate the effectiveness of different methods. The distance of the center of gravity of PCA plots was adopted to simplify the PCA plot and is shown in bar plots. Differential metabolites were screened out by using jointly the Students' t test (p<0.05) and random forest (mean decrease accuracy >1) [33, 34]

All the data processing and evaluation were completed using Matlab R2014a (Mathworks, US). The script of MCEE was provided in the ESM.

Results

Results on simulated datasets

As can be seen in the PCA scores plots derived from a basic model (Fig. 1a and e), there was a distinction between the two groups of *Y* (e.g. control and case groups) while the difference between the two groups of *F* (e.g. different race or gender) was obvious as well, indicating that the influence of *F* was noteworthy and may impact the true result. After MCEE processing, the separation of *Y* groups increased (Fig. 1b), with the group center distance altered from 3.64 to 5.14. As supposed, the confounding groups were well overlapped (Fig. 1f) indicating that there was no or dramatically decreased confounding effect in the processed data. Comparatively, the results of PCA-based method (Fig. 1c and g) were not as good as that of MCEE. Little change was found in the distance of *Y* group centers (before processing: 3.64 and after processing: 4.25) while the distance of *F* groups declined (from 1.59 to 0.9). For the DOSC-based method (Fig. 1d and h), after processing, the distances between *Y* and *F* group centers were both large indicating that the influence of *F* still existed. We recorded the CPU processing time (MacBook, Core i5, 8G RAM). The median processing time is 37.33 seconds. Additionally, Spearman correlation analysis was applied to *F* and every metabolite before and after MCEE processing. The correlation coefficients of *F* and almost all the modified metabolites were declined greatly. An example was shown as Fig. 2a and b. The Spearman correlation coefficient of *F* and a metabolite depleted from 1 to 0.21 by MCEE processing.

Further on, we generated 100 simulated basic datasets to validate the effectiveness and advantages of MCEE. Consistent with above findings, the distances of *Y* group centers derived from PCA scores plots enlarged greatly after the processing by MCEE and DOSC based methods, comparing with those of raw datasets (Fig. 2c). The distances of *F* group centers in the PCA scores plots were the lowest in the MCEE processed datasets (Fig. 2d). Consequently, the good performance of MCEE remained stable in our tests.

The above results were based on the basic model. More models were used to explore the overall performance of MCEE. Figure 3 illustrated the group center distance

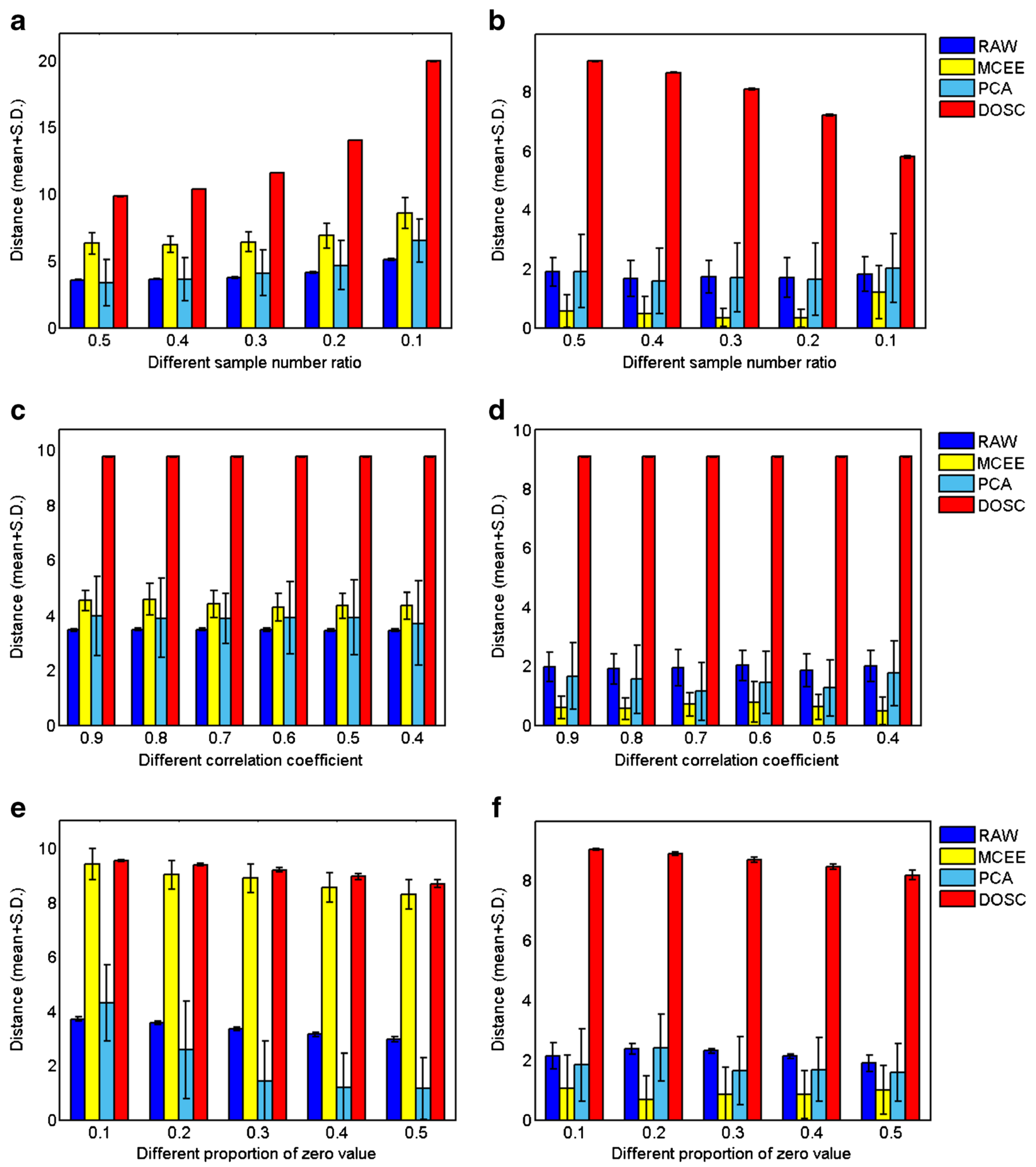


Fig. 3 Group center distances (mean with S.D.) of **(a)** different sample ratio model, group by *Y*; **(b)** different sample ratio model, group by *F*; **(c)** different correlation coefficient model, group by *Y*; **(d)** different correlation coefficient model, group by *F*; **(e)** Sparse model, group by *Y*; and **(f)** Sparse model, group by *F*

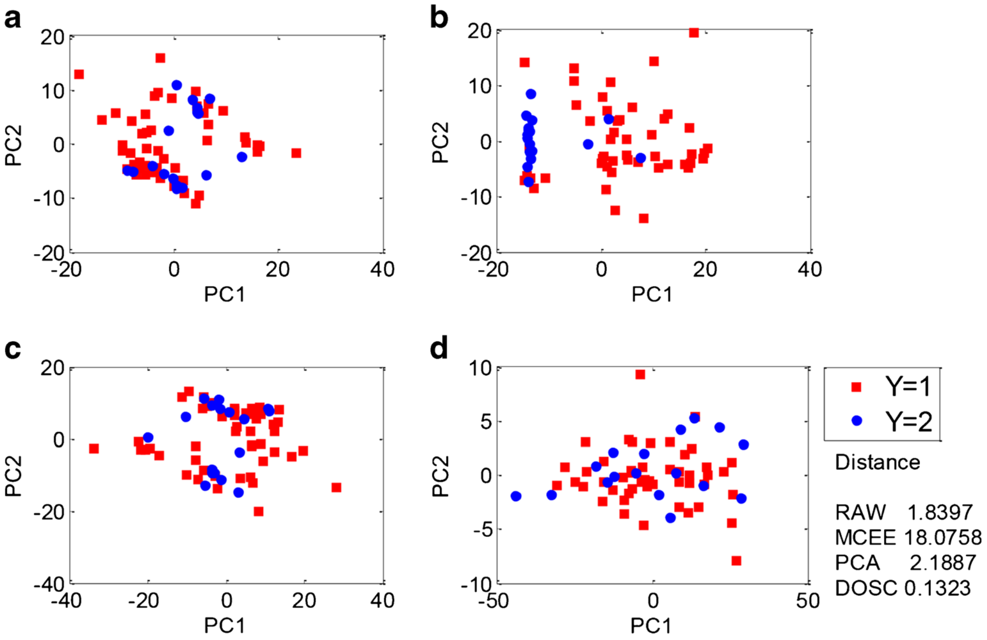
(mean with S.D., 50 times) of PCA scores plots based on the raw and processed datasets by MCEE and PCA and DOSC based methods, with the alterations of sample number ratio (Fig. 3a and b), correlation coefficient (Fig. 3c and d), and proportion of zero value (Fig. 3e and f) respectively. Once again, the performances of MCEE (the yellow bars) were superior to the other two methods in all the tests with large distances between *Y* groups (Fig. 3a, c, e) and small ones (Fig. 3b, d, f) between *F* groups.

Results on real datasets

Improved group separation was achieved by MCEE in the HCC dataset

A dataset derived from serum metabolic profiles of 48 HCC patients and 17 healthy controls was processed by the 3 methods respectively. For the clinical cases, PCA scores plot did not reveal differences between the disease and the control group (Fig. 4a).

Fig. 4 PCA scores plots of raw liver cancer data **(a)** controls in blue and cases in red, MCEE processed liver cancer data **(b)**, and datasets processed by PCA **(c)** and DOSC **(d)** based method; the center distance is 1.8397, 18.0758, 2.1887, and 0.1323 respectively



After MCEE processing, the influence of Hepatitis B, a confounding factor, was removed and better separation of the two groups was observed (Fig. 4b). No clear improvement was found in the results of PCA- based method (Fig. 4c). Notably, the group gap was extremely large in that of DOSC-based method (Fig. 4d). This might be caused by the undesirable over-fitting problem which frequently comes along with DOSC when coping with metabolomics data [31, 32].

MCEE played an important role in the detection of potential biomarkers in the arthritis dataset

MCEE and the other two methods were also applied to an arthritis dataset comprising two types of arthritis, AS (n=25) and GA (n=26). Similar to the results of the HCC dataset, the two types of arthritis can be separated clearly by PCA after MCEE processing (the influence of age was eliminated) (Fig. 5b).

In addition to better group separation, confounding effect elimination also contributed to the detection of biomarkers which was helpful in better classifying of arthritis subtypes. The mean decrease value (MDA, calculated by random forest) of differential metabolites (between AS and GA) derived from the original and the MCEE processed datasets were illustrated in Fig. 6. After MCEE processing, 8 additional differential metabolites (highlighted in red), adipic acid, alloxanoic acid, beta-glutamic acid, galactose, isoleucine, lysine, N-acetylglutamine and N-methyl-glutamic acid were screened out and only 1 metabolite, malic acid, was eliminated as a differential metabolite(highlighted in blue). The increasing/decreasing trends of the differential metabolites were unchanged although the fold change values were altered.

The performance of MCEE was consistently good in a dataset derived from UPLC-QTOF/MS platform (Fig. 5e, f, g, and h).

Discussion

Theoretically speaking, MCEE has some advantages over the other two methods. The basic principle of the DOSC method involves separation of the data into two parts, according to the grouping variable, and removal of the part that is unrelated (orthogonal) to the instructor group, *Y* from the matrix *M*. It means that all the information irrelevant to *Y* will be removed, including effects that were not confounding as well as valuable data that should have been saved. Hence, it is an efficient (sometimes over-efficient) way for distinguishing groups [31] and is usually used together with PLS. As such, it is not an appropriate technique for preidentified confounding effect elimination. On the other hand, removal of confounding factors using the PCA-based method is suboptimal as the decomposed PCs are components that have large variance but without any direct association with a confounding factor. As shown in Fig. 1e, the PC2 may contain not only a portion of effects of the confounding factor but also some other useful information although a clear separation of *F* group was observed in the scores plots. After processing, PC2 and many other PCs that were significantly associated with *F* were removed and thus both some influence of *F* and some useful information for the separation of *Y* groups were removed simultaneously. In contrast, MCEE focuses specifically on removing the specified confounding effect and only components of some metabolites definitely influenced by the factor

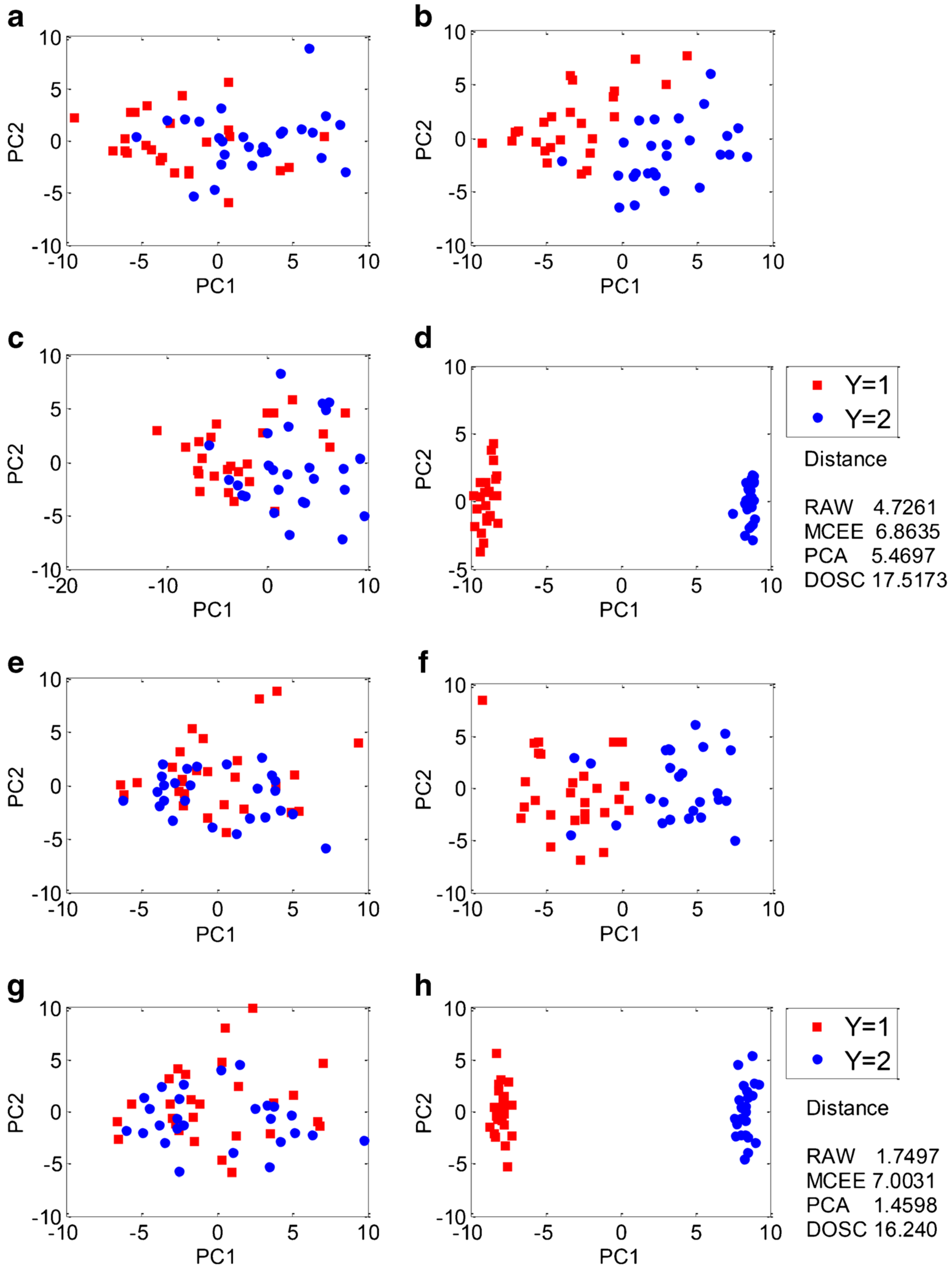


Fig. 5 PCA scores plots of raw arthritis data **(a)**, GA in blue, and AS in red, MCEE processed arthritis data **(b)**, and datasets processed by PCA **(c)** and DOSC **(d)** based method, and the center distance is 4.7261, 6.8635, 5.4697, and 17.5173, respectively. PCA scores plots of the arthritis dataset (derived from UPLC-QTOF/MS platform) based on raw data **(e)**, AS in blue and RA in red, MCEE processed data **(f)**, and datasets processed by PCA **(g)** and DOSC **(h)** based methods; the center distance is 1.7497, 7.0031, 1.4598, and 16.2400, respectively

will be corrected. Therefore, its performance was evaluated to be the best among the three methods tested.

Gout (GA), a chronic inflammatory disease, affects more than 1% of people in the USA [35], and 1.4% of people in the UK and Germany [25], and it is the most common form of inflammatory arthritis among men with metabolic hyperuricemia and deposition of urate crystals in articular and periarticular tissues [26, 36]. Ankylosing spondylitis (AS), a common inflammatory rheumatic disease, affects the axial skeleton, leading to characteristic inflammatory back pain, which can cause structural and functional impairments [24]. Malic acid is the intermediate product of the citric acid cycle and the body synthesizes it during the process of converting carbohydrates to energy. There is a generalized decline in oxidative activity in muscles with age [37]. The citrate synthase and isocitrate dehydrogenase, which are key enzymes in the citric acid cycle, decrease in activity with age [38]. The age of onset of AS and GA is not the same. The incidence of AS is mainly in young people, whereas GA is mainly in middle-aged people [25, 27, 35]. This may explain why the difference in malic acid level disappeared after removal of the confounding factor, the effect of age, using MCEE. Uric acid is the ultimate metabolite of the purine nucleotide cycle (PNC). The branched chain amino acids (BCAAs) are essential amino acids making up skeletal muscles in the human body and play an important role in protein synthesis.

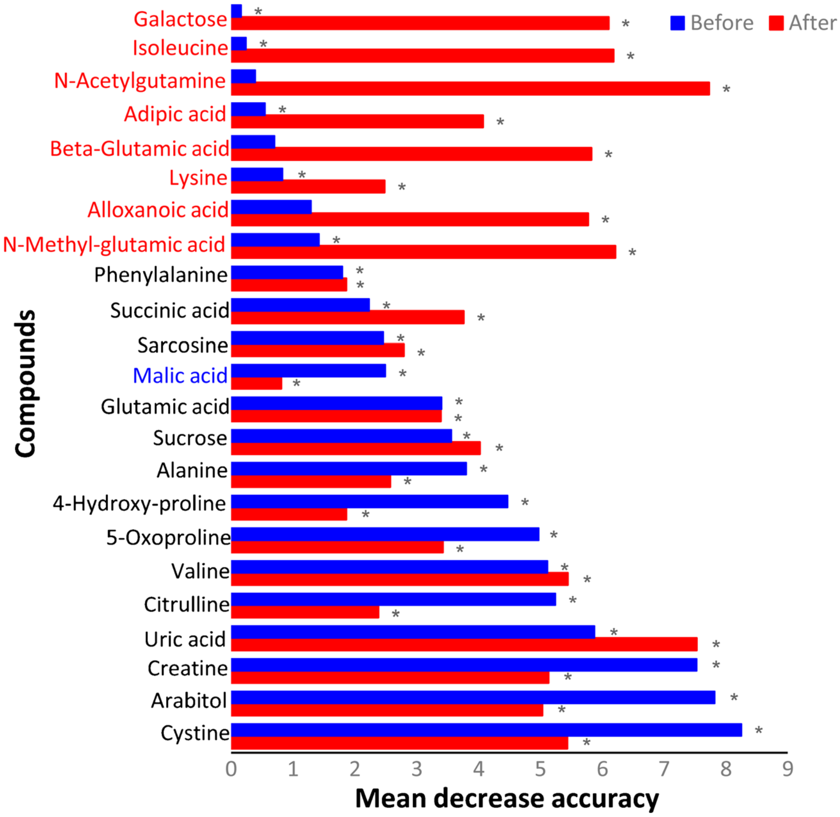


Fig. 6 Bar plots of the mean decrease accuracy (MDA, calculated by random forest) of potential biomarkers before (blue) and after (red) MCEE processing. Random forest (MDA > 1) and Students' *t*-test ($p < 0.05$) were applied jointly for potential biomarkers identification. New biomarkers identified from dataset processed by MCEE are highlighted in red. The metabolite that is significant before processing but is non-significant after processing is highlighted in blue. * Indicates Students' *t*-test $p < 0.05$

BCAAs and uric acid can interact with each other. BCAAs decrease uric acid production and reduce the incidence of gout in a person engaging in endurance exercise [39]. This may be a possible reason for the difference in the level of the BCAAs, isoleucine, between AS and GA. Lysine was identified as a differential metabolite between AS and GA, after MCEE processing. It is reported as the most important amino acid of type I collagen, the amplest structural protein in vertebrates [40]. Nonsteroidal anti-inflammatory drugs (NSAIDs), widely used by patients with AS, inhibit the synthesis of collagen and proteoglycan in varying degrees [24, 41, 42]. The increase or decrease of collagen synthesis directly affects the content of lysine in the blood**.** In summary, after MCEE processing, more differential metabolites supplying potential diagnostic information on arthritis were determined in the metabolomic profiles for the arthritis dataset.

Conclusions

It is highly recognized that confounding effects can influence and bias data acquired using metabolomic techniques and, therefore, it is advantageous to remove them before sample analysis or via reliable post-analysis data preprocessing methods [1]. Currently, confounding effects are considered as much as possible in the experimental design stage and there are few reliable methods available for controlling or removing confounding effect directly. Here we proposed, for the first time, a data preprocessing method, MCEE, for confounding effect elimination. Two typical methods and simulated and real datasets were adopted for its performance evaluation. MCEE was found to be superior to the other two methods in group classification and specified factor elimination with the following advantages. (1) Over-modification is well-controlled as classification (*Y*) and confounding variables (*F*) are both involved and only effects of specified factors were eliminated and all the other information was retained. (2) It was safe without the loss of sample and variable numbers. No additional experimental work is required. (3) It was independent of association degree, sample numbers, and missing values, and thus was adaptive to metabolomics data analysis. Limitations of the MCEE method do exist and further research must be done to perfect the performance. First, there is a need to develop a generalized version that can remove confounding effects of multiple factors at one time. MCEE is currently a linear method, although non-normal distributions are used in the GLM. Second, the confounding effect may persist, even after MCEE processing. More statistical strategies and criteria, besides the PCA scores plot and correlation analysis used here, may need to be applied for a more precise assessment. Third, before applying MCEE, confounding factors need to be decided and quantified correctly and accurately. This sometimes is a very difficult and complex issue and currently this relies mostly on a priori knowledge [43, 44]. Nevertheless, we highly recommend, based on the results of this study, that metabolomics datasets be processed by MCEE before statistical analysis when confounding effects may exist.

Acknowledgments This work was supported by the National Natural Science Foundation of China (31501079, 31500954 and 81772530), the National Key R&D Program of China (2017YFC0906800), and the Seventh Framework Programme of the European Union (294923). The authors thank the support of Biobank of Shanghai 6th People's Hospital.

Compliance with ethical standards

The protocol of HCC was approved by the Zhongshan Hospital Institutional Review Board and written consents were signed by all participants before the study. The protocol of arthritis was approved by the Review Board in Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, and all participants gave informed consent before they were involved in the study.

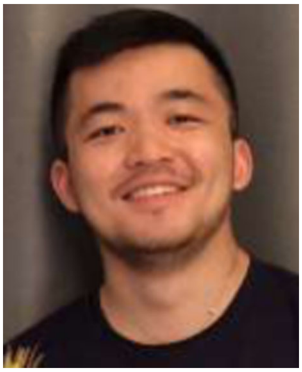
Conflict of Interest The authors declare that they have no competing interests.

References

1. Jager KJ, Zoccali C, Macleod A, Dekker FW. Confounding: what it is and how to deal with it. Kidney Int. 2008;73(3):256–60.
2. Hodson MP, Dear GJ, Roberts AD, Haylock CL, Ball RJ, Plumb RS, Stumpf CL, Griffin JL, Haselden JN. A gender-specific discriminator in Sprague-Dawley rat urine: the deployment of a metabolic profiling strategy for biomarker discovery and identification. Anal Biochem. 2007;362(2):182–92.
3. Moore SC, Matthews CE, Sampson JN, Stolzenberg-Solomon RZ, Zheng W, Cai Q, Tan YT, Chow WH, Ji BT, Liu DK, Xiao Q, Boca SM, Leitzmann MF, Yang G, Xiang YB, Sinha R, Shu XO, Cross AJ. Human metabolic correlates of body mass index. Metabolomics. 2014;10(2):259–69.
4. Slupsky CM, Rankin KN, Wagner J, Fu H, Chang D, Weljie AM, Saude EJ, Lix B, Adamko DJ, Shah S, Greiner R, Sykes BD, Marrie TJ. Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. Anal Chem. 2007;79(18):6995–7004.
5. Oberbach A, Bluher M, Wirth H, Till H, Kovacs P, Kullnick Y, Schlichting N, Tomm JM, Rolle-Kampczyk U, Murugaiyan J, Binder H, Dietrich A, von Bergen M. Combined proteomic and metabolomic profiling of serum reveals association of the complement system with obesity and identifies novel markers of body fat mass changes. J Proteome Res. 2011;10(10):4769–88.
6. Xie G, Ma X, Zhao A, Wang C, Zhang Y, Nieman D, Nicholson JK, Jia W, Bao Y, Jia W. The metabolite profiles of the obese population are gender-dependent. J Proteome Res. 2014;13(9):4062–73.
7. Xie G, Wang Y, Wang X, Zhao A, Chen T, Ni Y, Wong L, Zhang H, Zhang J, Liu C, Liu P, Jia W. Profiling of serum bile acids in a healthy Chinese population using UPLC-MS/MS. J Proteome Res. 2015;14(2):850–9.
8. Xie G, Wang S, Zhang H, Zhao A, Liu J, Ma Y, Lan K, Ni Y, Liu C, Liu P, Chen T, Jia W. Poly-pharmacokinetic study of a multicomponent herbal medicine in healthy Chinese volunteers. Clin Pharmacol Ther. 2017. https://doi.org/10.1002/cpt.784.
9. Zheng X, Chen T, Zhao A, Wang X, Xie G, Huang F, Liu J, Zhao Q, Wang S, Wang C, Zhou M, Panee J, He Z, Jia W. The brain metabolome of male rats across the lifespan. Sci Rep. 2016;6:24125.
10. Chen T, Ni Y, Ma X, Bao Y, Liu J, Huang F, Hu C, Xie G, Zhao A, Jia W, Jia W. Branched-chain and aromatic amino acid profiles and diabetes risk in Chinese populations. Sci Rep. 2016;6:20594.
11. Wei J, Xie G, Zhou Z, Shi P, Qiu Y, Zheng X, Chen T, Su M, Zhao A, Jia W. Salivary metabolite signatures of oral cancer and leukoplakia. Int J Cancer. 2011;129(9):2207–17.
12. Pourhoseingholi MA, Baghestani AR, Vahedi M. How to control confounding effects by statistical analysis. Gastroenterol Hepatol Bed Bench. 2012;5(2):79–83.
13. Christenfeld NJ, Sloan RP, Carroll D, Greenland S. Risk factors, confounding, and the illusion of statistical control. Psychosom Med. 2004;66(6):868–75.
14. Calderon-Santiago M, Lopez-Bascon MA, Peralbo-Molina A, Priego-Capote F. MetaboQC: a tool for correcting untargeted metabolomics data with mass spectrometry detection using quality controls. Talanta. 2017;174:29–37.
15. Thonusin C, IglayReger HB, Soni T, Rothberg AE, Burant CF, Evans CR. Evaluation of intensity drift correction strategies using MetaboDrift, a normalization tool for multi-batch metabolomics data. J Chromatogr A. 2017;1523:265–74.
16. van der Kloet FM, Bobeldijk I, Verheij ER, Jellema RH. Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping. J Proteome Res. 2009;8(11):5132–41.
17. Kamleh MA, Ebbels TM, Spagou K, Masson P, Want EJ. Optimizing the use of quality control samples for signal drift correction in large-scale urine metabolic profiling studies. Anal Chem. 2012;84(6):2670–7.
18. Wang SY, Kuo CH, Tseng YJ. Batch Normalizer: a fast total abundance regression calibration method to simultaneously adjust batch and injection order effects in liquid chromatography/time-of-flight mass spectrometry-based metabolomics data and comparison with current calibration methods. Anal Chem. 2013;85(2):1037–46.
19. Huan T, Li L. Counting missing values in a metabolite-intensity data set for measuring the analytical performance of a metabolomics platform. Anal Chem. 2015;87(2):1306–13.
20. Chen T, Xie G, Wang X, Fan J, Qiu Y, Zheng X, Qi X, Cao Y, Su M, Wang X, Xu LX, Yen Y, Liu P, Jia W. Serum and urine metabolite profiling reveals potential biomarkers of human hepatocellular carcinoma. Mol Cell Proteomics. 2011;10(7):M110.004945.
21. Rehermann B, Nascimbeni M. Immunology of hepatitis B virus and hepatitis C virus infection. Nat Rev Immunol. 2005;5(3):215–29.
22. Arzumanyan A, Reis HM, Feitelson MA. Pathogenic mechanisms in HBV- and HCV-associated hepatocellular carcinoma. Nat Rev Cancer. 2013;13(2):123–35.
23. Jiang M, Chen T, Feng H, Zhang Y, Li L, Zhao A, Niu X, Liang F, Wang M, Zhan J, Lu C, He X, Xiao L, Jia W, Lu A. Serum metabolic signatures of four types of human arthritis. J Proteome Res. 2013;12(8):3769–79.
24. Braun J, Sieper J. *Ankylosing spondylitis*. Lancet. 2007;369(9570):1379–90.
25. Annemans L, Spaepen E, Gaskin M, Bonnemaire M, Malier V, Gilbert T, Nuki G. Gout in the UK and Germany: prevalence, comorbidities, and management in general practice 2000–2005. Ann Rheum Dis. 2008;67(7):960–6.
26. Terkeltaub R. Update on gout: new therapeutic strategies and options. Nat Rev Rheumatol. 2010;6(1):30–8.
27. Wright KA, Crowson CS, Michet CJ, Matteson EL. Time trends in incidence, clinical features, and cardiovascular disease in *Ankylosing spondylitis* over three decades: a population-based study. Arthritis Care Res (Hoboken). 2015;67(6):836–41.
28. Scott DL, Wolfe F, Huizinga TW. Rheumatoid arthritis. Lancet. 2010;376(9746):1094–108.
29. Vignoli A, Tenori L, Luchinat C. Age and sex effects on plasma metabolite association networks in healthy subjects. J Proteome Res. 2018;17(1):97–107.
30. McCullagh P. Generalized linear models. Eur J Oper Res. 1984;16(3):285–92.
31. Luypaert J, Heuerding S, de Jong S, Massart DL. An evaluation of direct orthogonal signal correction and other preprocessing

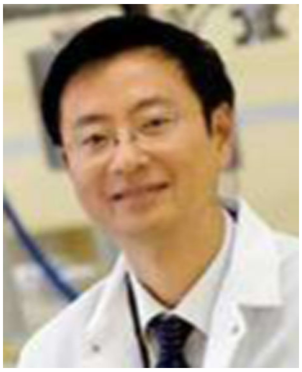
methods for the classification of clinical study lots of a dermatological cream. J Pharm Biomed Anal. 2002;30(3):453–66.

32. Westerhuis JA, Jong SD, Smilde AK. Direct orthogonal signal correction. Chemomet Intel Lab Syst. 2001;56(1):13–25.
33. Chen T, Cao Y, Zhang Y, Liu J, Bao Y, Wang C, Jia W, Zhao A. Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. Evid Based Complement Alternat Med. 2013;2013:298183.
34. Ma Y, Ding Z, Qian Y, Shi X, Castranova V, Harner EJ, Guo L. Predicting cancer drug response by proteomic profiling. Clin Cancer Res. 2006;12(15):4583–9.
35. Saag KG, Choi H (2006) Epidemiology, risk factors, and lifestyle modifications for gout. Arthritis Res Ther 8(Suppl 1:S2)
36. Terkeltaub RA. Clinical practice. Gout. N Engl J Med. 2003;349(17):1647–55.
37. Hansford RG, Castro F. Age-linked changes in the activity of enzymes of the tricarboxylate cycle and lipid oxidation, and of carnitine content, in muscles of the rat. Mech Aging Dev. 1982;19(2):191–200.
38. Vitorica J, Cano J, Satrustegui J, Machado A. Comparison between developmental and senescent changes in enzyme activities linked to energy metabolism in rat heart. Mech Aging Dev. 1981;16(2):105–16.
39. Tang FC, Chan CC. Contribution of branched-chain amino acids to purine nucleotide cycle: a pilot study. Eur J Clin Nutr. 2017;71(5):587–93.
40. Yamauchi M, Sricholpech M. Lysine post-translational modifications of collagen. Essays Biochem. 2012;52:113–33.
41. Fujii K, Tajiri K, Kajiwara T, Tanaka T, Murota K. Effects of NSAID on collagen and proteoglycan synthesis of cultured chondrocytes. J Rheumatol Suppl. 1989;18:28–31.
42. Palka J, Galewska Z. The effect of some antiinflammatory drugs on collagen of rat skin. Pol J Pharmacol Pharm. 1990;42(1):39–42.
43. Greenland S, Morgenstern H. Confounding in health research. Annu Rev Public Health. 2001;22:189–212.
44. McNamee R. Confounding and confounders. Occu Environ Med. 2003;60(3):227–34. quiz 164, 234

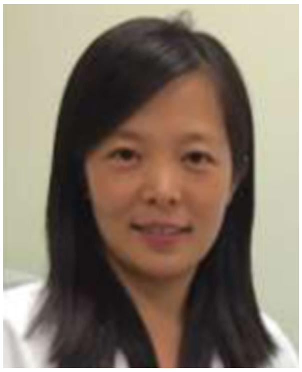


Yitao Li is a graduate student in the Center for Translational Medicine of Shanghai Jiao Tong University Affiliated Sixth People’s Hospital. He mainly engages in the study of metabolomics and bioinformatics.

Mengci Li is a PhD candidate of the Center for Translational Medicine at Shanghai Jiao Tong University Affiliated Sixth People's Hospital and the School of Biomedical Engineering and Med-X Research Institute at Shanghai Jiao Tong University in Shanghai, China. She majors in bioinformatics, including development of metabolomics and metagenomics softs and database as well as x-omics data analysis.



Wei Jia is the Associate Director for Shared Resources of the University of Hawaii Cancer and Director of the Center for Translational Medicine of Shanghai Jiao Tong University Affiliated Sixth People’s Hospital. His research interest involves carbon source metabolism and its regulation in cancer cells as well as the molecular mechanisms that link metabolic disruptions in gut microbial-host co-metabolism to metabolic disorders and gastrointestinal cancer.



Yan Ni is an assistant specialist of Cancer Epidemiology Program and Metabolomics Shared Resources at the University of Hawaii Cancer Center. She has been working for over 10 years in the field of mass spectrometry-based metabolomics and its applications.

Tianlu Chen is an associate researcher of the Center for Translational Medicine of Shanghai Jiao Tong University Affiliated Sixth People’s Hospital. Her research interest involves chemometrics and biological/medical information mining, especially metabolic diseases-related x-omics data processing and management.