Metabolomics (2016) 12:89
DOI 10.1007/s11306-016-1026-5

ORIGINAL ARTICLE

# Normalization and integration of large-scale metabolomics data using support vector regression

Xiaotao Shen[1] · Xiaoyun Gong[2] · Yuping Cai[1] · Yuan Guo[1] · Jia Tu[1] · Hao Li[1] · Tao Zhang[2] · Jialin Wang[3] · Fuzhong Xue[2] · Zheng-Jiang Zhu[1]

Received: 26 November 2015 / Accepted: 22 March 2016 / Published online: 26 March 2016

**Abstract**
*Introduction* Untargeted metabolomics studies for biomarker discovery often have hundreds to thousands of human samples. Data acquisition of large-scale samples has to be divided into several batches and may span from months to as long as several years. The signal drift of metabolites during data acquisition (intra- and inter-batch) is unavoidable and is a major confounding factor for large-scale metabolomics studies.
*Objectives* We aim to develop a data normalization method to reduce unwanted variations and integrate multiple batches in large-scale metabolomics studies prior to statistical analyses.
*Methods* We developed a machine learning algorithm-based method, support vector regression (SVR), for large-scale metabolomics data normalization and integration. An R package named MetNormalizer was developed and provided for data processing using SVR normalization.
*Results* After SVR normalization, the portion of metabolite ion peaks with relative standard deviations (RSDs) less than 30 % increased to more than 90 % of the total peaks, which is much better than other common normalization methods. The reduction of unwanted analytical variations helps to improve the performance of multivariate statistical analyses, both unsupervised and supervised, in terms of classification and prediction accuracy so that subtle metabolic changes in epidemiological studies can be detected.
*Conclusion* SVR normalization can effectively remove the unwanted intra- and inter-batch variations, and is much better than other common normalization methods.

**Keywords** Metabolomics · Data normalization · Data integration · Support vector regression · Quality control

## 1 Introduction

Metabolites are defined as the collection of small molecules that are produced during metabolism (Nicholson and Lindon 2008; Patti et al. 2012a; Rabinowitz and Silhavy 2013; Fiehn 2002). Mass spectrometry-based untargeted metabolomics has enabled simultaneous quantitative measurements of thousands of metabolites using minimal amounts of biological samples, providing functional readouts of physiological and pathological states of biological individuals at the systems level (Patti et al. 2012b). Although relatively new compared to genomics and proteomics, metabolomics has revealed new metabolic pathways in cell biology and improved our understanding of disease pathogenesis (Weiss and Kim 2012; Griffin et al. 2011; Patti et al. 2012a; Long et al. 2011). To enable a better understanding of physiological and pathological

**Electronic supplementary material** The online version of this article (doi:10.1007/s11306-016-1026-5) contains supplementary material, which is available to authorized users.

✉ Zheng-Jiang Zhu
jiangzhu@sioc.ac.cn

[1] Interdisciplinary Research Center on Biology and Chemistry, and Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, People's Republic of China

[2] Department of Epidemiology and Biostatistics, School of Public Health, Shandong University, Jinan 250012, People's Republic of China

[3] Shandong Cancer Hospital affiliated to Shandong University, and Shandong Academy of Medical Sciences, Jinan 250117, People's Republic of China

changes related to diseases and to define biomarkers for diagnosis, large-scale human samples are generally required for metabolomics studies to the extent of the epidemiological level in which thousands of samples are studied (Wang et al. 2011; Mapstone et al. 2014; Mayers et al. 2014). The validly large scale of human samples in biomarker discovery studies effectively averages out substantial variations observed in the human metabolome that are caused by differences in age, gender, diet, medication, lifestyle, stress and many additional factors (Wang et al. 2011; Mapstone et al. 2014; Mayers et al. 2014). Recent studies have advanced the application of the liquid chromatography-mass spectrometry (LC–MS) technique to large-scale studies of human biofluid samples (such as serum, plasma, and urine) (Wang et al. 2011; Mapstone et al. 2014; Luan et al. 2015). The untargeted metabolic profiling of one sample can be completed within a few minutes, enabling an analytical throughput of >100 samples per day (Evans et al. 2009; Lv et al. 2011). This capability of high-throughput analysis provided by LC–MS enables to analyze thousands to tens of thousands of samples within several months to 1 year.

However, the analysis of large-scale samples requires that great care is taken in experimental design, data acquisition, quality control, and subsequent data analysis. Obviously, not all of the samples can be analyzed in a single batch; therefore, one metabolomics study is usually divided into several batches, and data acquisition may span several months (Wang et al. 2011; Mapstone et al. 2014; Mayers et al. 2014; Bijlsma et al. 2006). The signal intensity drift of metabolites over time and across different batches is a major confounding factor in large-scale metabolomics studies. The unwanted variations in the measurements of metabolite ion peaks during data acquisition (intra- and inter-batch) are unavoidable and arise from sample handling and preparation, LC column degradation, matrix effects, MS instrument contamination and nonlinear drift over long runs (Leek et al. 2010; Burton et al. 2008; De Livera et al. 2015). Therefore, the development of a normalization method is necessary to remove the unwanted analytical variations occurring in intra- and inter-batch measurements and to integrate multiple batches forming an integral data set for subsequent statistical analysis (De Livera et al. 2015; De Livera et al. 2012). Effective removal of unwanted analytical variations helps increase the power of statistical analysis so that subtle metabolic changes in epidemiological studies can be detected (Veselkov et al. 2011).

For this purpose, several attempts have been made in the past several years to normalize data in large-scale metabolomics studies (van der Kloet et al. 2009; Veselkov et al. 2011). One of the most common normalization methods utilizes internal standard metabolites, which are added to the biological subject samples before or after extraction, for data normalization, such as ratio response (Bijlsma et al. 2006), NOMIIS (Sysi-Aho et al. 2007) and CCMN (Redestig et al. 2009). However, it is difficult to select one or several internal standards to normalize all metabolites that feature different polarities and functional groups in metabolic profiling. In addition, overlapped chromatographic peaks and ion suppression effects introduce biases into the internal standard-based normalization method (van der Kloet et al. 2009). Therefore, internal standards are usually added to monitor the reproducibility of sample preparation and LC–MS analysis, which is not an ideal choice for data normalization. Other methods, such as sum (Cairns et al. 2008), median (Wang et al. 2003) or L2 (Scholz et al. 2004) normalizations, use sample-wise scalar corrections for data normalization. But these scalar correction based methods are not applicable to most metabolomics experiments, as they heavily rely on the self-averaging property (Sysi-Aho et al. 2007). The pros and cons of all normalization methods, including internal standard based normalization, sample-wises scalar normalization, and variance based normalization (De Livera et al. 2012; Huber et al. 2002), have been comprehensively discussed and compared in several recent articles (van den Berg et al. 2006; Kamleh et al. 2012; De Livera et al. 2015).

Recently, utilizing quality control (QC) samples for data normalization become more popular (van der Kloet et al. 2009; Dunn et al. 2011, 2012; Kamleh et al. 2012; Wang et al. 2013). QC samples are prepared by pooling aliquots of biological subject samples in the study that are representative of the sample type under analysis, and then, periodically analyzing these samples over the entire data acquisition time course (Dunn et al. 2012). Intensity drifts of metabolites in biological subject samples can be detected by observing the signal changes of the same metabolite in the QC samples. For example, batch ratio based normalization method uses the mean (or median) intensity of all QC samples for each batch as correction factor to normalize dataset (Kamleh et al. 2012). Alternatively, QC sample also can be used to build regression as correction factor. A regression model is built based on the intensity drift of each metabolite in the QC samples and is used to predict and correct peak intensities of the same metabolite in subject samples. Linear (Kamleh et al. 2012; Wang et al. 2013) and non-linear (van der Kloet et al. 2009; Dunn et al. 2011) regressions are the two most common regression methods. However, linear regression with least squares cannot fit QC samples well, because most of the signal drifts have non-linear changes (Fig. 1). LOESS (locally estimated scatterplot smoothing) curve fitting regression combines the simplicity of classical least squares-based regression with the flexibility of nonlinear regression,

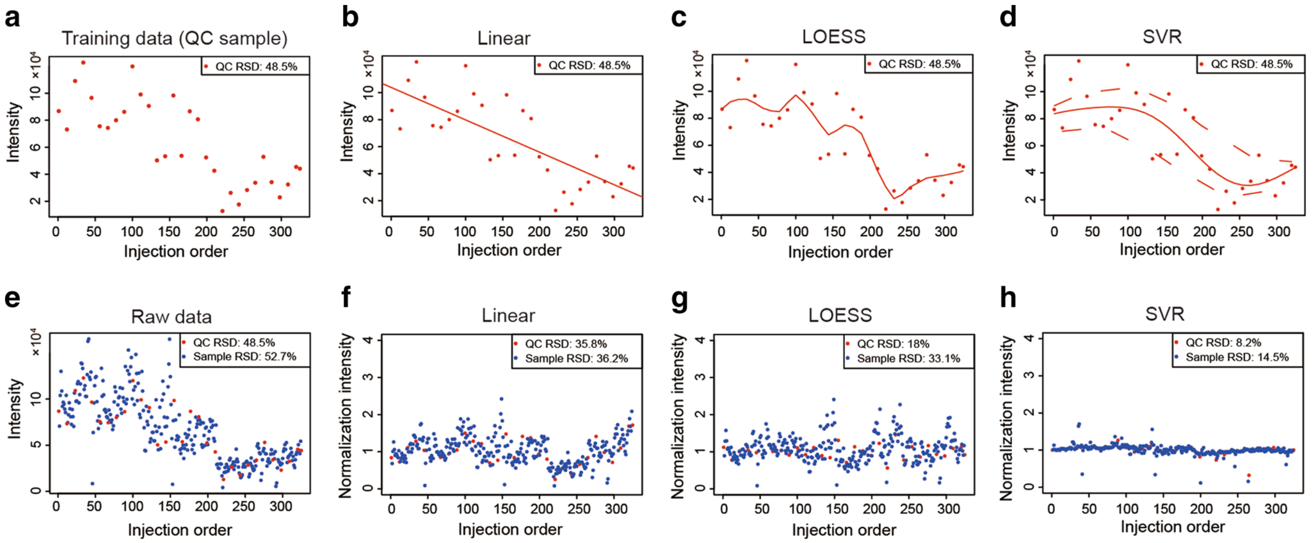

**Fig. 1** Comparison of data normalization methods for removing unwanted analytical variations: **a** non-linear signal drift of training samples (QC samples) during data acquisition; **b–d** applications of linear (**b**), LOESS (**c**), and SVR- (**d**) based data normalization methods for fitting non-linear signal drifts of training data; **e** signal drift of subject samples during data acquisition; and **f–h** normalized QC and subject samples using linear (**f**), LOESS (**g**), and SVR- (**h**) based data normalization methods. Parameters for LOESS normalization were set as follows: degree = 1 and span = 0.4. SVR normalization used Gaussian RBF kernel with C = 1 and ε = 0.1. The two *red dotted lines* are the margin lines of the support vectors. The selected metabolite peak is M623T423 (*m/z* 623.256; retention time 423 s) (Color figure online)

which can fit non-linear signal drift well (van der Kloet et al. 2009; Dunn et al. 2011). However, the principle of local non-linear regression in LOESS-based normalization limits its generation ability to predict the values of subject samples, and overfitting of the training data (i.e., QC samples) happens quite frequently in the presence of QC sample outliers. Therefore, a good regression model for QC-based normalization should meet two important criteria: (1) accurate non-linear fitting for signal drift and (2) excellent generation ability for predicting the values of subject samples and robust capability of removing the effects of outliers in training data. Both criteria can be evaluated by reducing the relative standard deviations (RSDs) of peaks in samples and improving classification and predictive accuracy in subsequent multivariate statistical analyses (Veselkov et al. 2011; Wang et al. 2013).

In this work, we introduce one of the most widely used machine learning methods, support vector regression (SVR), for the purpose of data regression in metabolomics (Brereton and Lloyd 2010; Ren et al. 2015). SVR maintains all of the main features that characterize the maximal margin algorithm in the support vector machine (SVM) algorithm (Cortes and Vapnik 1995), and a non-linear kernel function (e.g., radial basis function) is used for margin optimization and maximization (Steinwart and Christmann 2008). In SVM, dependent variables are discrete and the goal is to find the hyperplane which can separate observations into different classes. But in SVR, dependent variables are continuous and the hyperplane is used to predict the distribution of observations (Steinwart and Christmann 2008; Brereton and Lloyd 2010; Ren et al. 2015). Therefore, SVR normalization can accurately fit non-linear signal drift, is not susceptible to the presence of outliers, and thus has excellent generation capability for prediction when compared with linear and LOESS-based normalizations. In the past decade, the SVM method has been used in the fields of genomics (Fujarewicz et al. 2007) and metabolomics (Guan et al. 2009) for biomarker selection but has not been used for normalization of large-scale metabolomics data. By introducing the SVR method for normalization and integration of large-scale metabolomics data, the unwanted intra- and inter-batch variations were largely reduced, and the performances of multivariate statistical analyses, both unsupervised and supervised (e.g., PCA and OPLS-DA), have been largely improved in terms of classification and prediction accuracy. In addition, although the SVR normalization method is demonstrated on the normalization and integration of LC–MS-based metabolomics data in this work, the principle and method can be easily applied to GC–MS data. Finally, an R package named MetNormalizer was developed and provided for data processing using SVR normalization. R package MetNormalizer is available at http://www.metabolomics-shanghai.org/.

## 2 Materials and methods

### 2.1 Chemicals

LC–MS grade methanol (MeOH), water ($H_2O$), acetonitrile (ACN), water with 0.1 % formic acid (FA), and acetonitrile with 0.1 % FA were purchased from Honeywell (Muskegon, MI, USA). Ammonium fluoride ($NH_4F$) was purchased from Sigma (St. Louis, MO, USA). Commercial human serum sample was purchased from Equitech-Bio, Inc (Kerrville, TX, USA).

### 2.2 Sample preparation and LC–MS analysis

Serum samples were extracted using Bravo liquid handling system (Agilent Technologies, USA), and LC–MS analyses were performed using a UHPLC system (1290 series, Agilent Technologies, USA) coupled to a quadruple time-of-flight (Q-TOF) mass spectrometer (Agilent 6550 iFunnel Q-TOF, Agilent Technologies, USA). The details are provided in the supplementary material.

### 2.3 Data analysis

MS raw data (.d) files were converted to the mzXML format using ProteoWizard, and processed by XCMS (Smith et al. 2006; Tautenhahn et al. 2008). R package CAMERA (Kuhl et al. 2012) was used for peak annotation after XCMS data processing. The details of data analysis are provided in the supplementary material.

### 2.4 SVR-based data normalization

SVR based machine learning method is developed for the purpose of data regression in this work. SVR maintains all of the main features that characterize the maximal margin algorithm in SVM, and has a non-linear kernel function for margin optimization and maximization (Steinwart and Christmann 2008; Brereton and Lloyd 2010; Ren et al. 2015). SVM method uses supervised learning models for classification, however, it is usually called as SVR when it is used for regression analysis. Therefore, SVR has excellent learning and prediction abilities. A tutorial review article that explains SVR for application in analytical chemistry has been recently introduced (Brereton and Lloyd 2010).

The format of subject sample dataset was defined in an S×P matrix. S stands for S subject samples and P stands for P peaks. The format of QC sample dataset was defined as a Q×P matrix, where Q represents Q QC samples and P represents P peaks. The most important concept of QC sample-based normalization is that the signal drift of peaks in QC samples can represent instrument drift over the injection order, and the intensity drift of the QC and subject samples are affected by instrument drift in the same way. In this work, an SVR model of each peak in QC samples was first constructed. For one peak M = {$x_i$, $y_i$}, x is the independent viable such as injection order, and y is dependent viable which is the peak intensity, i = 1, 2, …, P. Therefore, SVR function can be written as Formula (1) shows.

$$f(x) = w\Phi(x) + b \tag{1}$$

$w$ is the normal vector and $\Phi$ represents a non-linearity transformation from $R^n$ (n-dimension real number space) to higher dimensional feature space. The goal of support vector-based regression is to find the $w$ and $b$ that make the minimum structure risk function, which is quite different from other traditional regressions. The solution to the optimization problem towards minimum structure risk function can be written as Formula (2) shows.

$$\begin{aligned}&\text{minimize } \frac{1}{2}||w||^2 + C\sum_{i=1}^{P}(\xi_i + \xi_i^*)\\ &\text{subject to}\begin{cases} y_i - f(x_i) - b \le \varepsilon + \xi_i \\ f(x_i) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0\end{cases}\end{aligned} \tag{2}$$

$||w||$ is the norm of normal vector so the width of margin is $2/||w||$. $\xi_i$ and $\xi_i^*$ represent the slack variables of upper and lower boundaries for the margin, respectively. $\varepsilon$ is the error tolerance. To get the minimum structure risk function which is fitted data well, so we must minimize $\frac{1}{2}||w||^2 + C\sum_{i=1}^{P}(\xi_i + \xi_i^*)$ under the prerequisites that all the data points in training dataset (i.e., peaks in QC sample dataset) are in the defined region. An in-depth theoretical background about SVR can be found in the literature (Steinwart and Christmann 2008; Brereton and Lloyd 2010; Ren et al. 2015).

In this work, we selected injection order or the top most correlated ion peaks (based on Pearson's correlation) of QC samples as independent variables to construct SVR function, and compared their performances on data normalization. As the results shown in supplementary Fig. 1, the use of top most correlated ion peaks as independent variables for regression is better than the use of injection order. Therefore, in our work, we choose top five most correlated ion peaks as independent variables for regression. Therefore, the SVR model built on QC samples can be written as Formula (3) shows.

$$f(x_{Q.cor}) = w\Phi(x_{Q.cor}) + b \tag{3}$$

$x_{Q.cor}$ is the top five most correlated peaks of QC samples. Then, the built SVR model for peak M was used to predict the intensities of the same peak in S subject samples.

$$y_{predict} = w\Phi(x_{S.cor}) + b \quad (4)$$

$x_{S.cor}$ is the same five most correlated peaks in the subjected samples. The values of w and b are solved from Formula (3). Finally, the intensities of each peak in the subject samples ($y$) were divided by the predictive peak intensities ($y_{predict}$) for normalization to remove unwanted intensity drift and analytical variations during data acquisition. Peaks in QC samples were normalized in the same way.

$$y' = y/y_{predict} \quad (5)$$

The SVR-based normalization used in this work has been developed as an R package named MetNormalizer, which is available in the supplementary and at our group website (http://www.metabolomics-shanghai.org/). MetNormalizer can be installed in Windows, Linux and Mac OS. The example serum data from study 1 (see below) are also provided. After installation of MetNormalizer, one can learn to run MetNormalizer according to the instruction in the help document using the example data provided.

### 2.5 Study designs

In this work, two metabolomics studies were designed to evaluate the performance of the SVR-based normalization method.

(1) Study 1 was an experiment which is used to assess the performance of normalization methods for removing intra-batch variations. The batch contained 37 QC samples, 288 subject samples, ten column conditioning samples and 43 blank samples. Column conditioning samples were injected ten times at the beginning of analysis to equilibrate the column. The QC samples were analyzed after every eight subject samples in the entire batch. The detailed run order is provided in supplementary Table 1. The entire batch required approximately 74 h in total for LC–MS analysis. Both the QC and subject samples consisted of the same commercial human serum samples.

(2) Study 2 was a large-scale metabolomics study that aimed to discover metabolite biomarkers for early diagnosis of esophagus cancer. The study was approved by the ethics committee of Shandong Cancer Hospital affiliated to Shandong University, and written informed consents were obtained from all participants involved in this study. There were 768 human serum samples in this study, and basic metrics regarding the cohort are provided in supplementary Table 2. All of the serum samples were collected at the Tumor Preventative and Therapeutic Base of Shandong Province (Feicheng People's Hospital), and the participants were screened using endoscope and iodine staining for esophagus cancer (golden standard for diagnosis of esophagus cancer). The participants were divided into two classes according to their reaction to iodine staining: screening positive and screening negative. Screening positive patients further received histopathology diagnosis to confirm cancer progression stages. Serum samples were analyzed using the LC–MS method described above. All of the 768 samples were randomly divided into four batches. Each batch contained 192 subject samples, 25 QC samples, ten column conditioning samples, and 31 blank samples. The running order within each batch was also randomized. One batch required approximately 52 h in total for either positive or negative MS analyses. The analyses of the four batches spanned approximately 2 months, including instrument maintenance and unexpected repairs. The detailed run order is provided in supplementary Table 3. This study was used to evaluate the performance of the SVR-based normalization method to remove inter-batch variations and to assess the improvement in the accuracy of statistical analyses after data normalization.

## 3 Results and discussion

To demonstrate the excellent performance of the SVR method for normalization and integration of large-scale untargeted metabolomics data, we designed two studies in this work (see the Materials and methods section for details).

### 3.1 Removal of unwanted variations using SVR-based normalization

SVR can effectively remove unwanted analytical variations during data acquisition compared with other QC-based regression methods such as linear and LOESS regressions. For the purpose of data normalization, a regression model was first built based on the intensity drift of each metabolite using training data (i.e., QC samples) (Fig. 1a–d). Then, the intensities of the same peaks in subject samples were predicted. Therefore, the unwanted analytical variations during data acquisition for each peak were removed. A commonly adapted criterion to assess the reproducibility of bioanalytical methods provided by the FDA is that the relative standard deviation (RSD) for a single analyte test should be within 15 % of QC samples (FDA 2013). In biomarker discovery studies, peaks with RSDs less than 30 % are typically accepted (Wang et al. 2013). Therefore, having RSDs less than 30 % was adopted as one of the most important criteria for performance evaluation of normalization methods.

As shown in Fig. 1, a peak (M623T423: m/z 623.256; retention time 423 s) was chosen as an example to demonstrate the non-linear signal drift over the injection order during data acquisition. The RSDs of this peak were 48.5 and 52.7 % in QC and subject samples, respectively, highlighting the large analytical variations between the samples (Fig. 1e). Here, we applied linear, LOESS and SVR normalization methods to fit the signal drift and calculated the RSDs of the M623T423 peak in QC and subject samples after data normalization (Fig. 1f–h). In general, the linear regression did not fit the signal drift in the QC samples very well (Fig. 1b). Therefore, RSDs for the M623T423 peak in the QC and subject samples slightly decreased to 35.8 and 36.2 %, respectively, after linear normalization (Fig. 1f), larger than the acceptable level recommended by the FDA. LOESS normalization fitted the signal drift in the QC samples better than linear normalization (Fig. 1c), as indicated by the significant decrease in the RSD of QC samples from 48.5 to 18.0 % (Fig. 1g). However, LOESS normalization had a limited generation capability for subject samples. The RSD of subject samples only decreased from 52.7 to 33.1 % (Fig. 1g), which is still higher than the acceptable level. The results showed that LOESS normalization has a good capability to fit non-linear signal drift and reduce variations in QC samples but has a relatively poor generation capability because variations between the subject samples could not be effectively removed.

In contrast, SVR had excellent generation capability and robustness (Fig. 1d, h). SVR normalization fit the signal drift in the QC samples well, as indicated by the significant decrease in the RSDs of the M623T423 peak in the QC samples from 48.5 to 8.2 %. Additionally, SVR normalization also effectively reduced the RSDs of the M623T423 peak in the subject samples from 52.7 to 14.5 % (Fig. 1h), highlighting the excellent generation capability of SVR normalization. In the metabolomics data, the QC outliers greatly influence the accuracy of non-linear fitting and are deleterious to perform normalization. When SVR fits the signal drift of the QC samples, it uses the margin lines to define the non-linear signal drift with extended tolerance to outliers and errors (such as the red dashed lines in Fig. 1d). Therefore, compared with LOESS normalization, SVR is not susceptible to QC outliers, which makes it more robust for data normalization and more effective at removing unwanted variations. Other examples to demonstrate that the SVR normalization can normalize different kinds of signal drifts are provided in supplementary Fig. 2. These results showed that SVR normalization can successfully remove unwanted variations in measurements of individual metabolites in both QC and subject samples compared with linear and LOESS normalization methods.

We further evaluated the overall performance of the SVR normalization method for normalization of the entire data set for study 1. After untargeted metabolomics profiling, there were 8413 ion peaks (or called feature) detected by XCMS in total. After XCMS processing, CAMERA was used to annotate the ion peaks. Ion peaks such as isotopic ions, adduct ions, multiple charged ions, and un-annotated ion were removed and discarded for further analysis. As a result, 1197 monoisotopic ion peaks (most are metabolite ion peaks) were chosen for data normalization and subsequent statistical analysis, and referred as “metabolite ion peaks”. The RSDs of each metabolite ion peak are shown as two-dimensional heat plots in Fig. 2a-d. In the heat plots, each point represents a metabolite ion peaks and the colors indicate the scale of the RSDs, with red representing lower RSDs. Clearly, a small portion of the metabolite ion peak achieved lower RSDs after linear and LOESS normalizations (Fig. 2b, c) compared with the raw data (Fig. 2a). The median RSD of the raw data only slightly decreased from 27.2 (IQR: 19.6–41.7 %) to 24.7 % (IQR: 18–37.8 %) and to 20.4 % (IQR: 15.3–32.8 %) after linear and LOESS normalizations, respectively. As a comparison, the RSDs of peaks were significantly decreased across the entire range using the SVR normalization method (Fig. 2d). The medium RSD significantly decreased to 9.7 % (IQR: 5.5–19.0 %) after SVR data normalization. Further data analysis showed that 1194 of 1197 metabolite ion peaks, as much as 99.7 %, had decreased RSDs after SVR normalization (Fig. 3a).

In the raw data, 677 of 1197 metabolite ion peaks (56.6 %) had RSDs less than 30 %. After linear and LOESS normalizations, the percentages increased to 62.7 and 75.5 %, respectively (Fig. 2e, f). However, the percentage of peaks with RSDs less than 30 % significantly increased to 90.7 % after SVR normalization (Figs. 2h, 3b). Therefore, qualified numbers of metabolite ion peaks (i.e., RSDs $<$ 30 %) for subsequent statistical analyses increased from 677 to 751, 904, and 1086 after linear, LOESS, and SVR normalizations, respectively. More importantly, after SVR normalization, the proportion of peaks with RSDs less than 10 % achieved the largest percentage, as high as 50.8 % of the total peaks. As a comparison, the proportions of peaks with RSDs less than 10 % were only 4.0 and 8.9 % after linear and LOESS normalizations, respectively (Figs. 2e–h, 3b). Furthermore, we evaluated the performance of SVR normalization for different abundant peaks, as shown in Fig. 3c, d. After SVR normalization, the proportion of peaks with RSDs less than 30 % was increased in peaks in a very broad range regardless of peak intensities (Fig. 3c, d and supplementary Table 4). The results demonstrated that the SVR normalization reduced the RSDs of metabolite ion peaks independently from peak intensities.

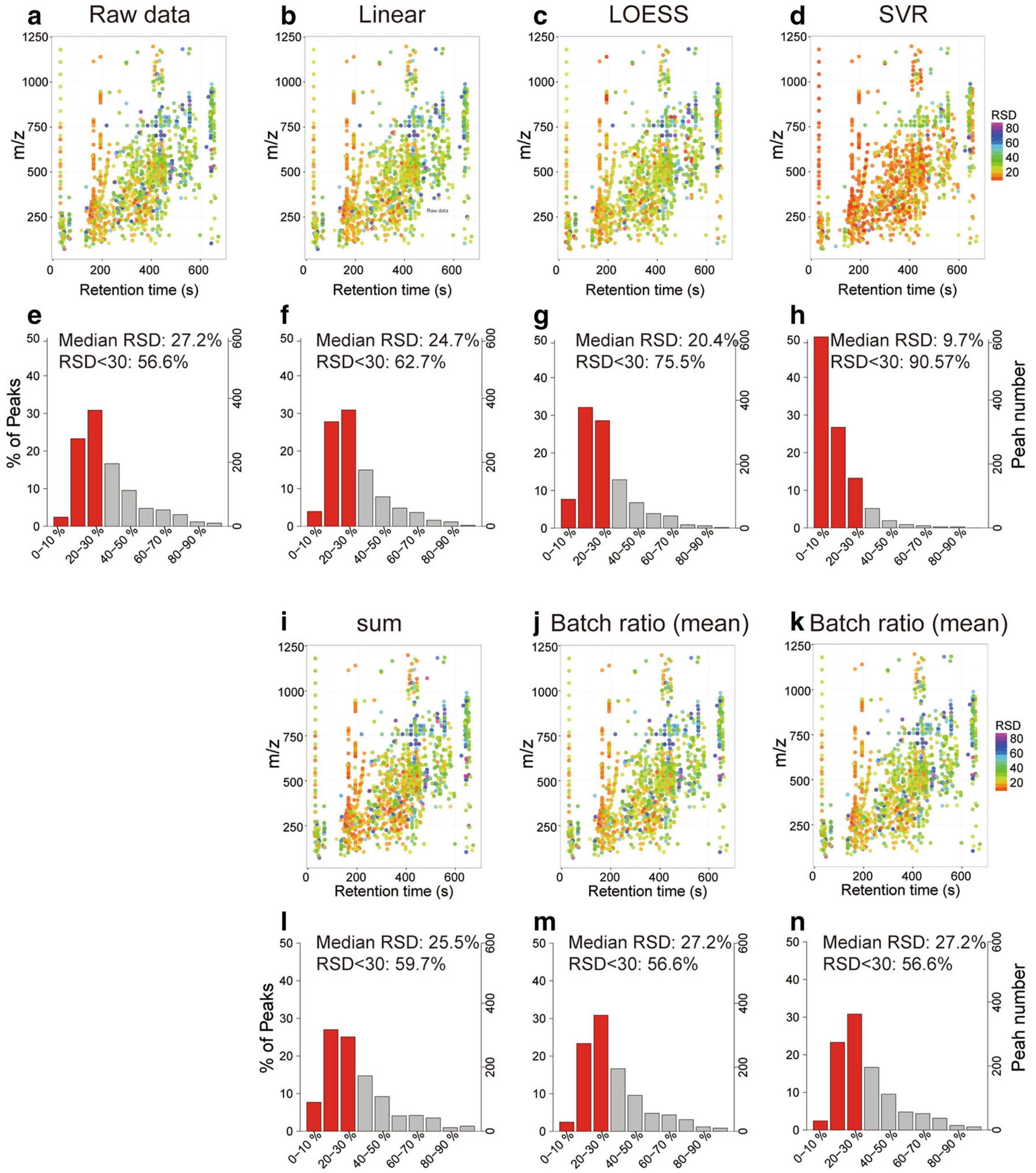

**Fig. 2** The distributions of the RSDs of metabolite ion peaks after data normalization. **a–d**, **i–k** Two-dimensional heat plots of peaks before and after data normalization: raw data (**a**), after linear normalization (**b**), after LOESS normalization (**c**), after SVR normalization (**d**), after sum normalization (**i**), after batch ratio (mean) normalization (**j**), and after batch ratio (median) normalization (**k**). *Each point* represents a metabolite ion peak, and the *colors* indicate the scale of the RSD. **e–h, l–n** Bar plots of the RSD distributions across all samples: raw data (**e**), after linear normalization (**f**), after LOESS normalization (**g**), after SVR normalization (**h**), after sum normalization (**l**), after batch ratio (mean) normalization (**m**), and after batch ratio (median) normalization (**n**)

We further compared SVR normalization against commonly used sample-wise scalar correction based normalization, such as sum normalization, and other QC-based batch ratio normalization (Fig. 2i–n). Sum normalization uses the total intensity of all peaks in one sample as correction factor. Batch ratio based normalization method uses the mean (or median) intensity of all QC samples for each batch as correction factor. Here, both batch ratio (mean) and batch ratio (median) methods are compared. After sum and batch ratio (mean and median) normalizations, the percentages of metabolite ion peaks with RSDs less than 30 % are 59.7, 56.6 and 56.6 %, respectively (Fig. 2i–n).

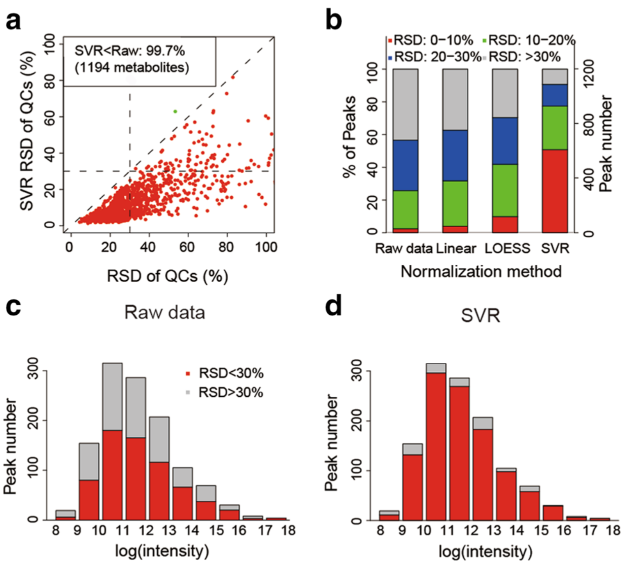

**Fig. 3** The performance of SVR normalization for reducing variations of metabolite ion peaks. **a** The change in RSDs in QC samples after SVR normalization. **b** The RSD distributions after different normalization methods. **c**, **d** The RSD distributions for different abundant peaks: raw data (**c**) and after SVR normalization (**d**)

The little changes indicate these methods cannot reduce analytical variations during metabolomics analysis.

Principal component analysis (PCA) is a commonly used unsupervised statistical method that can measure the similarity of samples through the tightness of sample clustering in PCA score plots. Here, we used a PCA score plot to evaluate the reproducibility of metabolomics data in study 1 and compared the performance of different data normalization methods. As shown in Fig. 4, after SVR normalization, QC and subject samples cluster much more tightly than the raw data, linear normalization, and LOESS normalization. We also calculated the median distances of QC and subject samples in PCA score plot to quantitatively evaluate the tightness degree of the clustering. The values for median distances are provided in supplementary Table 5. In brief, the median distance of QC samples decreased from 22.3 to 1.68 after SVR normalization. Similarly, the median distance between subject samples decreased from 22.9 to 5.12 after SVR normalization. As a comparison, the median distance of QC samples slightly decreased from 22.3 to 18.9 and 15.9 after linear and LOESS normalizations, respectively. Both QC and subject samples clustered much more tightly after SVR normalization compared with linear and LOESS normalizations, which indicates the excellent generation capability of SVR normalization. Therefore, we conclude that the SVR normalization method can significantly remove unwanted analytical variations and improve data reproducibility.

As a comparison, after sum and batch ratio (mean and median) normalizations, the median distances of QC samples in PCA plots are 22.4, 22.3 and 22.3, respectively (Fig. 4e–g and supplementary Table 5). And the median distances of subject samples in PCA plots are 22.3, 22.9 and 22.9, respectively (Fig. 4e–g and supplementary Table 5). The median distances are not changed after data normalization compared to raw data (22.3 for QC sample, and 22.9 for subject samples), and the results are consistent with the previous RSD reduction results, proving that these methods cannot reduce unwanted analytical variations during analysis.

## 3.2 Normalization and integration of multiple metabolomics datasets

One large-scale metabolomics study has to be divided into several batches, and data acquisition may span as long as several months to years (Bijlsma et al. 2006). The signal intensity drift across different batches is another major confounding factor for metabolomics studies. Here, we designed metabolomics study 2 to demonstrate that SVR normalization can successfully remove inter-batch variations and integrate multiple batches from one laboratory to form an integral dataset for subsequent statistical analysis. The 768 subject samples and 100 QC samples that were divided into four batches were processed together by XCMS. There were 9976 and 3592 metabolic features in positive and negative modes, respectively. After CAMERA annotation, there were 1024 and 497 metabolite ion peaks in positive and negative modes, respectively. Then, metabolite ion peaks in positive and negative modes were combined (1521 metabolite ion peaks in total) and subjected to SVR-based data normalization to remove inter-batch variations.

First, we utilized the PCA score plot to assess the performance of SVR normalization to remove inter-batch variations. As shown in the PCA score plot (Fig. 5a), the subject samples were distributed in four different clusters before data normalization. However, after SVR normalization, the subject samples from the four batches were clustered tightly in the PCA score plot (Fig. 5b). All of the QC samples from the four batches cluster very tightly. These results show that the inter-batch variations are effectively removed after data normalization.

Secondly, auto-scaled intensity plots were also used to visualize the reduced inter-batch variations after SVR normalization. Each of the 1521 peaks was subtracted by mean (also known as centering) and divided by standard deviation all of the QC samples in the four batches (van den Berg et al. 2006). Thus, all of the peaks had the same unit scale. Then, box plots were prepared to describe the standard deviations (SDs) of the peaks in each QC sample. If the inter-batch variations were removed, then the box plots for the QC samples of these centered peaks should

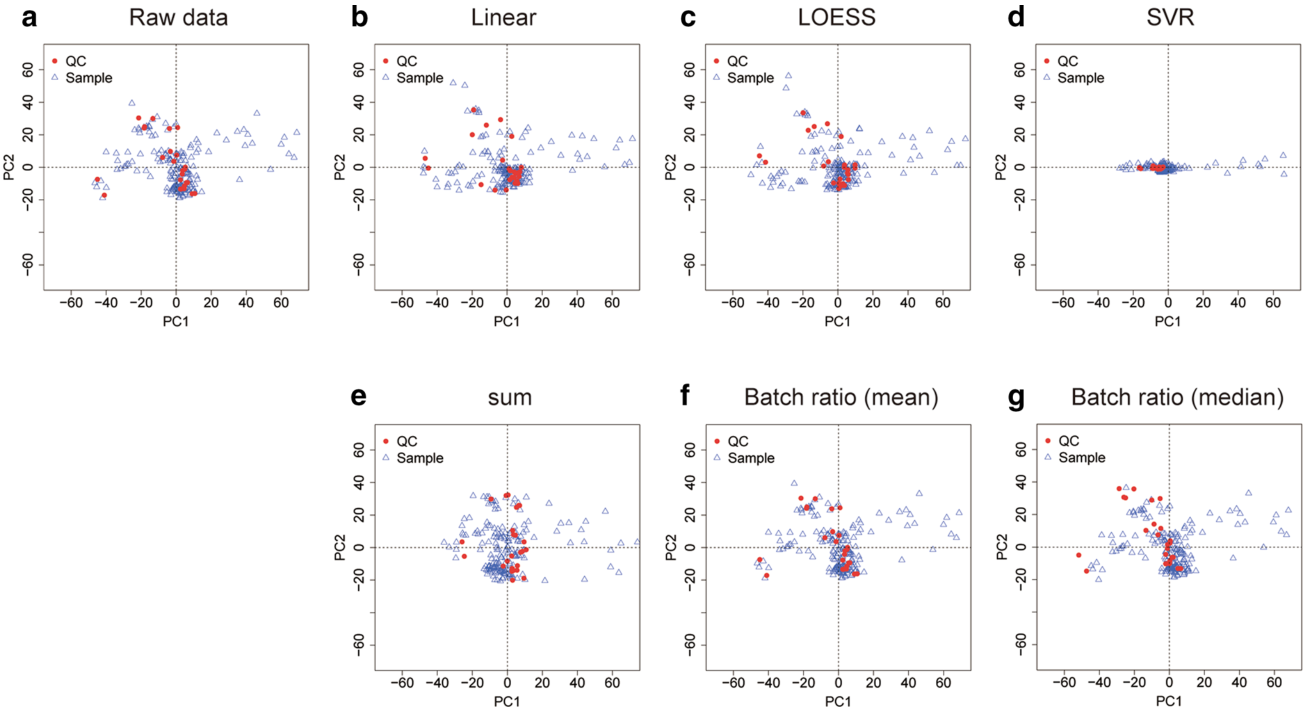

**Fig. 4** PCA score plots obtained from the raw data (**a**), after linear normalization (**b**), after LOESS normalization (**c**), after SVR normalization (**d**), after sum normalization (**e**), after batch ratio (mean) normalization (**f**), and after batch ratio (median) normalization (**g**). *Red circles* represent the QC samples, and *blue triangles* represent the subject samples (Color figure online)

have a mean close to zero with small variations around the mean. As shown in Fig. 5c, d, the standard deviations of the peaks in the QC samples were largely reduced after SVR normalization.

Finally, the percentage of peaks with RSDs less than 30 % were also calculated before and after SVR normalization. As shown in the cumulative RSD curves (Fig. 5e) and RSD distribution bar plots (Fig. 5f), the proportion of peaks with RSDs less than 30 % increased significantly to 1421 peaks (93.5 %) after SVR normalization compared with the raw data (499 peaks, 32.8 %). The proportion of peaks with RSDs less than 10 % achieved the largest percentage after SVR normalization, with as many as 48.6 % of the total peaks (739 peaks). In contrast, in the raw data, the proportion of peaks with RSDs less than 10 % was only 2.1 % of the total peaks (32 peaks). These results demonstrated that SVR normalization can successfully remove variations and integrate multiple batches into one integral dataset.

### 3.3 Improved classification accuracy with multivariate statistical analysis

The SVR-based data normalization method effectively reduces unwanted analytical variations and improves the power of statistical analyses so that subtle metabolic changes in metabolomics studies can be detected. Here, we used metabolomics study 2 as an example to demonstrate how the SVR normalization method improves classification and predictive accuracy in biomarker discovery. Study 2 had 768 subject serum samples, with 236 and 532 screening negative and positive samples, respectively.

First, we utilized a supervised multivariate analysis method (partial least squares, PLS) to select potential biomarkers that can distinguish between screening negative and positive subject samples in study 2. After PLS analysis, each of the 1521 metabolite ion peaks had a calculated value called variable importance in projection (VIP) to assess its contribution to classification. Higher VIP values indicate higher contribution; therefore, we selected the top 30 ranked peaks as potential biomarkers according to their VIP value ranks in the raw data and the data after SVR normalization (supplementary Tables 6, 7). Then, the 30 peaks were used as independent variables to construct the discrimination analysis model to demonstrate their performance on classification. Here, orthogonal partial least squares-discrimination analysis (OPLS-DA) was performed to visualize the classification performance, as shown in Fig. 6a, b. It was clear that the 30 peaks selected in the raw data did not well separate the screening positive and negative groups. The values of $R^2X$ and $Q^2cum$ for the raw data in the OPLS-DA model were 0.6 and 0.59, respectively. On the contrary, after SVR data normalization, the screening positive and negative groups were distinguished very well

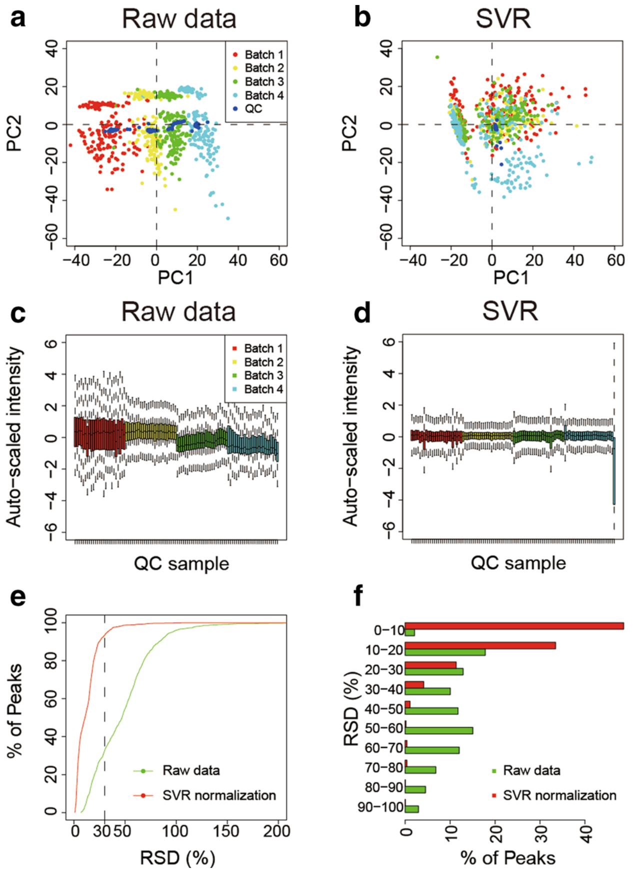

**Fig. 5** The performance of SVR normalization for removing inter-batch variations and integrating multiple metabolomics datasets. **a**, **b** The PCA score plots of metabolomics datasets in study 2 before and after SVR normalization. **c**, **d** The auto-scaled intensity box plots before and after SVR normalization. **e** The cumulative RSD curves of QC samples before and after SVR normalization. **f** The RSD distribution bar plots before and after SVR normalization

by the 30 selected peaks. The values for $R^2X$ and $Q^2$cum significantly increased to 0.99 and 0.99, respectively. We then use PLS-DA with double cross validation (Rosenberg et al. 2010) to construct receiver operating characteristic curves (ROCs) for raw and SVR normalized datasets (Fig. 6c; supplementary Fig. 4). The areas under the curves (AUCs) were 0.869 and 0.945 for the raw and SVR normalized datasets, respectively. This demonstrated that the SVR normalization can significantly improve the predictive accuracy of metabolomics datasets. In addition, we further compared the VIP values, p values (after FDR correction) and RSDs of selected 30 potential markers before and after SVR normalization (supplementary Fig. 3 and supplementary Table 8). Compared to raw data, most of metabolite markers have increased VIP values, decreased p values and RSDs after SVR normalization. These results indicated that SVR normalization can help select the effective potential markers by removing unwanted variations and improving the classification and predictive accuracy of multivariate statistical analysis.

## 4 Concluding remarks

In large-scale metabolomics studies, the signal drift of metabolites during data acquisition (intra- and inter-batch) is a major confounding factor that affects the accuracy of subsequent statistical analyses for biomarker discovery purposes. In this work, we introduced a machine learning algorithm-based normalization method, SVR normalization, which accurately fit non-linear signal drift and had excellent capacity for data regression and normalization. The SVR normalization method effectively removed the

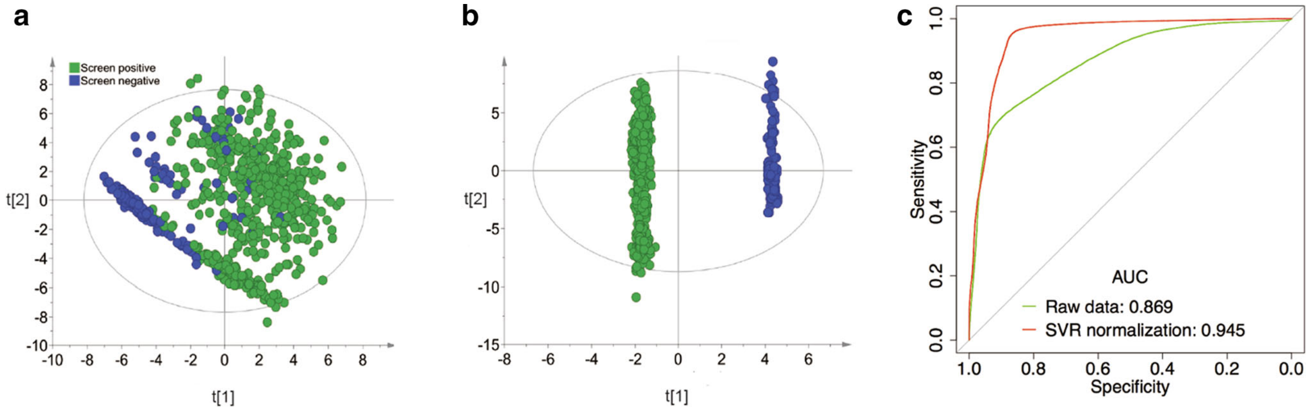

**Fig. 6** Improved classification and predictive accuracy with multivariate statistical analysis after SVR data normalization. **a**, **b** The OPLS-DA score plots using 30 metabolite ion peaks selected according to VIP values before and after SVR normalization. **c** The receiver operating characteristic curve (ROC) before and after SVR normalization

intra-batch and inter-batch variations during LC–MS analysis and enhanced the power of statistical analysis for biomarker discovery purposes. After SVR data normalization, the proportion of metabolite ion peaks with RSDs less than 30 % significantly increased to greater than 90 % of the total peaks, which is much better than other normalization methods such as linear and LOESS normalizations. Accurate selection of potential metabolite biomarkers by removing unwanted variations can improve classification accuracy using multivariate statistical analyses such as PCA and PLS-DA. The area under the curve (AUC) increased from 0.869 (raw dataset) to 0.945 (SVR normalized dataset), and this result showed that SVR normalization can significantly improve the predictive accuracy of statistical analyses.

**Acknowledgments** The work is financially supported by the funding from Interdisciplinary Research Center on Biology and Chemistry (IRCBC), Chinese Academy of Sciences (CAS), and the National Natural Science Foundation of China (Grants 21575151 and 81573246). Z.-J. Z. is supported by Thousand Youth Talents Program (The Recruitment Program of Global Youth Experts from Chinese government). This work is also partially supported by Agilent Technologies Thought Leader Award.

**Compliance with ethical standards**

**Conflict of interest** The authors declare no competing financial interest.

**Ethical approval** All institutional and national guidelines for the care and use of biological samples were followed. The data acquired were in accordance with appropriate ethical requirements.

**Research involving human participants** The human study was approved by the ethics committee of Shandong Cancer Hospital affiliated to Shandong University, Shandong Province, China.

**Informed consent** All written informed consents were obtained from all participants involved in this study.

## References

Bijlsma, S., Bobeldijk, L., Verheij, E. R., Ramaker, R., Kochhar, S., Macdonald, I. A., et al. (2006). Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Analytical Chemistry, 78*(2), 567–574.

Brereton, R. G., & Lloyd, G. R. (2010). Support vector machines for classification and regression. *Analyst, 135*(2), 230–267.

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*(2), 227–235.

Cairns, D. A., Thompson, D., Perkins, D. N., Stanley, A. J., Selby, P. J., & Banks, R. E. (2008). Proteomic profiling using mass spectrometry—does normalising by total ion current potentially mask some biological differences? *Proteomics, 8*(1), 21–27.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.

De Livera, A. M., Dias, D. A., De Souza, D., Rupasinghe, T., Pyke, J., Tull, D., et al. (2012). Normalizing and integrating metabolomics data. *Analytical Chemistry, 84*(24), 10768–10776.

De Livera, A. M., Sysi-Aho, M., Jacob, L., Gagnon-Bartsch, J. A., Castillo, S., Simpson, J. A., et al. (2015). Statistical methods for handling unwanted variation in metabolomics data. *Analytical Chemistry, 87*(7), 3606–3615.

Dunn, W. B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., et al. (2011). Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols, 6*(7), 1060–1083.

Dunn, W. B., Wilson, I. D., Nicholls, A. W., & Broadhurst, D. (2012). The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis, 4*(18), 2249–2264.

Evans, A. M., DeHaven, C. D., Barrett, T., Mitchell, M., & Milgram, E. (2009). Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Analytical Chemistry, 81*(16), 6656–6667.

FDA. (2013). *Guidance for industry, bioanalytical method validation*. Food and Drug Administration, Centre for Drug Valuation and Research (CDER).

Fiehn, O. (2002). Metabolomics—the link between genotypes and phenotypes. *Plant Molecular Biology, 48*(1–2), 155–171.

Fujarewicz, K., Jarzab, M., Eszlinger, M., Krohn, K., Paschke, R., Oczko-Wojciechowska, M., et al. (2007). A multi-gene approach to differentiate papillary thyroid carcinoma from benign lesions: gene selection using support vector machines with bootstrapping. *Endocrine-Related Cancer, 14*(3), 809–826.

Griffin, J. L., Atherton, H., Shockcor, J., & Atzori, L. (2011). Metabolomics as a tool for cardiac research. *Nature Reviews Cardiology, 8*(11), 630–643.

Guan, W., Zhou, M., Hampton, C. Y., Benigno, B. B., Walker, L. D., Gray, A., et al. (2009). Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. *BMC Bioinformatics, 10*, 259.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., & Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics, 18*(Suppl 1), 96–104.

Kamleh, M. A., Ebbels, T. M. D., Spagou, K., Masson, P., & Want, E. J. (2012). Optimizing the use of quality control samples for signal drift correction in large-scale urine metabolic profiling studies. *Analytical Chemistry, 84*(6), 2670–2677.

Kuhl, C., Tautenhahn, R., Bottcher, C., Larson, T. R., & Neumann, S. (2012). CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry, 84*(1), 283–289.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics, 11*(10), 733–739.

Long, J. Z., Cisar, J. S., Milliken, D., Niessen, S., Wang, C., Trauger, S. A., et al. (2011). Metabolomics annotates ABHD3 as a physiologic regulator of medium-chain phospholipids. *Nature Chemical Biology, 7*(11), 763–765.

Luan, H. M., Liu, L. F., Meng, N., Tang, Z., Chua, K. K., Chen, L. L., et al. (2015). LC MS-based urinary metabolite signatures in idiopathic Parkinson's disease. *Journal of Proteome Research, 14*(1), 467–478.

Lv, H. T., Palacios, G., Hartil, K., & Kurland, I. J. (2011). Advantages of tandem LC–MS for the rapid assessment of tissue-specific metabolic complexity using a pentafluorophenylpropyl stationary phase. *Journal of Proteome Research, 10*(4), 2104–2112.

Mapstone, M., Cheema, A. K., Fiandaca, M. S., Zhong, X. G., Mhyre, T. R., MacArthur, L. H., et al. (2014). Plasma phospholipids identify antecedent memory impairment in older adults. *Nature Medicine, 20*(4), 415.

Mayers, J. R., Wu, C., Clish, C. B., Kraft, P., Torrence, M. E., Fiske, B. P., et al. (2014). Elevation of circulating branched-chain amino acids is an early event in human pancreatic adenocarcinoma development. *Nature Medicine, 20*(10), 1193–1198.

Nicholson, J. K., & Lindon, J. C. (2008). Systems biology—metabonomics. *Nature, 455*(7216), 1054–1056.

Patti, G. J., Yanes, O., Shriver, L. P., Courade, J. P., Tautenhahn, R., Manchester, M., et al. (2012a). Metabolomics implicates altered sphingolipids in chronic pain of neuropathic origin. *Nature Chemical Biology, 8*(3), 232–234.

Patti, G. J., Yanes, O., & Siuzdak, G. (2012b). Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology, 13*(4), 263–269.

R Development Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria. http://www.R-project.org. Accessed 18 June 2015.

Rabinowitz, J. D., & Silhavy, T. J. (2013). Metabolite turns master regulator. *Nature, 500*(7462), 283–284.

Redestig, H., Fukushima, A., Stenlund, H., Moritz, T., Arita, M., Saito, K., et al. (2009). Compensation for systematic cross-contribution improves normalization of mass spectrometry based metabolomics data. *Analytical Chemistry, 81*(19), 7974–7980.

Ren, S., Hinzman, A. A., Kang, E. L., Szczesniak, R. D., & Lu, L. J. (2015). Computational and statistical analysis of metabolomics data. *Metabolomics, 11*(6), 1492–1513.

Rosenberg, L. H., Franzen, B., Auer, G., Lehtio, J., & Forshed, J. (2010). Multivariate meta-analysis of proteomics data from human prostate and colon tumours. *BMC Bioinformatics, 11*, 468.

Scholz, M., Gatzek, S., Sterling, A., Fiehn, O., & Selbig, J. (2004). Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics, 20*(15), 2447–2454.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry, 78*(3), 779–787.

Steinwart, I., & Christmann, A. (2008). *Support vector machines*. New York: Springer.

Sysi-Aho, M., Katajamaa, M., Yetukuri, L., & Oresic, M. (2007). Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC Bioinformatics, 8*, 93.

Tautenhahn, R., Bottcher, C., & Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics, 9*, 504.

van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics, 7*, 142.

van der Kloet, F. M., Bobeldijk, I., Verheij, E. R., & Jellema, R. H. (2009). Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping. *Journal of Proteome Research, 8*(11), 5132–5141.

Veselkov, K. A., Vingara, L. K., Masson, P., Robinette, S. L., Want, E., Li, J. V., et al. (2011). Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Analytical Chemistry, 83*(15), 5864–5872.

Wang, S. Y., Kuo, C. H., & Tseng, Y. F. J. (2013). Batch normalizer: a fast total abundance regression calibration method to simultaneously adjust batch and injection order effects in liquid chromatography/time-of-flight mass spectrometry-based metabolomics data and comparison with current calibration methods. *Analytical Chemistry, 85*(2), 1037–1046.

Wang, T. J., Larson, M. G., Vasan, R. S., Cheng, S., Rhee, E. P., McCabe, E., et al. (2011). Metabolite profiles and the risk of developing diabetes. *Nature Medicine, 17*(4), 448–453.

Wang, W. X., Zhou, H. H., Lin, H., Roy, S., Shaler, T. A., Hill, L. R., et al. (2003). Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Analytical Chemistry, 75*(18), 4818–4826.

Weiss, R. H., & Kim, K. M. (2012). Metabolomics in the study of kidney diseases. *Nature Reviews Nephrology, 8*(1), 22–33.