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

San-Yuan Wang,^{†,‡} Ching-Hua Kuo,^{‡,§,||} and Yufeng J. Tseng^{*,†,‡,§,⊥}

[†]Department of Computer Science and Information Engineering, [‡]The Metabolomics Core Laboratory, Center of Genomic Medicine, [§]School of Pharmacy, College of Medicine, ^{||}Department of Pharmacy, National Taiwan University Hospital, [⊥]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan

S Supporting Information

ABSTRACT: Metabolomics is a powerful tool for understanding phenotypes and discovering biomarkers. Combinations of multiple batches or data sets in large cross-sectional epidemiology studies are frequently utilized in metabolomics, but various systematic biases can introduce both batch and injection order effects and often require proper calibrations prior to chemometric analyses. We present a novel algorithm, Batch Normalizer, to calibrate large scale metabolomic data. Batch Normalizer utilizes a regression model with consideration of the total abundance of each sample to improve its calibration performance, and it is able to remove both batch effect and injection order effects. This calibration method was tested using liquid chromatography/time-of-flight mass spectrometry (LC/TOF-MS) chromatograms of 228 plasma samples and 23 pooled quality control (QC) samples. We evaluated the performance of Batch Normalizer by examining the distribution of relative standard deviation (RSD) for all peaks detected in the pooled QC samples, the average Pearson correlation coefficients for all peaks between any two of QC samples, and the distribution of QC samples in the scores plot of a principal component analysis (PCA). After calibration by Batch Normalizer, the number of peaks in QC samples with RSD less than 15% increased from 11 to 914, all of the QC samples were closely clustered in PCA scores plot, and the average Pearson correlation coefficients for all peaks of QC samples increased from 0.938 to 0.976. This method was compared to 7 commonly used calibration methods. We discovered that using Batch Normalizer to calibrate LC/TOF-MS data produces the best calibration results.

Metabolomics or metabonomics is known to quantitatively measure multiparametric metabolites to define pathophysiological responses in a biological system.^{1,2} Many endogenous metabolites can be quantified in a high-throughput manner using liquid chromatography/time-of-flight mass spectrometry (LC/TOF-MS).³ The epidemiological studies in metabolomics utilize large sample sizes with the LC/TOF-MS measurements often performed in batches. Batch effects can occur when measurements are affected by laboratory conditions, reagent lots, or intrabatch variability.^{4,5} Injection order effects are frequently observed in LC/TOF-MS data due to a build-up of contaminants that were not being removed during analysis of a batch of samples.⁶ When utilizing precious biofluids during long-term and large-scale metabolomics studies, it is important to have consistent measurements with the least amount of error from instrumental drift or offset, to avoid incorrect findings due to batch effects or injection order effects,⁴ and to calibrate the data by removing nonbiological systematic biases in LC/TOF-MS experiments. Nonbiological

systematic biases from instrument calibration or sample injection order are often observed in LC/TOF-MS data.

Experimentally, an internal standard (IS) calibration method is commonly used to remove nonbiological systematic biases. By selecting an IS displaying similar physicochemical properties to the metabolites, systematic batch-to-batch differences can be removed after calibrating and curve fitting with the IS. However, overlapping chromatographic peaks and ion suppression effects may introduce biases into the IS calibration method, especially in a metabolomics study.⁷ Chemicals coeluted with the IS will affect ionization efficiency; therefore, the peak intensity of ISs will be changed, resulting in imperfect calibration by the ISs. Besides, in metabolic profiling analysis, it is difficult to select ISs that can calibrate all analytes of interest.⁷ At the data analysis level, calibration methods derived from

Received: October 3, 2012

Accepted: December 13, 2012

Published: December 13, 2012

microarray, proteomics, and metabolomics studies, including centering,^{8–10} scaling,^{8,9,11} quantile,^{10,12,13} ratio-based,^{9,14,15} linear normalization,^{3,16} linear regression,^{4,10,15} and IS calibration,^{4,7,17} have been applied to remove nonbiological systematic biases. However, when the abundance variance of each peak from different batches is unequal, application of the centering method will not be sufficient.⁸ To reduce heteroscedastic effects, which occur when the variance varies between batches, centering is typically applied using a scaling method. Scaling methods can adjust the variability of an abundance by subtracting a constant value and then dividing with a scaling factor.⁸ The scaling factor usually is a statistical value calculated from all samples, such as the standard deviation of peak abundances across all samples. After dividing by the standard deviation of peak abundances across all samples, the standard deviation of calibrated peak abundances would be equal to 1. Common scaling methods, such as autoscaling,⁸ pareto scaling,¹⁸ or range scaling,¹⁹ all use a scaling factor to adjust each peak abundance for systematic variance. Autoscaling, the most often employed scaling method, uses the standard deviation of each peak of all samples as the scaling factor, whereas pareto scaling and range scaling use the square root of the standard deviation of each peak of all samples and the abundance range of each peak of all samples as the scaling factors. van den Berg et al.⁸ compared the performance of the above scaling methods for gas chromatography/mass spectrometry-based metabolomics data in detail and concluded that autoscaling and range scaling performed better than the other methods. To obtain scaling factors, the standard deviation and abundance range of each peak must be calculated from all samples. Thus, scaling methods have the general disadvantage that a recalibration must be performed between batches when obtaining a new data set.

Linear normalization methods, on the contrary, divide by a value which can be directly calculated from each sample, and thus, no recalibration is required after acquiring new data. Such normalization is usually applied to the data from each sample and comprises methods to make the data from all samples directly comparable with each other. Pluskal et al.²⁰ and Wang et al.²¹ are two examples of studies that use linear normalization methods to calibrate nonbiological systematic biases by dividing with the average of overall peak abundance, the median of peak abundance, or the abundance of a peak. The unit-norm method,²² another linear normalization method, calibrates each sample by setting the sum of the n th power of the abundances of all peaks of each sample equal to 1. Quantile normalization, which was proposed by Bolstad et al.,¹² is frequently used in microarray data analysis. In quantile normalization, the expression of an entire data set is calibrated with a reference expression distribution. Draisma et al.¹³ improved quantile normalization for LC/TOF-MS data and developed quantile equating which corrects each peak for the linear and the nonlinear differences in the distribution among each batch for metabolomics studies. Another common normalization approach is to calibrate the LC/TOF-MS data using ISs. For example, Kloet et al.⁴ utilized a single point regression method with pooled quality control (QC) samples and multiple ISs to calibrate batch and injection order effects.

In this study, we developed a novel algorithm, Batch Normalizer, which is based on the single value regression model with the total abundance information of each sample to remove nonbiological systematic biases. We used a metabolomics data set containing a total of 251 plasma samples

analyzed in 4 batches with 23 QC samples to evaluate our method. We also implemented and compared current calibration methods, that is, centering, scaling, quantile, ratio-based, linear normalization, and linear regression methods, which were derived from microarray, proteomics, and metabolomics studies, to compare and understand the benefit of each calibration method. We assessed the performance of nonbiological systematic bias removal using principal component analysis (PCA) scores plots, the average Pearson correlation coefficients for all peaks between any two of QC samples, and the number of peaks in QC samples with relative standard deviations (RSDs) less than 15%.¹⁶

■ EXPERIMENTAL SECTION

Reagents and Materials. All solvents used were of MS grade. Water and acetonitrile (ACN) were obtained from Scharlau (Sentmenat, Spain) and J. T. Baker (Phillipsburg, NJ), respectively. Reagents were analytical grade and they were obtained from Sigma-Aldrich (Steinheim, Germany).

Sample Preparation. A total of 228 plasma samples consisting of three repetitions, each at two time points, from 38 patients were obtained from the National Taiwan University Hospital. Additionally, 23 QC samples were obtained by pooling aliquots from each plasma sample. A volume of 100 μL of human plasma sample was extracted with 400 μL of methanol containing 250 ng mL^{-1} of [D_8]-phenylalanine (IS1) and [$^{15}\text{N}_2$]-theophylline (IS2) as ISs and the extraction was performed by shaking for 2 min using a Geno/Grinder 2010 (OPS Diagnostics, LLC, NJ). The extract was then centrifuged at 15 000g for 5 min at 4 °C (Eppendorf Centrifuge 5415R). The supernatant (375 μL) was collected in a new eppendorf tube, and the pellet was re-extracted using the same protocol. The plasma extracts were pooled and dried using a centrifugal vaporizer for 4 h (Thermo SpeedVacSPD111 V, Waltham, MA). The residue was reconstituted with 200 μL of 50% methanol and centrifuged at 15 000g for 5 min. The supernatant was filtered with a 0.2- μm filter (Minisart RC 4, Sartorius, Goettingen, Germany) and subjected to LC/TOF-MS analysis.

LC/TOF-MS Analysis. All of the plasma samples were analyzed using an Agilent 1290 UHPLC system coupled to a 6540-QTOF (Agilent Technologies, Santa Clara, CA). An Acquity HSS T3 column (100 mm $\times$ 2.1 mm, 1.8 μm ; Waters, Milford, MA) was used for the separation and the column was maintained at 40 °C. The mobile phase was composed of solvent A, water/0.1% formic acid and solvent B, acetonitrile/0.1% formic acid. The gradient elution program was as the following: 0–1.5 min, 2% B; 1.5–9 min, linear gradient from 2 to 50% B; 9–14 min, linear gradient from 50 to 95% B; and hold at 95% B for 3 min. The flow rate was 300 $\mu\text{L min}^{-1}$. For sample ionization, a Jet Stream electrospray ionization source was used with a capillary voltage of 4.0 kV in positive mode. The MS parameters were set as follows: gas temperature, 325 °C; gas flow, 5 L/min; nebulizer, 40 psi; sheath gas temperature, 325 °C; sheath gas flow, 10 L/min. A scan range of 50–1700 m/z was set.

Data. In total, 251 plasma samples were separated into four batches and run in a random order. In the four batches, there were 60, 60, 60, and 48 plasma samples with 6, 6, 6, and 5 QC samples, respectively.

Data Analysis. MS data were converted to mzXML format using Trapper (ISB)²³ before generating peak tables with XCMS.²⁴ The centWave²⁵ method from XCMS was used to

detect the peaks with the following parameters, ppm = 25, peakwidth = (5, 20), and snthresh = 5. Peak grouping was carried out using the peak density grouping method with the following parameters, mzwid = 0.1 and bw = 10 (5 after retention time alignment). The retention time correction parameters were as follows, missing = 1, smooth = "loess", and family = "gaussian". After the second pass of peak grouping, the fillPeaks method was performed to fill the missing peak data with the default parameters. Finally, we ignored the peaks if it is not be detected in other QC data. The number of detected peaks from each QC samples was 2789, the median number of detected peaks from each sample was 2706, and three samples from the same patient were considered as outliers due to the average number of detected peaks of each sample was less than 1500. All calibration processes and further data analysis were performed using the R statistical environment (version 2.11.1)²⁶ calling compiled C++ objects for major computing processes. Thus, all computing were basically performed in the C++ programs compiled by a GCC compiler (version 4.4.3).²⁷

WORKFLOW AND METHODS

Batch Normalizer. We assume that the behavior between any two consecutive QC samples in the same batch is linear,⁴ that the analysis time of each sample is equivalent to the injection order, and that the total abundance of all metabolites should be equal across different experiments. The total abundance of all metabolites is the sum of total intensities of all detected peaks in a LC/TOF-MS experiment. According to the above properties, we designed a total abundance regression calibration model (Batch Normalizer) with m indicator variables, where m is the number of batches (see eqs 1 and 2 and Figure 1). In eq 1, $C_{p,q}$, TA_q , and $\text{median}(TA_*)$ are the abundance of peak p of the QC sample q , the total abundance of the QC sample q , and the median of the total abundance of each sample s , respectively. Thus, $TA.\text{scale}_q$ is the ratio of the median total abundance to the total abundance of QC sample q . Furthermore, $\text{injection order}_{q,b}$ and $X_{q,b}$ are the injection order of the QC sample q in batch b , and it is set as 0 if the QC sample q is not in batch b and the indicator variable that takes on the value 1 only if the QC sample q is in batch b . Finally, $\text{base}_{p,b}$ and $\text{RF}_{p,b}$ are the parameters of this regression model that are to be estimated for each batch b .

$$\begin{aligned}
 TA.\text{scale}_q \times C_{p,q} &\sim \sum_{b=1}^m (\text{base}_{p,b} \times X_{q,b} + \text{injection order}_{q,b} \times \text{RF}_{p,b}) \\
 X_{q,b} &= \begin{cases} 1 & \text{if QC sample } q \text{ is in batch } b \\ 0 & \text{otherwise} \end{cases} \\
 TA.\text{scale}_q &= \frac{\text{median}(TA_*)}{TA_q}
 \end{aligned} \quad (1)$$

After estimating the parameters in eq 1, $\hat{C}_{p,s}$, the calibrated abundance of peak p of the sample s with the abundance $C_{p,s}$, the injection order ($\text{injection order}_{s,b}$), and the total abundances TA_s can be calculated using eq 2. In eq 2, the $\text{median}(C_{p,qc,*})$ is the median abundance of peak p across all QC samples in all batches and is used to maintain calibrated data fitting in the original distribution.

Figure 1. Batch Normalizer workflow using the total abundance (TA) scale and fitting with the regression line. The solid and open circles are the QC and study samples, respectively. Colored circles refer to different batches, that is, black for the first batch and red for the second batch.

$$\begin{aligned}
 \hat{C}_{p,s} &= \frac{C_{p,s} \times TA.\text{scale}_s \times \text{median}(C_{p,qc,*})}{\sum_{b=1}^m (\text{base}_{p,b} \times X_{s,b} + \text{injection order}_{s,b} \times \text{RF}_{p,b})} \\
 X_{s,b} &= \begin{cases} 1 & \text{if sample } s \text{ is in batch } b \\ 0 & \text{otherwise} \end{cases} \\
 TA.\text{scale}_s &= \frac{\text{median}(TA_*)}{TA_s}
 \end{aligned} \quad (2)$$

Other Calibration Methods Used in This Work. Seven commonly used calibration methods, centering, autoscaling, quantile, Ratio-A, totL₁, single value regression, and IS calibration, were selected for comparison purposes and were implemented as follows. Table 1 lists all the calibration used for comparison in this work.

Centering Calibration Method. The centering method is widely used in the microarray, proteomics, and metabolomics. It centers each peak abundance value around a constant to adjust for the effects of corresponding systematic biases. After this calibration, the mean of each peak's abundance across all samples within each batch is set to a constant. In this study, the constant was set as the median of each peak's abundance across

Table 1. List of the Calibration Methods

method	correction factor	main application field
centering	mean intensity of each peak	transcriptomics, ⁹ proteomics, ¹⁰ and metabolomics ⁸
scaling	mean and standard deviation of each peak	transcriptomics ⁹ and metabolomics ^{8,11}
quantile	intensity distribution	transcriptomics, ¹² proteomics, ¹⁰ and metabolomics ¹³
ratio-A	mean intensity of QCs	transcriptomics ^{9,14} and metabolomics ¹⁵
totL ₁	total abundance of each sample	transcriptomics, ¹⁴ metabolomics, ¹⁵
regression	regression on all QCs	proteomics ¹⁰ and metabolomics ^{4,15}
IS	intensity of the IS	proteomics ¹⁷ and metabolomics ^{4,7}

all QC samples. For a peak p of sample s in batch b , the calibrated abundance $\hat{C}_{p,s,b}$ is calculated using eq 3.

$$\hat{C}_{p,s,b} = C_{p,s,b} - \text{mean}(C_{p,*},b) + \text{median}(C_{p,qc,*}) \quad (3)$$

In eq 3, the $\text{mean}(C_{p,*},b)$ is the average abundance of peak p across all samples in batch b .

Scaling Calibration Method and Autoscaling. The scaling calibration method divides every peak abundance value by a scaling factor and sets the average abundance of peak p across all samples in batch b as zero. The scaling factor might be the standard deviation, the interquartile range, the length of the range, or the square root of the standard deviation of the abundance of peak p across all samples in batch b . However, for those data with lesser intensity values than the average intensity value, adjusted abundance values become negative and cause problems in later statistical analyses. To overcome this adjusted negative abundance value issue, we calibrate data by multiplying the corresponding scaling factor and adding the mean abundance of the QC samples in the first batch. We also apply the autoscaling method to calibrate the raw data using eqs 4 and 5, where $\text{SF}_{p,*},b$ and $\text{SF}_{p,qc,1}$ are the scaling factor of the peak p in batch b and the scaling factor of the peak p across QC samples in the first batch, respectively. We set the standard deviation of the peak abundance in batch b as $\text{SF}_{p,*},b$ and the standard deviation of the peak abundance of all QC samples in the first batch as $\text{SF}_{p,qc,1}$. $\text{Mean}(C_{p,qc,1})$ is the average abundance of peak p across all QC samples in the first batch and is used to maintain calibrated data fitting in the original distribution.

$$C'_{p,s,b} = \frac{C_{p,s,b} - \text{mean}(C_{p,*},b)}{\text{SF}_{p,*},b} \quad (4)$$

$$\hat{C}_{p,s,b} = C'_{p,s,b} \times \text{SF}_{p,qc,1} + \text{mean}(C_{p,qc,1}) \quad (5)$$

Quantile Calibration Method. The aim of the quantile calibration method is to make the peak distributions of each sample identical, in terms of statistical properties. In this study, the function “normalize.quantiles”, which is included in the “preprocessCore” package of the R statistical environment, as proposed by Bolstad et al.,¹² has been applied to all study samples. This function initially assigns each sample to a column and places the abundance values for peaks common to all samples in the same row. In the second step, it assigns an index to each peak abundance value in the column. As a third step, it sorts each column and then replaces the peak abundance values in each row with the mean of each row. Finally, it resorts the original order of the assigned indexes for each sample.

Ratio-Based Calibration Method and Ratio-A. A ratio-based calibration calibrates all samples by a reference sample. In this study, we performed the Ratio-A method and used the arithmetic mean of multiple QC samples within each batch as the reference sample. In the Ratio-A method, the calibrated abundance $\hat{C}_{p,s,b}$ is obtained using eq 6, where $\text{mean}(C_{p,qc,b})$ is the average abundance of peak p across all QC samples in batch b .

$$\hat{C}_{p,s,b} = \frac{C_{p,s,b} \times \text{median}(C_{p,qc,*})}{\text{mean}(C_{p,qc,b})} \quad (6)$$

Linear Normalizer Calibration Method and totL₁. Linear normalizer calibration is commonly used to adjust a LC/TOF-MS data set. One of the most frequently utilized linear normalizer methods is the unit-norm method, totL₁.²² This method initially scales each sample so that the sum of all peak abundances equals 1. In this study, we multiply the totL₁ by the median sum of all peak abundances across all samples, $\text{median}(\text{TA}_*)$ (in eq 7).

$$\hat{C}_{p,s} = \frac{C_{p,s} \times \text{median}(\text{TA}_*)}{\text{TA}_s} \quad (7)$$

Regression Calibration Method and SPRC. The single value regression method assumes that the behavior between two consecutive QC samples is linear, that the analysis time of each sample is the same, and that the time stamp is equivalent to the injection order.⁴ We then use eq 8 to estimate the parameters $\text{base}_{p,b}$ and $\text{RF}_{p,b}$, which are dependent on the abundance of peak p and the injection order of QC sample q in batch b , respectively. Then, the calibrated abundance of peak p of the sample s in batch b with the abundance $C_{p,s,b}$ and the injection order $\text{injection order}_{s,b}$ will be calculated using eq 9.

$$C_{p,q,b} \sim \text{base}_{p,b} + \text{injection order}_{q,b} \times \text{RF}_{p,b} \quad (8)$$

$$\hat{C}_{p,s,b} = \frac{C_{p,s,b} \times \text{median}(C_{p,qc,*})}{\text{base}_{p,b} + \text{injection order}_{s,b} \times \text{RF}_{p,b}} \quad (9)$$

IS Calibration Method. The IS calibration method divides the peak abundance of each peak by the abundance of the IS in that sample. In this study, the calibrated abundance obtained by the IS calibration method $\hat{C}_{p,s}$ is calculated using eq 10, where $\text{median}(C_{IS,*})$ and $C_{IS,s}$ are the median abundance of IS across all samples in all batches and the abundance of IS of sample s , respectively.

$$\hat{C}_{p,s} = \frac{C_{p,s} \times \text{median}(C_{IS,*})}{C_{IS,s}} \quad (10)$$

Evaluation of LC/TOF-MS Calibration Performances. A commonly adapted criterion for the measurement of the reproducibility of the analysis results is found in *Guidance on Bioanalytical Method Validation* wherein the FDA recommends for single analyte tests that tolerance limits are set so that the measured response detected in two-thirds of QC samples is within 15% of the QC mean, except for compounds with concentrations at or near the limit of quantification (LOQ).^{16,28}

In this study, we adopt this criterion and evaluate the performance of each calibration method by using (1) the number of peaks in QC samples with RSDs less than 15% where the RSD of each peak p of QC samples is calculated using eq 11 in which $\sigma_{p,qc}$ is the standard deviation of peak p

Figure 2. PCA scores plots of the original LC/TOF-MS data set (A) with PC1, 2; (B) with PC1, 3; (C) with PC2, 3; and the plots after-applying Batch Normalizer: (D) with PC1, 2; (E) with PC1, 3; (F) with PC2, 3. The solid and open circles refer to QC and study samples, respectively. Colored circles refer to different batches, that is, black for batch 1, red for batch 2, green for batch 3, and blue for batch 4.

across all QC samples; (2) the average Pearson correlation coefficients for all peaks between any two QC samples; (3) the scores plots of PCA, a commonly used method in metabolomics, for visual inspection of clustering of QC peaks.

$$RSD_p = \frac{\sigma_{p,qc}}{\text{mean}(C_{p,qc})} \quad (11)$$

PCA has frequently been applied to evaluate the similarity between different data and can be used to check the repeatability of the LC/TOF-MS data since data that are

closely related to each other (highly repeated data or injections) would cluster together in a PCA plot.^{16,29} In theory, PCA projects the chromatographic data to a new set of orthogonal variables known as principal components (PCs). These PCs are related to the original LC/TOF-MS data set because each PC is a linear combination of the original variables. For highly correlated data variables, the reduced variable dimensions of PCs are able to describe the maximum variation within the data. PCs are ranked according to the amount of variance explained in the data. Therefore, the first

PC explains the maximum amount of variance in the data. A scores plot from PCA is a three-dimensional map as the function of the first three PCs results. If there are patterns or certain groupings (classes) in the data, samples of similar patterns will be close to each other in a scores plot. PCA were performed with respect to the LC/TOF-MS data set for each batch in this study. We then calculated the average distance of each QC sample pair in the 3D PCA scores plot to determine whether the QC samples from different or same batch were clustered. The PCA scores plots of the first three principal components were generated using the function “prcomp” of the R statistical environment (version 2.11.1).

RESULTS AND DISCUSSION

Batch Normalizer Calibrates Both the Injection and Batch Effects. Batch and injection effects are severe in the original data. PC1, PC2, and PC3 in the PCA scoring plots accounted for the three topmost varying patterns in the original LC/TOF-MS data (solid circles for QC samples and open circles for non-QC samples, black for batch 1, red for batch 2, green for batch 3, and blue for batch 4) before calibration. Figures 2A–C (marked as “Raw” in the figures) display the PCA scores plots for the original LC/TOF-MS data set before calibration with respect to PC1 and PC2 (Figure 2A), PC1 and PC3 (Figure 2B), PC2 and PC3 (Figure 2C). The PC1 and PC2 scores of the QC samples within each batch are distributed in the three regions in Figure 2A. The QC samples in the first and the fourth batches are located at the top and bottom of the PCA plot, and the QC samples in the other two batches are located at the center of the PCA plot. In Figure 2B,C, the QC samples in the first batch are separated from other QC samples and the PC3 scores of QC samples in the first batch increases with injection order. The distribution of QC samples in the first batch with different PC3 scores for different injection orders displays the injection order effects. In Figure 2C, the QC samples of different batches are separated into four regions by the PC2 scores. The distinct PC2 score for each batch shows a different pattern for each batch.

The occurrence of batch and injection order effects before calibration can also be detected by investigating the RSD values of peak abundance of the QC samples and the CV of the replicate injections. The average RSDs of peak abundance for the QC samples is 54.6% with a standard deviation of 27.6%. There are only 11 peaks (out of 2789) with RSDs that fall within the 15% criteria which indicated the variation was high before calibration. Also, the appearance of significant injection order effects, which can lead to a biased statistical analysis before calibration, can also be found by comparing the replicate injections. In the study samples (not the QC samples), we calculate the coefficient of variation (CV) of each peak in the study samples from the same plasma (replicate injections). The CV of each peak of the study sample is calculated to evaluate the reproducibility of the continuous replicate injections. The CV of peak p of the study sample s is the standard deviation of peak p across the three corresponding continuous replicate samples divided by the peak abundance of peak p of the study sample s . We find the number of peaks with average CV within 15% in replicate injections of study samples (Table S-1 in the Supporting Information) is much greater than the number of peaks with RSD (Table S-2 in the Supporting Information) within 15% QC samples in each batch (15 injections in between). One of the main causes of this phenomenon is from injection order effects.

Figure 2D–F displays the data distribution after calibration using Batch Normalizer. In theory, a good calibration should exhibit the following properties: an increase in the average Pearson correlation coefficients since all QC samples should be in high correlation and all QC samples are identical; smaller variation of peak abundance, the RSD value, showing the method reduced the variations in peak abundance of the same QC samples; a great increase in the number of peaks with RSDs falling within the 15% criteria to show the method adjusts more peaks to a reasonable variation range; last, QC samples should be clustered together in all scoring plots regardless of injection order and batch, meaning they are very similar or identical. In Figure 2D–F, we can see the significant improvement of the analytical resulting after calibrating with Batch Normalizer. The average RSDs of the peak abundance among all QC samples is reduced to 24.1% with a standard deviation of 20.0%. The number of peaks with the RSDs that fall within the 15% criteria is greatly increased from 11 to 914, indicating a significant improvement in data quality. Table 2 also shows that Batch

Table 2. Average Distance (D) in the PCA Scores Plot, Average Pearson Correlation Coefficients (P) for All Peaks, and Proportion of Peaks with RSDs (pRSD) within 15% of the QC Samples, Indicated As These Values Correspond to Each Calibration Method (M)

M	D	P	pRSD (%)
before calibration	32.4	0.938	0.39
batch normalizer	8.4	0.976	32.77
SPRC	18.8	0.951	8.46
centering	32.3	0.920	0.00
autoscaling	31.5	0.918	1.36
quantile	22.6	0.955	3.19
ratio-A	23.9	0.972	3.48
totL ₁	18.8	0.938 ^a	2.69
IS1	23.2	0.938 ^a	0.43
IS2	27.4	0.938 ^a	0.04

^aPearson correlation coefficients is unaffected by linear transformations.

Normalizer produces the highest average Pearson correlation coefficients for all peaks between any two of QC samples. All the QC samples from all batches (solid circles of all colors) are clustered together at each PC plot showing injection order and batch effect being adjusted after the Batch Normalizer.

Batch Normalizer Best Improves Repeatability of the QC Samples When Compared to Other Calibration Methods. We applied each calibration method to the data and calculated the RSD of each peak across all QC samples to check whether calibration helps to increase the number of “usable” peaks within the 15% criteria described earlier. Originally, there were 11 peaks with RSDs falling within 15% in the raw data (Figure 3A). Batch Normalizer has the strongest performance overall with 914 (32.77%) of the total 2789 peaks falling below the 15% RSD threshold (Figure 3B). The peaks that cannot be adjusted by Batch Normalizer arise from nonlinear fluctuation between any two consecutive QC samples in the sample batch. One interesting observation is that the second best performing method, SPRC, is also regression based. SPRC results in a greatly improved reproducibility over other methods (Table S-2 in the Supporting Information).

Poorly performing methods were observed in Figure 3D; the centering method does not help to adjust the LC/TOF-MS

Figure 3. RSD distribution across all QC samples after each calibration method. The black bars indicate where the RSD falls within 15%. Of 2789 detected peaks, the number falling within 15% is (A) 11, (B) 914, (C) 236, (D) 0, (E) 38, (F) 89, (G) 97, (H) 75, (I) 12, and (J) 1. The gray bars indicate where the RSD exceeds 100%.

peaks back within the 15% RSDs criteria. The autoscaling method shows limited improvement, because the RSD distribution is almost identical to the original (see Figure 3E). Those two methods, in theory, might be problematic by generating negative abundance peaks, because some peaks' abundance values are less than the average abundance value. Continuing, the quantile, Ratio-A, and totL₁ methods display modest calibration ability, and the corresponding peak numbers with RSDs less than 15% criteria are 89, 97, and 75, respectively

(Figure 3F,G,H). The totL₁ method that scales each sample to make the sum of all peak abundances equal to a constant would fail to scale in the case of excessive abundance of certain peaks.

The IS calibration method did not help to calibrate the LC/TOF-MS data (Table S-2 in the Supporting Information). The number of peaks with RSDs falling within 15% is less than in the original data set after applying IS calibration (see Figure 3I,J and also Tables S-1 and S-2 in the Supporting Information). Since metabolites that coelute with the ISs would affect their ionization efficiency, the peak intensity of the ISs will be affected by individual variance of the coeluted metabolites. This might be the reason that the data quality does not show improvement using the IS calibration method.

Figure 4 shows the performance of each calibration method by the cumulative RSD curve of QC samples. Those cumulative

Figure 4. Performance of each calibration method presented by the cumulative RSD curve of QC samples.

RSD curves indicate that Batch Normalizer reduces the variances due to batch and/or injection effects by reducing the variations of QC peaks, and that it outperforms the other calibration methods. After applying the Batch Normalizer method, the proportion of peaks that satisfy the FDA 15% criteria for measuring sample repeatability is increased about 85 times (Table 2). Furthermore, the distribution of peaks with RSDs within 15% grouped by level of peak abundances shows that the significant increase in number of peaks in QC samples with RSD within 15% was from those of high abundances (Figure S-2 in the Supporting Information).

Batch Normalizer Can Best Cluster all QC Samples in PCA Scores Plots. PCA scores plots could be used to display the repeatability and similarity of the QC samples. Highly similar samples will form a tight cluster in a PCA scores plot. Figure 5A shows that the QC samples of each batch in the data before calibration (marked as raw) are separated into four parts in the 3D PCA score plot, showing that the four batches are considered different. To be more precise, the PC3 scores of the QC samples before calibration increase with the injection order, demonstrating significant injection order effects. The four distinct values of PC2 scores of the QC samples also indicate batch effects (see Figure S-1 in the Supporting Information). Batch Normalizer reduced these effects. The PCA scores plot indicates that the QC samples are more aggregated after calibration by Batch Normalizer (Figure 5B and also Figure S-

Figure 5. PCA scores plots obtained before and after applying each calibration method. The solid and open circles refer to QC and study samples, respectively. Colored circles refer to different batches, that is, black for batch 1, red for batch 2, green for batch 3, and blue for batch 4.

1B in the Supporting Information for each PC). Among all the methods, only Batch Normalizer can efficiently adjust batch and injection order effects. Batch Normalizer also calibrates the QC samples to a shortest average distance in the PCA scores plot and generates the highest average Pearson correlation coefficient for all peaks (Table 2).

Some good aggregation results of PCA scores plots can be observed by calibrating with SPRC (Figure 5C) and Ratio-A (Figure 5G) methods. However, considering PC1 scores of QC samples, poor calibration is achieved on the most explained variation, PC1, (not within the ± 1 standard deviation region, see the gray lines in each subfigures of Figure S-1B in the Supporting Information) after applying SPRC (Figure S-1C in the Supporting Information) and Ratio-A (Figure S-1G in the Supporting Information). The fluctuation of PC1 scores of the QC samples indicates that SPRC and Ratio-A calibration methods only partially mitigate batch and injection order effects. The PCA results obtained when calibrating with the centering (Figure 5D) and autoscaling (Figure 5E) methods are very similar to the original data (Figure 5A), and the average distances of QC samples in the PCA scores plots are almost equal to the raw data's (Table 2). Hence, the centering and

autoscaling methods may not improve the repeatability and similarity of QC samples; moreover, they may reduce the repeatability and similarity of QC samples due to reduction of the average Pearson correlation coefficient for all peaks of the QC samples to 0.92 and 0.918 (Table 2). The quantile (Figure 5F) and totL_1 (Figure 5H) methods display an efficient calibration within each batch, that is, the QC samples within each batch are clustered; however, the QC samples from each batch are separately clustered in the PCA scores. Thus, the quantile method can only calibrate injection order effects (Figure S-1F in the Supporting Information). Figure S-1I,J shows that the QC samples from each batch are separately clustered in the PCA scores plots after applying the IS calibration method, and Figure S-1I,J also shows high fluctuation in the PC1 scores of QC samples. Hence, the IS calibration method cannot efficiently calibrate batch and injection order effects in this data set.

Table 2 also shows that Batch Normalizer produces the highest average Pearson correlation coefficients for all peaks between any two QC samples. Though the Ratio-A calibration method also produces high average Pearson correlation

coefficients, it results in only a small proportion of peaks with RSDs within 15%.

In general, two regression methods, Batch Normalizer and SPRC, outperformed the other methods for calibrating batch and injection order effects, though at the expense of computing time. Moreover, Batch Normalizer performed much faster than SPRC, especially when applied to larger sample numbers and batches. Normal regression methods such as SPRC require $np \times nb$ computation time, where np and nb are the number of peaks and the number of batches, respectively. Computational complexity will significantly increase with respect to the number of batches. In this demo case, the total execution time was 15.09 s (Batch Normalizer) and 24.28 s (SPRC) when applying those two regression calibration methods on a personal computer with an Intel Core i7-2600 processor and 16 GB RAM separately. The real computing time of Batch Normalizer is 62% of SPRC due to the $(np-1)$ indicator variables used in the algorithm. Each indicator variable records the sample and the batch pairs in each regression model; therefore, Batch Normalizer is faster by performing np times regression. In addition to its decreased computational cost, Batch Normalizer considers the total abundance of each sample to improve its performance and is able to remove injection order effects.

CONCLUSIONS

The batch and injection order effects in metabolomic analyses can lead to severe problems in later chemometric analyses, post LC–MS measurements. In this study, we proposed an algorithm to remove batch and injection order effects, Batch Normalizer, to reduce biases due to batch and injection order effects in LC/TOF-MS data. Batch Normalizer uses the total abundance of each sample and a linear regression model to calibrate both batch and injection order effects and has been demonstrated to effectively improve the reproducibility of peak abundance of QC samples. Furthermore, the similarity of QC samples can be improved and QC samples can be closely clustered in a PCA scores plot after Batch Normalizer treatment. Compared with current batch and injection order effects removal algorithms, the proposed Batch Normalizer method provides the best calibration results.

ASSOCIATED CONTENT

Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*Phone: +886.2.3366.4888 ext. 529. Fax: +886.2.23628167. E-mail: yjtseng@csie.ntu.edu.tw.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was funded by the Taiwan National Science Council, Grant Numbers NSC99-2627-B002-018, NSC99-2321-B-002-042, 100-2627-B-002-016-, and NSC100-2325-B-002-041. Resources of the National Taiwan University's Laboratory of Computational Molecular Design and Metabolomics, Metabolomics Core Laboratory, and Department of Computer

Science and Information Engineering were used in performing these studies.

REFERENCES

- (1) Robertson, D. G. *Toxicol. Sci.* **2005**, *85*, 809–822.
- (2) Weljie, A. M.; Newton, J.; Mercier, P.; Carlson, E.; Slupsky, C. M. *Anal. Chem.* **2006**, *78*, 4430–4442.
- (3) Castillo, S.; Gopalacharyulu, P.; Yetukuri, L.; Orešič, M. *Chemom. Intell. Lab. Syst.* **2011**, *108*, 23–32.
- (4) van der Kloet, F. M.; Bobeldijk, I.; Verheij, E. R.; Jellema, R. H. *J. Proteome Res.* **2009**, *8*, 5132–5141.
- (5) Leek, J. T.; Scharpf, R. B.; Bravo, H. C.; Simcha, D.; Langmead, B.; Johnson, W. E.; Geman, D.; Baggerly, K.; Irizarry, R. A. *Nat. Rev. Genet.* **2010**, *11*, 733–739.
- (6) Lai, L.; Michopoulos, F.; Gika, H.; Theodoridis, G.; Wilkinson, R. W.; Odedra, R.; Wingate, J.; Bonner, R.; Tate, S.; Wilson, I. D. *Mol. Biosyst.* **2010**, *6*, 108–120.
- (7) Redestig, H.; Fukushima, A.; Stenlund, H.; Moritz, T.; Arita, M.; Saito, K.; Kusano, M. *Anal. Chem.* **2009**, *81*, 7974–7980.
- (8) van den Berg, R. A.; Hoefsloot, H. C.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. *BMC Genomics* **2006**, *7*, 142.
- (9) Luo, J.; Schumacher, M.; Scherer, A.; Sanoudou, D.; Megherbi, D.; Davison, T.; Shi, T.; Tong, W.; Shi, L.; Hong, H.; Zhao, C.; Elloumi, F.; Shi, W.; Thomas, R.; Lin, S.; Tillinghast, G.; Liu, G.; Zhou, Y.; Herman, D.; Li, Y.; Deng, Y.; Fang, H.; Bushel, P.; Woods, M.; Zhang, J. *Pharmacogenomics J.* **2010**, *10*, 278–291.
- (10) Callister, S. J.; Barry, R. C.; Adkins, J. N.; Johnson, E. T.; Qian, W.-j.; Webb-Robertson, B.-J. M.; Smith, R. D.; Lipton, M. S. *J. Proteome Res.* **2006**, *5*, 277–286.
- (11) Xia, J.; Psychogios, N.; Young, N.; Wishart, D. S. *Nucleic Acids Res.* **2009**, *37*, W652–660.
- (12) Bolstad, B. M.; Irizarry, R. A.; Astrand, M.; Speed, T. P. *Bioinformatics* **2003**, *19*, 185–193.
- (13) Draisma, H. H. M.; Reijmers, T. H.; van der Kloet, F.; Bobeldijk-Pastorova, I.; Spies-Faber, E.; Vogels, J. T. W. E.; Meulman, J. J.; Boomsma, D. I.; van der Greef, J.; Hankemeier, T. *Anal. Chem.* **2010**, *82*, 1039–1046.
- (14) Quackenbush, J. *Nat. Genet.* **2002**, *32* (Suppl.), 496–501.
- (15) Kamleh, M. A.; Ebbels, T. M. D.; Spagou, K.; Masson, P.; Want, E. J. *Anal. Chem.* **2012**, *84*, 2670–2677.
- (16) Zelena, E.; Dunn, W. B.; Broadhurst, D.; Francis-McIntyre, S.; Carroll, K. M.; Begley, P.; O'Hagan, S.; Knowles, J. D.; Halsall, A.; Wilson, I. D.; Kell, D. B. *Anal. Chem.* **2009**, *81*, 1357–1364.
- (17) Yan, Y.; Weaver, V. M.; Blair, I. A. *J. Proteome Res.* **2005**, *4*, 2007–2014.
- (18) Isao, N. *J. Mol. Struct.* **2008**, *883–884*, 216–227.
- (19) Smilde, A. K.; van der Werf, M. J.; Bijlsma, S.; van der Werff-van der Vat, B. J. C.; Jellema, R. H. *Anal. Chem.* **2005**, *77*, 6729–6736.
- (20) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinf.* **2010**, *11*, 395.
- (21) Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T. A.; Hill, L. R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. H. *Anal. Chem.* **2003**, *75*, 4818–4826.
- (22) Crawford, L. R.; Morrison, J. D. *Anal. Chem.* **1968**, *40*, 1464–1469.
- (23) Keller, A.; Eng, J.; Zhang, N.; Li, X. J.; Aebersold, R. *Mol. Syst. Biol.* **2005**, *1*, 2005–0017.
- (24) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779–787.
- (25) Tautenhahn, R.; Bottcher, C.; Neumann, S. *BMC Bioinf.* **2008**, *9*, 504.
- (26) R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, <http://www.R-project.org>, 2010.
- (27) Richard, M. S. *Using the GNU Compiler Collection*. Boston, <http://gcc.gnu.org>, 2010.
- (28) FDA. *Guidance for Industry, Bioanalytical Method Validation*, Food and Drug Administration, Centre for Drug Valuation and Research (CDER), May 2001.

(29) Wong, M. C.; Lee, W. T.; Wong, J. S.; Frost, G.; Lodge, J. J. *Chromatogr. B* **2008**, *871*, 341–348.