Talanta 130 (2014) 442–448

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

Detection of batch effects in liquid chromatography-mass spectrometry metabolomic data using guided principal component analysis

J. Kuligowski [a,1], D. Pérez-Guaita [b,1], I. Lliso [a], J. Escobar [a], Z. León [c], L. Gombau [d], R. Solberg [e], O.D. Saugstad [e], M. Vento [a,f], G. Quintás [d,*]

[a] Neonatal Research Unit, Health Research Institute La Fe, Valencia, Spain
[b] Department of Analytical Chemistry, University of Valencia, Burjassot, Spain
[c] Analytical Unit, Health Research Institute La Fe, Valencia, Spain
[d] Leitat Technological Center, Bio In Vitro Division, Valencia, Spain
[e] Department of Pediatric Research, Institute for Surgical Research, Oslo University Hospital – Rikshospitalet, Oslo, Norway
[f] Division of Neonatology, University & Polytechnic Hospital La Fe, Valencia, Spain

ARTICLE INFO

Article history:
Received 29 April 2014
Received in revised form
8 July 2014
Accepted 10 July 2014
Available online 18 July 2014

Keywords:
Liquid chromatography-mass spectrometry (LC-MS)
Metabolomics
Guided principal component analysis
Batch effect

ABSTRACT

Metabolomics based on liquid chromatography-mass spectrometry (LC-MS) is a powerful tool for studying dynamic responses of biological systems to different physiological or pathological conditions. Differences in the instrumental response within and between batches introduce unwanted and uncontrolled data variation that should be removed to extract useful information. This work exploits a recently developed method for the identification of batch effects in high throughput genomic data based on the calculation of a δ statistic through principal component analysis (PCA) and guided PCA. Its applicability to LC-MS metabolomic data was tested on two real examples. The first example involved the repeated analysis of 42 plasma samples and 6 blanks in three independent batches, and the second data set involved the analysis of 101 plasma and 18 blank samples in a single batch with a total runtime of 50 h. The first and second data set were used to evaluate between and within-batch effects using the δ statistic, respectively. Results obtained showed the usefulness of using the δ statistic together with other approaches such as summary statistics of peak intensity distributions, PCA scores plots or the monitoring of IS peak intensities, to detect and identify instrumental instabilities in LC-MS.

1. Introduction

Metabolomics is the comprehensive and simultaneous determination of metabolite concentrations from the metabolome and their changes over time as a consequence of stimuli. Considering the outstanding complexity of the metabolome, the use of ultra-performance liquid chromatography - mass spectrometry (UPLC-MS) is gaining importance in metabolomics due to its increased sensitivity and high throughput as compared to other techniques such as nuclear magnetic resonance (NMR) or gas chromatography-MS (GC–MS) [1].

UPLC-MS analysis in metabolomic studies are frequently performed in batches if they are run over long periods of time, involve a high number of samples and also if the study is carried out across different laboratories or instruments. Therefore, as in other high-throughput techniques used for example to assay gene or protein expressions, experimental data might include additional unwanted non-biological variation derived from laboratory or instrumental conditions [2–4]. Even under repeatability conditions, instrumental variation in UPLC-MS might arise from a number of sources including drifts in sensitivity, ionization efficiency, and gradual changes in column performance over short analysis periods (tens of injections) [5]. This type of batch effects leads to increased variability and decreased power to detect biologically meaningful responses [3]. Consequently, batch effects have to be reliably detected and eventually removed to avoid impact on repeatability and reproducibility of results across independent studies.

Normalization in metabolomics is used to remove the systematic variation unrelated to the biological difference among samples. Data normalization strategies to overcome batch effects can be

[*] Corresponding author. Tel.: +34 93 788 23 00.
E-mail address: guillermo.r.quintas@uv.es (G. Quintás).
[1] Both authors contributed equally to this work.

http://dx.doi.org/10.1016/j.talanta.2014.07.031
0039-9140/© 2014 Elsevier B.V. All rights reserved.

clustered in two groups [6]: (i) statistical models based on scaling factors calculated using complete data sets, such as normalization by 1-norm [7], median value [8] or quantile [9], and (ii) normalization using spiked internal standards (IS) with similar physicochemical properties as the analytes of interest (IS-norm) [10].

Detecting the presence of batch effects as well as the assessment of the performance of normalization methods can be accomplished in a number of ways. A simple method is the use of summary statistics of peak intensity distributions of pooled samples used as quality controls (QCs), such as the average Pearson correlation coefficients for all peaks between any two QC samples [11] or the distribution of RSD% in QCs. Alternatively, the visual inspection of principal component analysis (PCA) scores plots is a common practice [11]. However, PCA is an unsupervised method and batch effects can easily remain undetected if it they are not the largest source of variability in the data. Recently, Reese et al. [12] proposed a δ statistic based on PCA and guided PCA (gPCA) for the statistical evaluation of batch effects in multivariate data sets and demonstrated its applicability for high throughput genomic data using both, real and simulated data. In this study, we evaluate the usefulness of the δ statistic for the identification of between- and within-batch effects in metabolomic UPLC-MS datasets. To verify its applicability, the method is tested on real data. The first data set used was acquired in the frame of a study aiming at the identification of plasma biomarkers for the diagnosis of gastric cancer occurring via the sequence of molecular events known as Correa's Cascade (i.e. acute gastritis > chronic gastritis > precancerous lesions > gastric cancer). This data set was obtained from the repeated analysis of 42 plasma samples and 6 blanks in three independent batches. The second data set employed for evaluating the use of the δ statistic was obtained from a study of the effect of hypoxia and resuscitation in a piglet model involving the analysis of 101 plasma and 18 blank samples in a single batch with a total runtime of 50 h. The first and second data set were used to evaluate between and within-batch effects using the δ statistic, respectively. Results obtained showed the usefulness of the δ statistic to determine whether a between-batch effect exists in UPLC-MS data sets and also to detect and identify instrumental instabilities during a single batch measurement.

2. Batch effect estimation using PCA and guided PCA

Principal component analysis (PCA) is a widely used unsupervised method to detect batch effects in metabolomics. However, if batch effects are not the largest source of variability in the data, these effects can be easily overlooked. Recently, Reese et al. [12] developed a δ statistic to detect and quantify the significance of batch effects, based on PCA and an extension of PCA, namely guided PCA (gPCA). For a detailed description of the method see reference [12]. Nevertheless, it is worth paying attention to the basics of the method.

The δ statistic is defined as the ratio of the variance of the first principal component calculated using guided-PCA to the variance of the first principal component calculated from PCA:

$$\delta = \frac{var(XV_{g1})}{var(XV_{u1})}$$

where X $(n \times p)$ is the experimental data set with n samples and p variables, $V_{u1}(p \times p)$ is the matrix of right singular vectors calculated by singular value decomposition (SVD) of matrix X from 'unguided' PCA and $V_{g1}(b \times b)$ is the matrix of right singular vectors calculated by SVD of the product $(Y'X)$ from gPCA. Here, the matrix Y $(n \times b)$ is an indicator matrix with elements $y_{ik} = 1$ if sample i is in batch k, otherwise $y_{ik} = 0$. As shown in [12], large singular values in gPCA indicate that the batch effect is important for the corresponding principal component and so, high values of δ (i.e. close to 1) are indicative of a batch effect. The statistical significance of the batch effect can be estimated by a permutation test in which the rows of the Y matrix are randomly permuted M times (in this work, $M = 1000$). For each permutation, a δ statistic is calculated (δ_p). Then, the δ value calculated using the real batch ordering is compared to the reference distribution of δ_{perm} values and a one-sided p-value is estimated as the proportion of times that the δ statistic is in the extreme tail of the reference null distribution.

$$\delta = \frac{\sum_{m=1}^{M} (\hat{\delta} < \hat{\delta}_p)}{M}$$

This procedure was slightly modified for the evaluation of changes in the instrument performance within a single batch. In this case, the batch integrated by N samples is artificially split into k smaller subsets (k-fold) of N/k contiguous samples and a δ statistic, based on local PCA and gPCA models of sample subsets, is calculated.

Data analysis was performed using MATLAB 2012b (The Mathworks, Natick, USA), the PLS Toolbox 7.0 (Eigenvector Res.Inc., Wenatchee, USA) and in house written MATLAB scripts. Data (.netCDF,.mzXML and.mat files) and MATLAB scripts included in this work are available from the authors. The calculation of δ was performed according to Reese et al. [12] and the gPCA R package available via CRAN (http://cran.r-project.org/web/packages/gPCA/).

3. Results and discussion

3.1. Between-batch effect evaluation

Collection, pretreatment, storage and analysis of plasma samples of Data set I are described in the Supplementary Material. Briefly, data set I involved the analysis of a single set of plasma samples collected from one individual on the same day to minimize biological variation and facilitate the batch effect detection. After plasma collection, the samples were aliquoted and kept at −80 °C for 1, 3 and 7 days until analysis by UPLC-ESI(+)–TOF-MS in batches 1, 2 and 3, respectively. The sample ordering between batches was replicated to facilitate the identification of instrumental batch effects. Between batches, the ESI/MS detector inlet interface was cleaned and the MS was calibrated.

First, the three batches were initially compared using: (i) the number of features detected; (ii) the distribution of RSD% values in QC samples and the number of peaks with RSD $\leq$ 15%; (iii) the absolute values and stability of the intensities of the internal standards (ISs), and (iv) the presence of sample clustering or trends in scores plots from a PCA model calculated using the whole sample set.

After peak detection and alignment of the UPLC-MS data a total of 4294, 4534 and 3815 features were detected in batches 1, 2 and 3, respectively. As these figures were inflated by contaminants, non-relevant features with mean intensity values in blanks higher than 10% of the mean value in samples were removed, leaving a total of 2600, 2745 and 2582 in batches 1, 2 and 3, respectively. Then, variables were grouped among batches using the 'nearest' method with the following parameters, mzVsRT = 1 and RT and m/z tolerances of 5 s and 2 mDa, respectively. The number of common features was comparable among batches (see Fig. 1A). Fig. 1B also showed stable and comparable intensity profiles for the ISs PheAla-D$_5$, LeuEnk and Reserpine in the three batches.

The set of 1186 variables detected in the three batches was retained for further analysis. Highly comparable RSD% distributions and percentages of variables with RSD $\leq$ 15% (66.6, 60.0 and 60.7% for batches 1, 2 and 3, respectively) (see Fig. 1C). In spite of that, an exploratory analysis by PCA revealed a clear clustering as

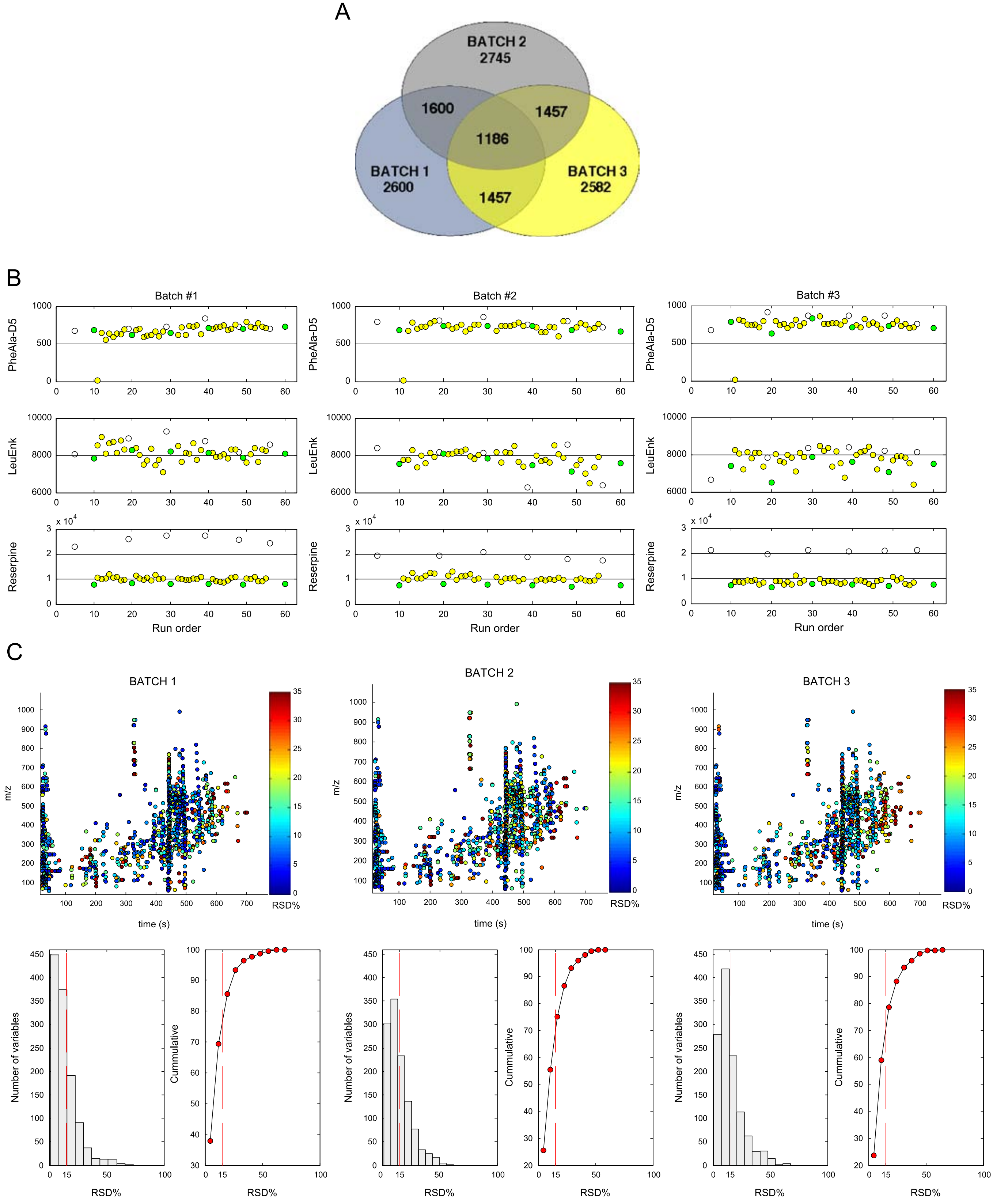

Fig. 1. Initial analysis of batch effects for data set I. (A) Venn diagram showing the number of variables detected; (B) intensities of the internal standards PheAla-D_5, LeuEnk and Reserpine (bottom) in batches 1, 2 and 3 of data set I. Yellow, white and green circles indicate plasma, blank and QC samples, respectively; (C) Distribution of features and RSD% in QC samples in batches 1, 2 and 3 of data set I. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

shown in Fig. 2A. The normalization of the peak intensities by 1-norm resulted only in slight changes in the PC scores clustering (see Fig. 2C). By using IS-norm, samples from batches 1 and 2, measured on days 1 and 3, were tightly clustered and clearly separated along PC1 from a second cluster of batch 3 samples, measured on day 7 (see Fig. 2D). On the contrary, Fig. 2B shows

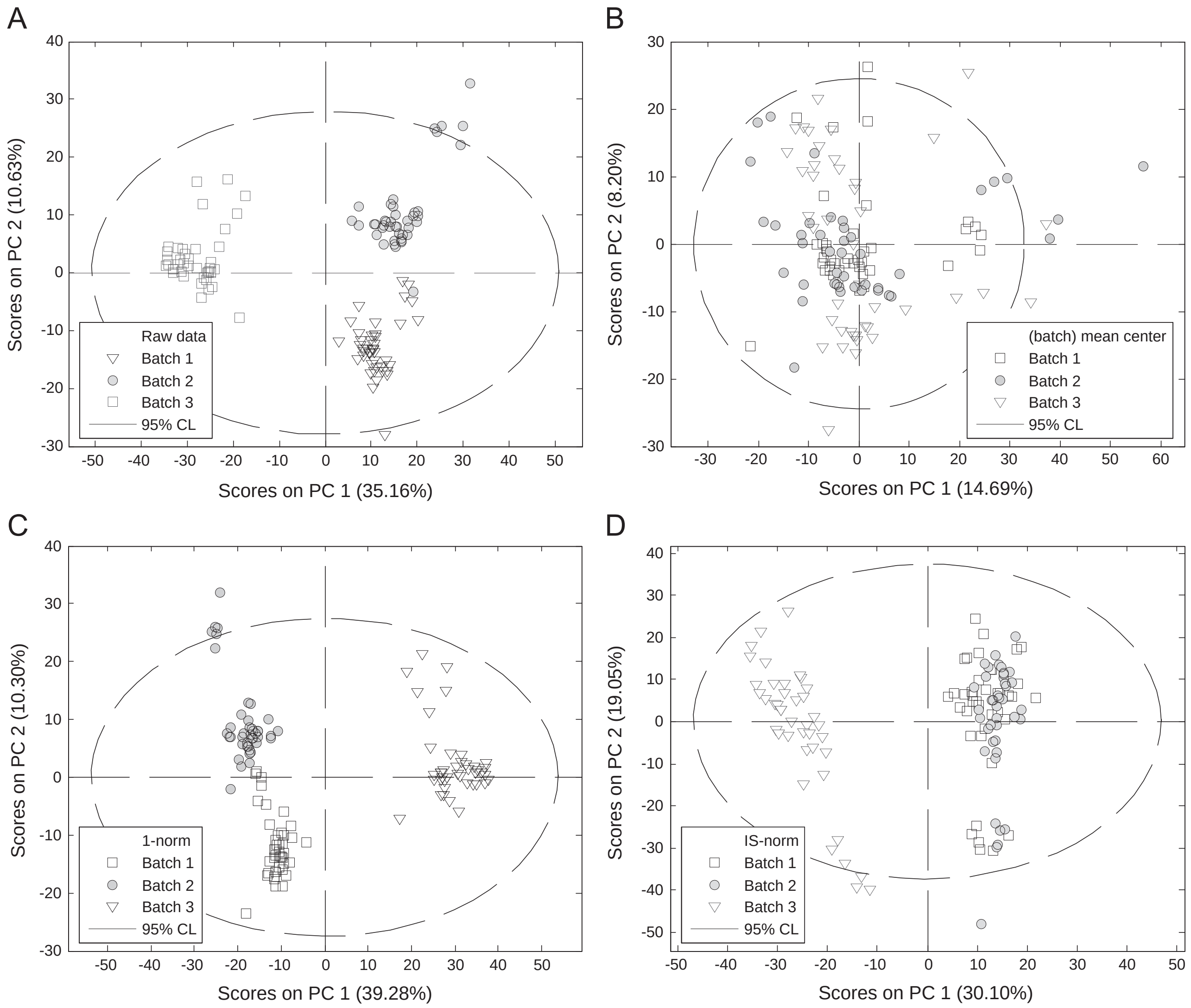

Fig. 2. PCA scores plots obtained using raw (A), mean centered (B) 1-norm (C) and IS normalized (D) data set I.

that prior batch mean centering successfully removed the batch clustering.

Results from visual inspection of the PCA scores plot were in good agreement with those found using the δ statistic (see Fig. 3). Whilst a statistically significant batch effect was found using raw data (δ=0.994, p-value < 0.001), mean centering of each batch prior to PCA decreased the δ value and their statistical significance (δ=0.006, p-value > 0.05).

Then, a pair-wise analysis among the three batches was carried out. The δ statistic between batches 1 and 2 calculated using raw data (δ=0.957) indicated a somewhat low batch effect difficult to identify from the PCA scores plot in Fig. 2A. After IS-norm, the δ=0.669 (p-value > 0.05) obtained between batches 1 and 2 was also in good agreement with the observed clustering in Fig. 2D, where samples from batches 1 and 2 were clearly separated from batch 3 samples. Results showed that the use of the δ statistic is useful to quantify the statistical significance of batch effects in UPLC-MS.

3.2. Within-batch effect evaluation

Data set II comprised a set of plasma samples (n=85), QCs (n=16) and blanks (n=18), collected within a study of the effect of hypoxia and resucistation in a piglet model. This data set was selected to evaluate the use of δ to detect instrumental effects within a single batch. Detailed description of the collection, pretreatment, storage and analysis procedures followed for the analysis of samples are included in the Supplementary Material. After peak detection and alignment, a total of 471 variables were retained for further analysis.

Signal stability was initially analyzed using the intensities of the ISs PheAla-D_5 and Methionine-D_3 and PCA scores plots to detect trends and sample clustering within the batch. The plot of the IS intensities showed a decreasing trend and allowed the identification of a set of 11 outlying samples that were removed from the data set (see Fig. 4). The PC1, PC2 and PC3 scores plot of the PCA obtained using pareto scaling as data pretreatment, is shown in Fig. 5(A). The decreasing trend over time observed for PC1 and PC2 scores plots was indicative of an instrumental effect that was analyzed using the δ statistic. Accordingly, the data set was split into k sample subsets (k={3, 4, 6, 9}) and each subset was then considered as an independent (sub)batch. From results summarized in Table 1, the statistical significance of the batch effect in raw data was confirmed (δ p-value$\ll$0.05) for all the considered k-fold splits.

An evaluation of batch normalization methods was out of the scope of this work. However, three common methods were applied to test the usefulness of the δ statistic to compare their

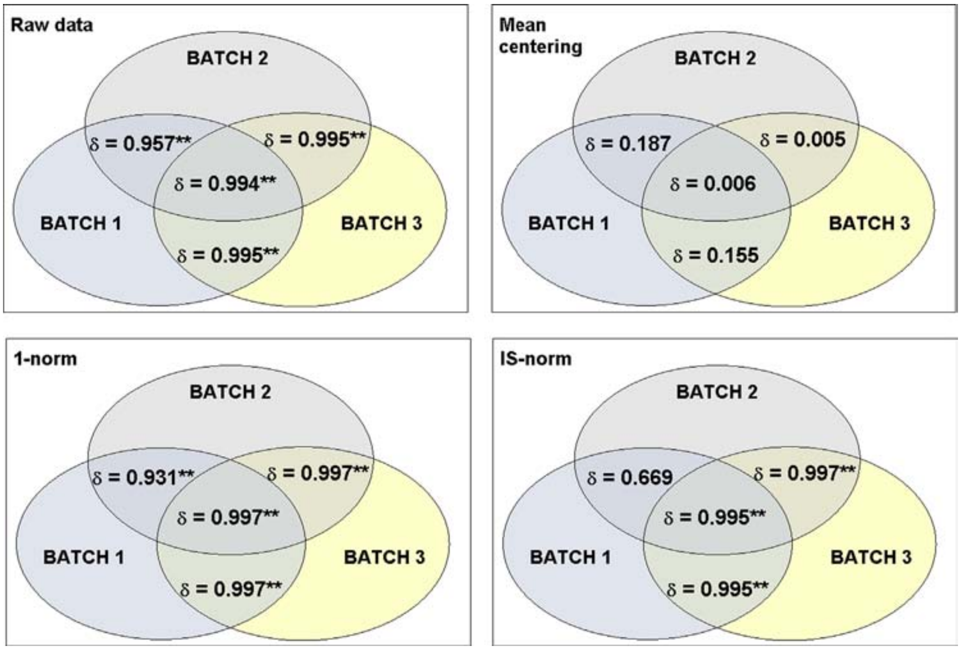

Fig. 3. Venn diagrams showing the δ values calculated for data set I using raw data and after using mean centering, 1-norm and IS-norm as data pretreatment. Note: **indicates δ *p*-values < 0.001.

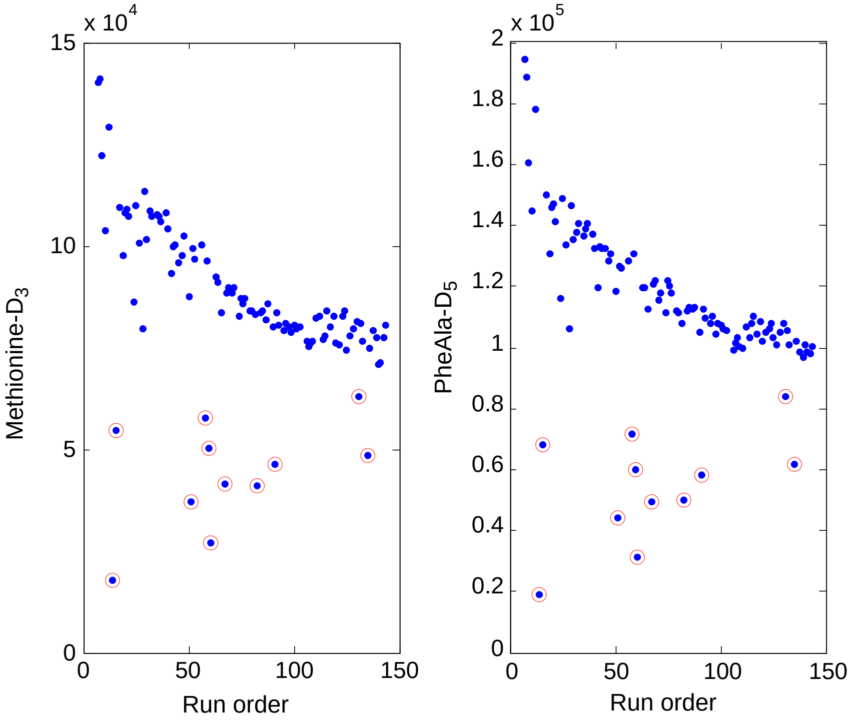

Fig. 4. Intensity of ISs in data set II. Red circles indicate samples classified as outliers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

performance. The first one (1-norm) divides all the peak intensities by the sum of all intensities in the sample. The second (IS-norm_A) and the third (IS-norm_B) method are based on the use of the intensities of the ISs. In IS-norm_A, Methionine-D_3 (RT=155 s) and Phenylalanine-D_5 (RT=295 s) were used to normalize variables eluting at RT≪225 s and RT > 225 s, respectively. In IS-norm_B, variables with RT < 120 s were excluded from normalization. Fig. 5 (B–D) shows the variation of the PC1, PC2 and PC3 scores obtained by PCA after data normalization. While it appears that the decreasing trend in the scores was removed in a great extent using all three normalization methods, it was difficult to conclude which one provided a better effect removal. However, from results summarized in Table 1 using the δ statistic, it can be seen that although normalization using both, 1-norm and IS-norm_A partially removed the effect, as shown by the decrease in the δ values, IS-norm_B normalization was the only approach

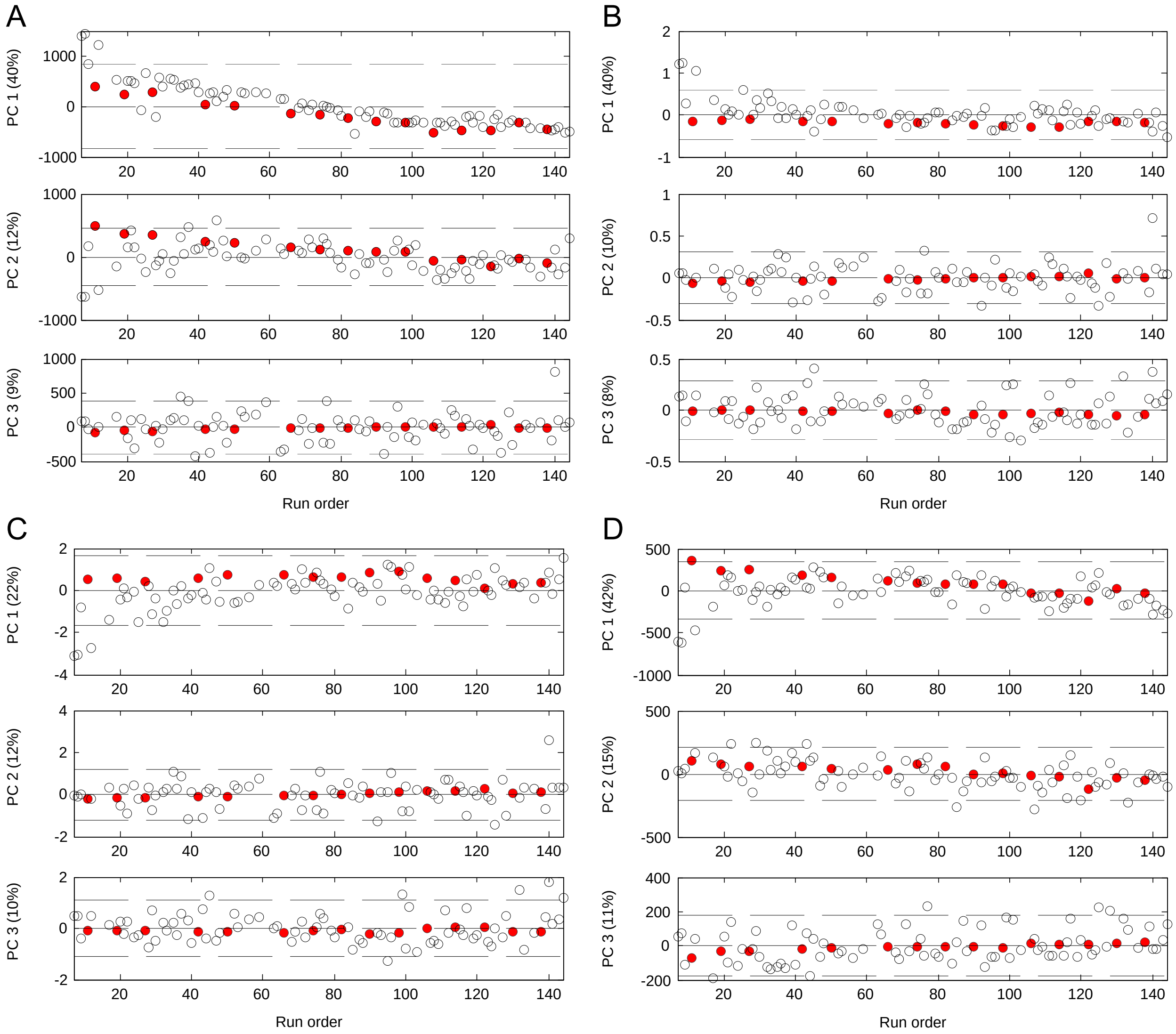

Fig. 5. Analysis of batch effects in data set II using PCA scores from raw data (A) and after normalization: 1-norm (B) IS-normA (C) and IS-normB (D). Note: Red circles indicate QC samples. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
δ values calculated for data sets II before and after data normalization and using different k-fold splits of the data set.

	***k*-fold**			
Normalization	**3**	**4**	**6**	**9**
Raw data	0.959 (0.001)	0.975 (0.001)	0.969 (0.002)	0.979 (0.002)
1-norm	0.875 (0.015)	0.938 (0.001)	0.937 (0.013)	0.961 (0.006)
IS- normA	0.781 (0.038)	0.866 (0.003)	0.856 (0.031)	0.905 (0.019)
IS-normB	0.574 (0.39)	0.726 (0.36)	0.669 (0.68)	0.824 (0.57)

providing a statistically significant removal of the effect (δ p-value≥0.05). The use of IS-norm assumes that variance exhibited by the IS is caused exclusively by systematic error [13]. However, ion suppression might be very different between front and late eluting peaks and so, the use of Methionine-D$_3$ for normalization of peaks eluting at RT < 120 s may introduce bias. The relation between the RT and the value of the δ statistic was evaluated by calculating the δ statistic in subsets of variables clustered in RT windows of 100 s. Results depicted in Fig. 6 show that 1-norm and IS-norm_A increased the value of the δ statistic in the RT window between 50 s and 150 s. Moreover, IS-norm_B performed better in the 150–250 s intervals, where the ISs elute supporting the conclusion that the poor results obtained using IS-norm_A were due to an improper selection of the IS used for normalization. This shows that the δ statistic can also be used to gain further insights into the variables responsible for the local batch effect and so, it could be used to optimize the normalization conditions.

4. Conclusions

Batch effects among UPLC-MS metabolomic data sets are very frequent and, if they are not recognized or if they are not properly removed, they can make it almost impossible to separate batch and biologically relevant effects. Results found on real data showed that the δ statistic can be used to estimate the statistical significance of batch effects and to compare the performance of normalization methods in LC-MS metabolomic data sets. The analysis in RT windows and smaller sample subsets in the case

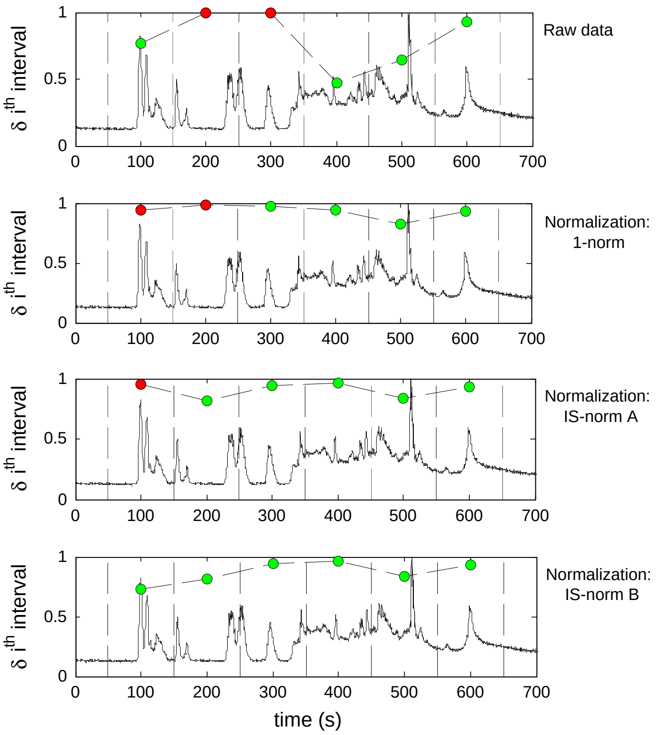

Fig. 6. Analysis of within-batch effect in data set II as a function of RT using (from top to bottom) raw, 1-norm, IS-norm_A and IS-norm_B data. *Note*: Vertical dotted lines indicate the limits of the RT windows used for the calculation of the within-batch δ values Red dots indicate RT windows showing statistically significant batch effect. δ was calculated using a 4-fold batch split. The chromatogram (black solid line) is a randomly selected TIC of a QC sample. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of single batch analysis can provide a better insight into the source of batch effects. Although the power of the approach for detecting batch effects depends on the ratio of variables affected and the size of the effect, this statistic can be a useful tool to be used together with other approaches such as summary statistics of peak intensity distributions, PCA scores plots and the monitoring of IS peak intensities.

Acknowledgments

JE and JK acknowledge the “Sara Borrell” grants (CD11/00154 and CD12/00667) from the *Instituto Carlos III* Ministry of Economy and Competitiveness. IL is grateful for the financial support from the Laeerdal Foundation. MV acknowledges the FISPI11/0313 grant from the *Instituto Carlos III* (Ministry of Economy and Competitiveness). We acknowledge financial support from the Research Network on Maternal and Child Health and Development *Retic Red SAMID* RD12/0022/12 (*Instituto Carlos III,* Ministry of Economy and Competitiveness, Spain). GQ acknowledges financial support from the Spanish Ministry of Economy and Competitiveness (SAF2012-39948).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.talanta.2014.07.031.

References

[1] T.O. Metz, Q. Zhang, J.S. Page, Y. Shen, S.J. Callister, J.M. Jacobs, R.D. Smith, Biomark. Med. 1 (2007) 159–185.

[2] F.M. van der Kloet, I. Bobeldijk, E.R. Verheij, R.H. Jellema, J. Proteome Res. 8 (2009) 5132–5141.

[3] J.T. Leek, R.B. Scharpf, H.C. Bravo, D. Simcha, B. Langmead, W.E. Johnson, D. Geman, K. Baggerly, R.A. Irizarry, Nat. Rev. Genet. 11 (2010) 733–739.

[4] L. Lai, F. Michopoulos, H. Gika, G. Theodoridis, R.W. Wilkinson, R. Odedra, J. Wingate, R. Bonner, S. Tate, I.D. Wilson, Mol. Biosyst. 6 (2010) 108.

[5] W.B. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, M. Brown, J.D. Knowles, A. Halsall, J.N. Haselden, A.W. Nicholls, I.D. Wilson, D.B. Kell, R. Goodacre, Nat. Protoc. 6 (2011) 1060–1083.

[6] M. Sysi-Aho, M. Katajamaa, L. Yetukuri, M. Orešič, BMC. Bioinform. 8 (2007) 93.

[7] M. Scholz, S. Gatzek, A. Sterling, O. Fiehn, J. Selbig, Bioinform. Oxf. Engl. 20 (2004) 2447–2454.

[8] W. Wang, H. Zhou, H. Lin, S. Roy, T.A. Shaler, L.R. Hill, S. Norton, P. Kumar, M. Anderle, C.H. Becker, Anal. Chem. 75 (2003) 4818–4826.

[9] J. Lee, J. Park, M. Lim, S.J. Seong, J.J. Seo, S.M. Park, H.W. Lee, Y.-R. Yoon, Anal. Sci. Int. J. Jpn. Soc. Anal. Chem. 28 (2012) 801–805.

[10] S. Bijlsma, I. Bobeldijk, E.R. Verheij, R. Ramaker, S. Kochhar, I.A. Macdonald, B. van Ommen, A.K. Smilde, Anal. Chem. 78 (2006) 567–574.

[11] S.-Y. Wang, C.-H. Kuo, Y.J. Tseng, Anal. Chem. 85 (2013) 1037–1046.

[12] S.E. Reese, K.J. Archer, T.M. Therneau, E.J. Atkinson, C.M. Vachon, M. de Andrade, J.-P.A. Kocher, J.E. Eckel-Passow, Bioinformatics 29 (2013) 2877–2883.

[13] H. Redestig, A. Fukushima, H. Stenlund, T. Moritz, M. Arita, K. Saito, M. Kusano, Anal. Chem. 81 (2009) 7974–7980.