BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 20 2014, pages 2899–2905
doi:10.1093/bioinformatics/btu423

Gene expression

Advance Access publication July 2, 2014

Intensity drift removal in LC/MS metabolomics by common variance compensation

Francesc Fernández-Albert[1,2,*], Rafael Llorach[2,3], Mar Garcia-Aloy[2,3], Andrey Ziyatdinov[1], Cristina Andres-Lacueva[2,3] and Alexandre Perera[1]

[1]Department d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, CIBER-BBN, [2]Biomarkers & Nutrimetabolomic Lab., Department of Nutrition and Food Science-XaRTA, INSA, Faculty of Pharmacy, Food and Nutrition Torribera Campus, University of Barcelona, Av. Prat de la Riba 171, 08921, Sta Coloma de Gramenet and [3]INGENIO-CONSOLIDER Program, FUN-C-Food CSD2007-063, Av Joan XXIII s/n 08028, Barcelona, Spain

Associate Editor: Ziv Bar-Joseph

ABSTRACT

Liquid chromatography coupled to mass spectrometry (LC/MS) has become widely used in Metabolomics. Several artefacts have been identified during the acquisition step in large LC/MS metabolomics experiments, including ion suppression, carryover or changes in the sensitivity and intensity. Several sources have been pointed out as responsible for these effects. In this context, the drift effects of the peak intensity is one of the most frequent and may even constitute the main source of variance in the data, resulting in misleading statistical results when the samples are analysed. In this article, we propose the introduction of a methodology based on a common variance analysis before the data normalization to address this issue. This methodology was tested and compared with four other methods by calculating the Dunn and Silhouette indices of the quality control classes. The results showed that our proposed methodology performed better than any of the other four methods. As far as we know, this is the first time that this kind of approach has been applied in the metabolomics context.

Availability and implementation: The source code of the methods is available as the R package intCor at http://b2slab.upc.edu/software-and-downloads/intensity-drift-correction/.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 11, 2014; revised on June 20, 2014; accepted on June 25, 2014

1 INTRODUCTION

Metabolomics aims to asses the metabolic changes in a global way to infer biological functions and provide the detailed biochemical responses of cellular systems (Fiehn *et al.*, 2006). Liquid chromatography/mass spectrometry (LC/MS) devices are among the most used experimental set-ups in metabolomics. LC/MS analyses of biological samples such as urine or plasma give high-throughput data having a three index scheme: retention time, mass/charge ratio and intensity values (Aihua *et al.*, 2012; Xin *et al.*, 2008). In metabolomic data, as in other types of high-dimensional data such as gas sensor arrays or microarray data, the intensity values of the variables might be biased or might suffer from variations owing to external factors. Among these factors is a contribution from the drift of the experimental devices, owing to various causes such as column ageing in the case of LC/MS, temperature variations or contamination effects (Burton *et al.*, 2008; Wang *et al.*, 2013). The presence of peak intensity drift in the data is an important issue, as its effects can be important enough to mask the real statistical behaviour of the data and may indeed be the largest source of variance in the data (Veselkov *et al.*, 2011; Wang *et al.*, 2013).

In most LC/MS protocols, quality control (QC) samples are regularly injected to ensure good analytical device performance (Dunn *et al.*, 2011). In LC/MS metabolomics studies, the quality controls have been carried out using pools of biological samples, spikes with standards or Milli-Q water samples (Llorach *et al.*, 2009). These quality control samples consist either of a pooling of all the samples in the study or of a spike-in of some known metabolites (several classes having different types of QC samples might be injected). In the data preprocessing stage, one may distinguish two different steps: data normalization and data equalization. We understand the data normalization step as the mathematical process that makes the *variables* in the dataset comparable, whereas the data equalization step makes the *samples* from the dataset comparable. In the literature, many normalization and equalization methods, based on several different approaches and scopes, may be found. Regarding equalization methods, a methodology using certain internal known metabolites as quality standards to normalize the whole dataset has been reported (Sysi-Aho *et al.*, 2007). Another approach is to use the injected samples for internal control (i.e. QCs) to fit a smoothed model for the intensity levels of certain features, and then to correct all the biological samples accordingly (Dunn *et al.*, 2011). The R package sva includes the ComBat function, which compensates the batch effects on microarray data using an empirical Bayes approach (Johnson *et al.*, 2007; Leek *et al.*, 2012). This method has been applied to normalize gene expression and methylation data (Chen *et al.*, 2013; Leitch *et al.*, 2013). Equalization methods based on a sample-wise correction for LC/MS metabolomic data have also been tested and compared by Veselkov *et al.* (2011). Their results suggest that a variance stabilization transformation of the data, followed by a median fold change normalization, gives the best performance as compared

*To whom correspondence should be addressed.

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

with three other methods. Their method performs a normalization and an equalization step to give a robust output when having urine samples with different concentration values. Among the equalization methods, the one proposed by Artursson *et al*., based on component correction (CC), was developed in the sensor-array field (Artursson *et al*., 2000). This method is based on the assumption that, in multivariate data, the drift direction is the first principal component (PC) of a PCA decomposition for a class consisting of measurements of the same samples. Such samples are known as technical replicates (i.e. there is no biological or chemical variation in addition to the variability of the technical replication of the measure). Once the drift direction is computed, the drift is removed from the data by subtracting the data projection on the drift direction from the original data. However, if some between-class variability is aligned with the drift direction, it will also be subtracted, and some non-drift variability will be removed. A natural extension of the CC method is the one proposed by Ziyatdinov *et al*., which is based on a common principal component analysis (CPCA) decomposition (Ziyatdinov *et al*., 2010). This method proposes modelling the drift contribution in the data as the direction capturing maximum variance that simultaneously diagonalizes the covariance matrices of a set of classes. All the variability of the samples in that particular direction is considered to be drift-induced variability, and the projection of the data on that direction is subtracted from the data as in the CC method.

In this article, to find the drift model, we state the hypothesis that the intensity drift of the chromatograms is the common variance direction of all the QC classes that captures the maximum variance. In this context, we propose a preprocessing method based on a two-step approach by first equalizing the data through a CPCA, and then normalizing the data using a median fold change step.

2 MATERIALS AND METHODS

2.1 Description of the data

The samples were analysed by liquid chromatography coupled with a hybrid quadrupole time-of-flight (LC-q-TOF, Hybrid quadrupole TOF QSTAR Elite, AB/MDS Sciex) in positive mode using the protocol proposed by Tulipani *et al*. (2011). LC was performed in High-Performance Liquid Chromatography (HPLC) Agilent (Agilent 1200 Series Rapid Resolution HPLC system) using an RP 18 Luna column (50 × 2.0 mm, 5 m), with a sample injection volume of 15 μl. A linear gradient elution was performed consisting of [A] Milli-Q water 0.1% HCOOH (v/v) and [B] acetonitrile 0.1% HCOOH (v/v).The gradient elution (v/v) of [B] was as follows: (time, min; B, %): (0, 1), (4, 20), (6, 95), (7.5, 95),(8, 1), (12, 1). Q-TOF spray parameters were set as previously described (Llorach *et al*., 2009), and full data acquisition was performed scanning from 70 to 700 m/z. The TOF was calibrated with reserpine (1 pmol/μl). LC–MS data were acquired in random order to avoid possible bias and the batches equilibrated. Throughout all the analysis, data process quality control (QC) samples were analysed to monitor the stability and functionality of the system. The sample collecting span was of 18 days, and there was a replacement of the chromatographic column in the process on day 14. There were 994 study samples and 182 QC samples. Three classes of QC samples were used for each batch:

- Water: Milli-Q water samples (n = 96 samples).
- Spikes: Standard mixture solution (n = 48 samples) consisting of 12 metabolites at the final concentration of 5 ppm for all of them except for indole-3-acetic-2,2-d2 acid whose final concentration was of 10 ppm.
- Reference: Urine sample belonging to the one volunteer (n = 38 samples).

2.2 Preprocessing

All the methods were applied to the chromatograms without any prior feature detection. The R package XCMS was used to read the chromatograms of the mzXML files containing the sample data (Smith *et al*., 2006). The chromatograms were aligned using an in-house-developed R package (UB/UPC). The chromatographic data of all the files read were merged, creating an $n \times m$ chromatogram matrix X. This step required the binning of the retention time in m bins that were given by the XCMS package. Therefore, the chromatogram matrix had samples as rows and retention time as columns (in our case, n = 1176 samples and m = 441 retention time points). Thus, the i-th row of this matrix corresponds to the chromatogram of the i-th sample. From here on, the variable j refers to the retention time bins in the chromatogram matrix. A class-wise outlier detection and removal procedure was applied to the QC classes. This procedure was based on computing the score distance (SD) and an orthogonal distance (OD) in a PCA model using the pcaPP R package (Filzmoser *et al*., 2013; Hubert *et al*., 2005). QC samples having SD and OD distances greater than the suggested critical values were considered to be outliers and were discarded from the dataset (Filzmoser and Fritz, 2007). The critical values used in the package were (i) a quantile of the chi-squared distribution for the SD and (ii) a Wilson–Hilferty approximation for the scaled chi-squared distribution for the OD (Filzmoser and Fritz, 2007; Hubert *et al*., 2005). Using this approach, nine outlier samples were detected (four samples in class *reference*, three in class *water* and two in class *spikes*). As it is known that raw LC/MS metabolomic data suffer from multiplicative noise, we took the logarithm of the data to compensate for such error sources and to convert them into additive noise sources (Veselkov *et al*., 2011). Once the o outliers were removed, we could then define the quantity $p = n - o$ to be the new sample range. The resulting matrix $Y(p \times m)$ was used as an input parameter for all the normalization methods tested. This matrix contains the data for both the QC classes and the study class, and it can be divided into matrices corresponding to each class (i.e. $Y_{QC}(p_{QC} \times m)$ for the corresponding dataset for all QC classes, $Y_r(p_r \times m)$ for the corresponding dataset for the reference class, etc.).

2.3 Methods

The five methods compared in this article (CPCA, CC, median fold change, ComBat and our CPCA + median fold change) have different input parameters. The methods based on a CPCA decomposition or the CC method involve a class (or classes)-selection step to use them for the drift modelling. These methods also need as input the number of components of the drift decomposition, which are supposed to be captured. The ComBat method needs the batch relation for each class, whereas the median fold change method does not need any specific input parameter in addition to the data to be normalized.

2.3.1 Component correction The hypothesis underlying this method is that the drift direction is found in the first PC of a reference class. The methodology used to normalize these data is described in (Artursson *et al*., 2000). As the feature pattern of the QC samples was more complex than that of the other two QC classes, the reference class was selected to generate the PCA model. Because of this higher complexity, this class is better able to capture the drift in the data than would a class with a simpler feature pattern. Mathematically, the CC method can be expressed

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

as in Equation (1).

$$Y_r = S \cdot L^T + E \quad (1)$$

The methodology proposed by Artursson *et al.* removes one PC, but the method can be generalized to remove as many PCs as can be found in the data. If *Ncomps* is the number of components to be removed, then S ($p_r \times Ncomps$) is the scores matrix, L ($m \times Ncomps$) is the loadings matrix and E ($p_r \times m$) is the error matrix.

As only one PC is required to perform the normalization, no further increase in the dimensions of the matrices S and L is necessary. The drift direction is the first PC, which in this case corresponds to the matrix L. The next step is to project the dataset Y onto this direction to obtain the drift component in the data. This projection is mathematically expressed as the Equation (2).

$$Y_d^{cc} = (Y \cdot L) \cdot L^T \quad (2)$$

where the superscript *cc* refers to CC and the subscript *d* to drift. Once the drift component Y_d^{cc} ($p_r \times m$) of the data is computed, the last step in removing the drift is to subtract it from the data. Equation (3) shows this last step, and the resulting matrix Z^{cc} ($p_r \times m$) is the corrected matrix using the CC method.

$$Z^{cc} = Y - Y_d^{cc} \quad (3)$$

2.3.2 Median fold change The median fold change method is not focused on finding the drift direction. Its objective is to rescale the data to make the median fold changes of the variables close to zero. The methodology followed in applying this method is the one of Veselkov *et al.*, based on a sample-wise approach (Veselkov *et al.*, 2011). The first step of this method, shown in equation (4), is to compute the median for each variable, thus obtaining a vector $\hat{y}_i(1 \times m)$. This vector is used to rescale the original dataset Y into a new one, $\hat{Y}(p \times m)$ IG(IG) [see Equation (4)].

$$\hat{Y}_{ij} = \frac{Y_{ij}}{\hat{y}_i} \text{ where } \hat{y}_i = median_i(Y_{ij}) \quad (4)$$

To obtain the normalized dataset $Z^M(p \times m)$, the dataset Y is divided by the sample median of the matrix $\hat{Y}$ (defined as $\hat{w}_j(p \times 1)$) as shown in Equation (5).

$$Z_{ij}^M = \frac{Y_{ij}}{\hat{w}_j} \text{ where } \hat{w}_j = median_j(\hat{Y}_{ij}) \quad (5)$$

where the superscript M refers to median fold change method.

2.3.3 ComBat The ComBat method is a function of the R package sva. This function aims to correct the batch effects, which are known to be a source of bias, in gene expression experiments; its extension to LC/MS metabolomic datasets is both natural and straightforward. First, it is assumed that batch effects have multiplicative and additive contributions to the data, and that these effects can be variable dependent (gene or peak, respectively). We state a model following this hypothesis [see Equation (6)]

$$Y_{ijb} = \alpha_j + X\beta_j + \gamma_{jb} + \delta_{jb}\epsilon_{ijb} \quad (6)$$

where Y_{ijb} is the intensity value for sample i, variable j and batch b. α_j is the intensity value for variable j, X is the design matrix, β_j contains the regression coefficients of the model, γ_{jb} is a matrix containing the additive batch effects for variable j and batch b, δ_{jb} is a matrix containing the multiplicative batch effects for variable j and batch b and ϵ_{ijb} is the residual matrix of the model. Using either a parametric or a non-parametric empirical prior estimation, the distributions for γ_{jb} and δ_{jb}^2 are estimated. The conditional posterior probabilities (γ_{jb}^* and δ_{jb}^{2*}) can then be found, and the data are corrected for batch effects as shown in Equation (7). In the following, all the variables having a hat (ˆ) on them refer to their values estimated from the data.

$$Z_{ijb}^{CB} = \frac{\hat{\sigma}_j}{\hat{\delta}_{ij}^*}(Z_{ijb} - \hat{\gamma_{jb}}^*) + \hat{\alpha}_j + X\hat{\beta}_j \quad (7)$$

where the superscript *CB* refers to the ComBat function, Z_{ijb} is the standardized data and $\hat{\sigma}_j$ is the estimated standard deviation.

2.3.4 CPCA CPCA is a generalization of the PCA decomposition for different classes first introduced by Flury *et al.* (Flury, 1984). Say we have k classes and Σ_k ($p_k \times p_k$) are the set of their covariance matrices, then CPCA aims at finding a space such as the one defined by the V (in general, $p_k \times p_k$) matrix shown in Equation (8). In the space spanned by V, the covariance matrices for all the classes involved Σ_k are diagonal.

$$\Lambda_k = V^T \cdot \Sigma_k \cdot V \quad (8)$$

where Λ_k ($p_k \times p_k$) is the diagonalized covariance matrix for class k. Each one of the dimensions of this space is called a common principal component (CPC). The hypothesis underlying the CPCA method for drift correction is that the drift direction is contained in the CPC capturing the largest variance. The CPC will be computed by using the Y_{QC} dataset [i.e. there are three expressions like Equation (8), using the different covariance matrices for the QC classes: Σ_r, Σ_{water}, Σ_{spikes}]. In a similar way as in a PCA decomposition, given the desired number of CPCs and following a stepwise algorithm, it is possible to compute the number of CPCs one by one (Trendafilov, 2010). Setting the desired number of CPCs as *Ncomps*, the dimensionality of the V matrix is ($p_k \times Ncomps$). We have tested the values *Ncomps* = 1, 2, 3 separately for this method.

Once the CPCs are found, the dataset is projected onto this space as shown in (9)

$$Y_d^{CPCA} = (Y \cdot V) \cdot V^T \quad (9)$$

Y_d^{CPCA} ($n \times p$) contains the drift component in the data. To eliminate the drift from the data, the last step is to subtract this drift from the data [Equation (10)]

$$Z^{CPCA} = Y - Y_d^{CPCA} \quad (10)$$

where Z^{CPCA} ($n \times p$) is the corrected dataset using the CPCA method.

2.3.5 CPCA + median fold change The method we propose consists of a two-step approach. First, the data are equalized by removing the drift using CPCA and, in the second step, the data are normalized by applying the median fold change method. As the CPCA method was applied three times with different number of extracted CPCs (Ncomps in previous subsection), the proposed method will be computed for the same number of components (Ncomps = 1, 2, 3).

2.4 Validation

From the class definition in Section 2.1, it follows that a PCA score plot of all the classes should have the classes clearly separated in different clusters. We propose a quality measure for peak intensity drift correction methods based on the standard clustering internal measures Dunn and Silhouette for the QC classes in the principal plane (the plane explaining maximum variability of the data) score plot of all the classes (including the study class). The clustering technique used was k-means. The R package clValid was used to compute the quality indices (Brock *et al.*, 2008, 2011). In general, the greater the Dunn and Silhouette indices, the better the clustering, meaning that the QC classes are more easily separable in the principal plane and that the intra-class variance is lower.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

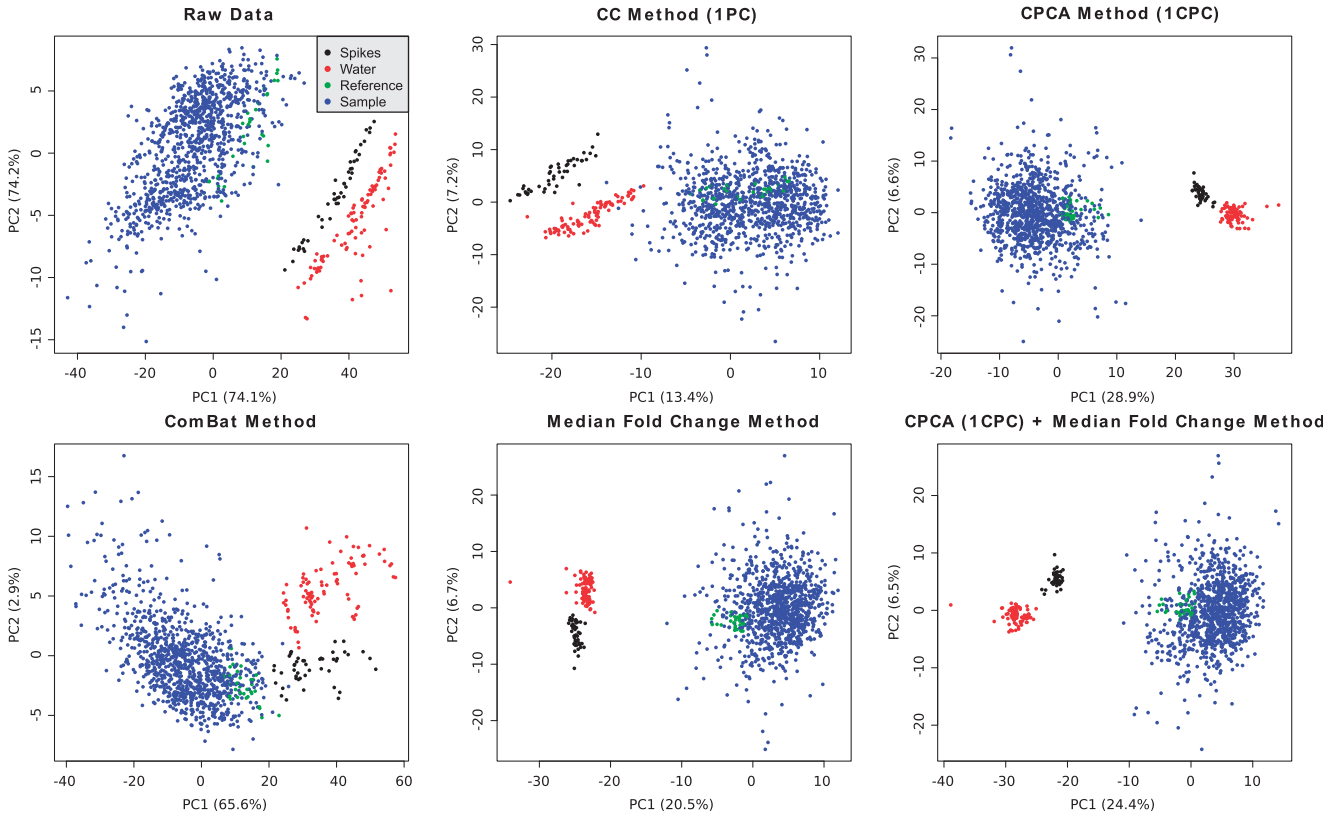

Fig. 1. Set of PCA Scoreplots showing the raw data and the effect on data for each method. The class labelled as sample is the study class. The numbers in brackets on the axes of the plots refer to the estimated variance for that particular direction in the data

Table 1. Dunn and Silhouette values for all the tested methods

Method/index	Dunn	Silhouette
None	0.029	0.560
CPCA (1CPC)	0.159	0.749
CPCA (2CPC)	0.129	0.725
CPCA (3CPC)	0.051	0.680
ComBat	0.074	0.553
CC (1PC)	0.182	0.573
CC (2PC)	0.249	0.600
CC (3PC)	0.209	0.566
Median fold change	0.171	0.719
CPCA (1CPC) + Median	0.208	**0.794**
CPCA (2CPC) + Median	**0.344**	0.690
CPCA (3CPC) + Median	0.101	0.631

Notes: The CPCA and CC methods were tested removing one, two and three components (the number in brackets refers to the components subtracted from the data). The last three entries of the table correspond to sequentially using the CPCA with one CPC and then the median fold change method. The highest clustering indices are shown in bold.

3 RESULTS AND DISCUSSION

The top left plot in Figure 1 depicts a PCA score plot for the raw data using all the classes. The Figure shows that one of the main sources of variance is the interclass variability. The drift effect on the variance becomes clear when the same PCA score plots are coloured according to the injection time or the batch of the sample. This effect is shown in Figure 2. From this Figure, we conclude that there is a clear drift component (having different sources) that is causing an important drift of the QC classes and which, in all likelihood, affects the samples in the study class as well.

Table 1 contains the Dunn and Silhouette values for all the methods used, whereas Figure 1 depicts the PCA Scoreplots for the same methods. The CPCA + median fold change method shows the highest clustering values (highest Dunn index when two components are removed and highest Silhouette index when one component is removed) and it has a slight advantage over the CPCA and the median fold change methods.

From the mathematical formulation of the Dunn index, it is important to note that its measures might give low values when there are a small number of samples some distance from the cluster centre, even if all the other samples and classes are tightly clustered (Dunn, 1973). This might be the case here, as applying the CC method removing two PCs had a higher Dunn index than the CPCA + median fold change removing one CPC, when the PCA score plots showed a lower drift clustering for the latter (compare Supplementary Fig. S1 versus the bottom right plot of Fig. 1).

The Silhouette index (Table 1) for the different CPCA methods applied suggests that the drift seems to be contained in just the first CPC, as the quality measures go down as more CPCs are removed from the data. Figure 3 depicts the origin

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

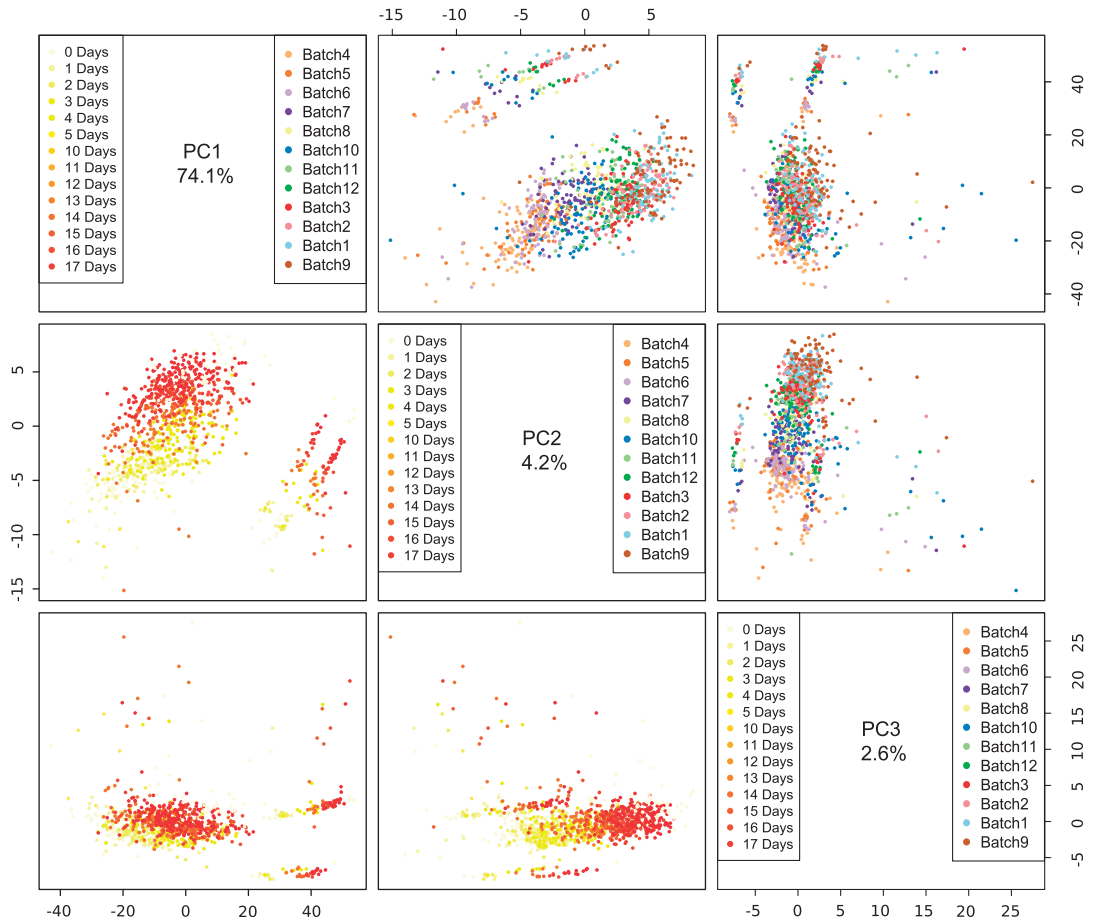

Fig. 2. PCA Scoreplot of all the classes in the data. The three plots in the lower left show the effect of the time elapsed since the first sample was injected, whereas the plots in the top right refer to the batch of the sample. The numbers in the diagonal plots correspond to the variance captured by each PC. The order in the legend of the batches corresponds to the real injection order of the samples

of the variance removed by each of the three CPCs and the mean chromatogram of the QC samples. Whereas the first CPC shows a smooth variation along the retention time, the second and third CPCs show abrupt changes in their values, especially close to major chromatographic peaks. In some of these peaks, the values of the second and third CPCs have a zigzag behaviour, suddenly going from negative to positive values or vice versa. This behaviour suggests that the variance captured by these two CPCs corresponds to not having the chromatograms perfectly aligned, meaning that the peaks in the chromatograms are not fully coincident across the samples. This fact further reinforces the hypothesis that the drift is contained in just the first CPC. Assuming this hypothesis, it can be noted from Figure 3 that the drift of the data is not higher close to the chromatographic peaks, apparently quite the opposite, meaning that the drift affects the baselines of the chromatograms more than their peaks.

The CC method corrects some of the drift in the data although a large drift component is still to be found in the data (Fig. 1). The larger Dunn index value for the CC method as compared with the raw data value is evidence for the drift correction (Table 1). However, this improvement is not validated by the Silhouette index, which remains practically unchanged as compared with the raw Silhouette value regardless of the number of components removed.

The ComBat method does not seem to be the most suitable method for correcting LC/MS metabolomic data despite being used widely and successfully in the field of gene expression and methylation data. Although it corrects some batch effects in the study samples, the batch effects are still important in the QC classes after the correction (see circles A–G for some *spikes* and *water* samples compared with the H circle containing some study samples in Supplementary Fig. S2). Furthermore, the resultant PCA score plot (lower left plot of Fig. 1) for the ComBat correction suggests that some between-variance component was removed in the correction process, as the different classes are closer than for the raw data.

The median fold change method considerably improves both the Dunn and the Silhouette indices. A visual inspection of the resulting PCA score plot for the median fold change method confirms this improvement (Fig. 1). Nevertheless, the PCA score plot also shows that the *spikes* and *water* classes have similar shapes and these long shapes turn out to be caused by residual uncorrected drift effects (Supplementary Fig. S3). This fact suggests that, as the median fold change method normalizes the data without specifically trying to remove the drift, there may

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

still be a source of variance in the data caused by the drift of the experimental device. On the other hand, because the methods based in the CPCA approach (CPCA and CPCA + median fold change methods) are developed to model the drift direction, their resultant datasets show less residual drift in their corresponding principal plane score plot (Supplementary Figs. S4 and S5 for the CPCA and the CPCA + median fold change methods, respectively).

The methods have been characterised for different dataset sizes (10, 25, 33 and 50% of the dataset) through a random subsampling stage (taking 100 iterations per subsample). For each subset, we applied the proposed methodology to test the compensation effect given dataset size and to find an estimate on the variability of this effect. We computed the Silhouette index for each drift-correction method as described in Supplementary Section S1. Figure 4 depicts the Silhouette values for all the methods and dataset sizes. The Figure shows that the CPCA + Medians method has the highest mean Silhouette values for all the tested values. To measure how the performance of the methods is modified as function of the dataset size, we have fitted a linear model using size as a cofactor and computed an ANalysis Of VAriance (ANOVA) test. Supplementary Table S1 contains the values of the slopes, the standard errors and the *P*-values of the ANOVA tests for all the methods. Results show that the median fold change and the CC methods suffer from performance for large datasets. On the other hand, the ComBat method improves its correction as data availability increases. This last result is probably because of a better estimation of the batch components when having larger sample sizes. The Supplementary Table S1 also shows that the performance of the CPCA and CPCA + Medians methods is insensitive to the considered dataset size. This suggests that the mathematical approach taken, where the drift component is extracted from a multi-class QC variance analysis is more resilient to the sample size variations.

Overall, in the context of LC/MS drift correction, the proposed two-step methodology shows better clustering properties of the QC samples for large metabolomic studies than the median fold change method. The method also shows a robust behaviour under small sample size conditions. Furthermore, unlike the median fold change method, the two-step method is able to capture intensity drifts that covariate with the retention time.

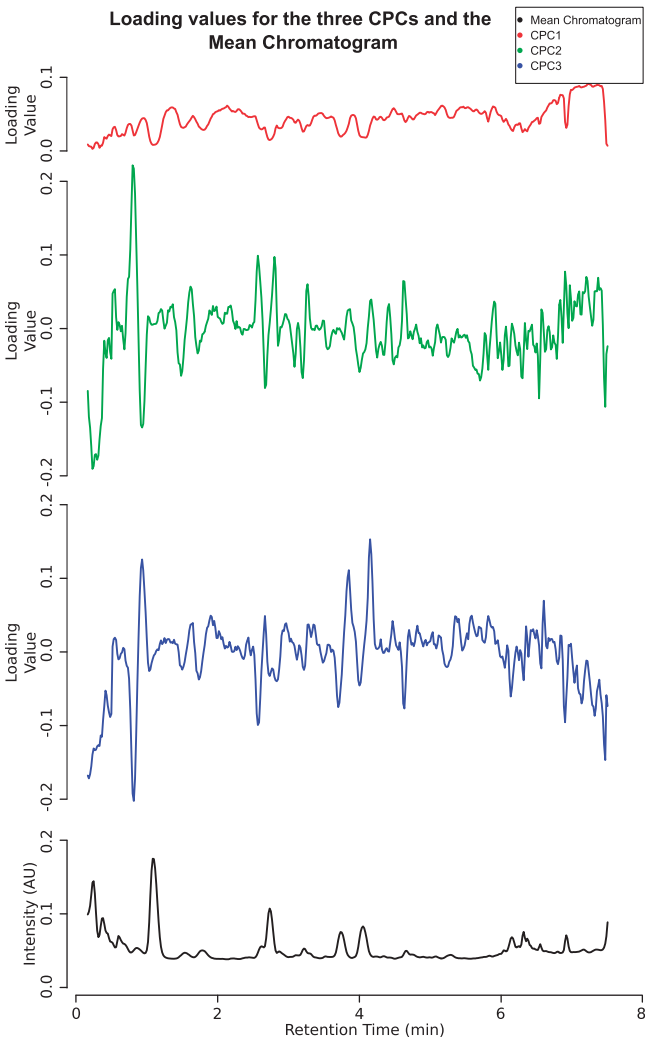

Fig. 3. CPCs loading values and mean chromatogram for the QC samples. The chromatogram is plotted in arbitrary units. The loadings of the CPCs correspond to the columns of the *V* matrix shown in Equation (8)

4 CONCLUSIONS

Applying the combined method CPCA and the median fold change, results in a dataset that contains less drift effects than the dataset corrected solely by the median fold change, and the

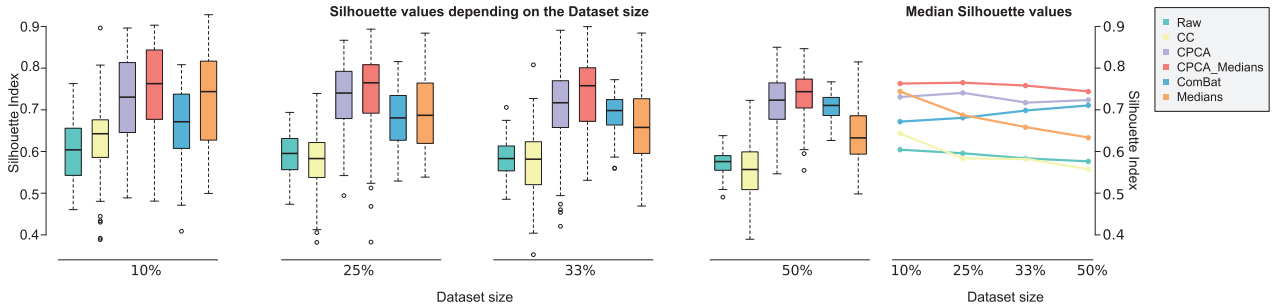

Fig. 4. PCA Scoreplot of all the classes in the data. The three plots in the lower left show the effect of the time elapsed since the first sample was injected, whereas the plots in the top right refer to the batch of the sample. The numbers in the diagonal plots correspond to the variance captured by each PC

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

QC class separability of the dataset is higher than if just the CPC method is applied.

Results show that, among all the methods tested to normalize the LC/MS metabolomic data, the best approach is to use a two-step method in which the first step is to remove the drift by finding the drift direction in a multivariate space using a CPCA approach. The second step, based on performing a median fold change to account for differences in concentration results, improved between-class separability and hence resulted in a better-normalized dataset overall. As far we know, this is the first time that this kind of approach has been applied in the metabolomics context. Applications such as the one proposed open the possibility of carrying out large epidemiological LC/MS metabolomics experiments with high guarantee of the control of the quality of the acquisition data step.

Funding: Spanish national grants (AGL2009-13906-C02-01/ALI, AGL2010-10084-E, 2014 SGR 1063, 2014 SGR 1566), the CONSOLIDER INGENIO 2010 Programme, FUN-C-FOOD (CSD2007-063) from the Spanish Ministry of Economy and Competitiveness (MINECO), as well as FEDER (Fondo Europeo de Desarrollo Regional) and Merck Serono 2010 Research Grants (Fundación Salud 2000). R.LL. thanks the MICINN and the European Social Funds for their financial contribution to the R.LL. Ramón y Cajal contract (Ramón y Cajal Programme, MICINN-RYC RYC-2010-07334). This work was partially funded by the Spanish Ministerio de Ciencia y Tecnología through the (TEC2010-20886-C02-02 and TEC2010-20886-C02-01) grants, and the Ramón y Cajal programme. A.P. is part of the (2009SGR-1395) consolidated research group of the Generalitat de Catalunya, Spain. CIBER-BBN is an initiative of the Spanish ISCIII. F. Fernández-Albert thanks EVALXARTA-UB and Agència de Gestió d'Ajuts Universitaris I de Recerca, AGAUR (Generalitat de Catalunya) for their financial support. M.G.-A. thanks the Generalitat de Catalunyas Agency for Management of University and Research Grants (AGAUR) for the predoctoral (FI-DGR 2011) fellowship.

Conflict of Interest: none declared.

REFERENCES

Aihua,Z. *et al.* (2012) Modern analytical techniques in metabolomics analysis. *Analyst*, **137**, 293–300.

Artursson,T. *et al.* (2000) Drift correction for gas sensors using multivariate methods. *J. Chemom.*, **14**, 711–723.

Brock,G. *et al.* (2008) clValid: an R package for cluster validation. *J. Stat. Softw.*, **25**, 1–22.

Brock,G. *et al.* (2011) *clValid: Validation of Clustering Results*. R package version 0.6-4. Comprehensive R Archive Network (CRAN).

Burton,L. *et al.* (2008) Instrumental and experimental effects in LC-MS-based metabolomics. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, **871**, 227–235.

Chen,K. *et al.* (2013) Gene expression profile analysis of human intervertebral disc degeneration. *Genet. Mol. Biol.*, **36**, 448–454.

Dunn,J.C. (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. Cybern.*, **3**, 32–57.

Dunn,W.B. *et al.* (2011) Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.*, **6**, 1060–1083.

Fiehn,O. *et al.* (2006) Establishing reporting standards for metabolomic and metabonomic studies: a call for participation. *OMICS*, **10**, 158–163.

Filzmoser,P. and Fritz,H. (2007) Exploring high-dimensional data with robust principal components. In: Aivazian,S. *et al.* (eds) *Proceedings of the eight International Conference on Computer Data Analysis and Modeling*. Vol. 1, Belarusian State University, Minsk, pp. 18–22.

Filzmoser,P. *et al.* (2013) *pcaPP: Robust PCA by Projection Pursuit*. R package version 1.9-49. Comprehensive R Archive Network (CRAN).

Flury,B.N. (1984) Common principal components in K groups. *J. Am. Stat. Assoc.*, **79**, 892–898.

Hubert,M. *et al.* (2005) ROBPCA: a new approach to robust principal component analysis. *Technometrics*, **47**, 1–34.

Johnson,W.E. *et al.* (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Leek,J.T. *et al.* (2012) The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28**, 882–883.

Leitch,H.G. *et al.* (2013) Naive pluripotency is associated with global DNA hypomethylation. *Nat. Struct. Mol. Biol.*, **20**, 311–316.

Llorach,R. *et al.* (2009) An LC-MS-based metabolomics approach for exploring urinary metabolome modifications after cocoa consumption. *J. Proteome Res.*, **8**, 5060–5068.

Smith,C.A. *et al.* (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, **78**, 779–787.

Sysi-Aho,M. *et al.* (2007) Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC Bioinformatics*, **8**, 93.

Tulipani,S. *et al.* (2011) Metabolomics unveils urinary changes in subjects with metabolic syndrome following 12-week nut consumption. *J. Proteome Res.*, **10**, 5047–5058.

Trendafilov,N.T. (2010) Stepwise estimation of common principal components. *Comput. Stat. Data Anal.*, **54**, 3446–3457.

Veselkov,K.A. *et al.* (2011) Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Anal. Chem.*, **83**, 5864–5872.

Wang,S.Y. *et al.* (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. *Anal. Chem.*, **85**, 1037–1046.

Xin,L. *et al.* (2008) LC-MS-based metabonomics analysis. *J. Chromatogr. B*, **866**, 64–76.

Ziyatdinov,A. *et al.* (2010) Drift compensation of gas sensor array data by common principal component analysis. *Sens. Actuators B Chem.*, **146**, 460–465.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016