

# Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction

Carl Brunius<sup>1,3</sup>  · Lin Shi<sup>1,3</sup> · Rikard Landberg<sup>1,2,3</sup>

Received: 13 June 2016 / Accepted: 15 September 2016 / Published online: 22 September 2016  
© The Author(s) 2016. This article is published with open access at Springerlink.com

## Abstract

**Introduction** Liquid chromatography-mass spectrometry (LC-MS) is a commonly used technique in untargeted metabolomics owing to broad coverage of metabolites, high sensitivity and simple sample preparation. However, data generated from multiple batches are affected by measurement errors inherent to alterations in signal intensity, drift in mass accuracy and retention times between samples both within and between batches. These measurement errors reduce repeatability and reproducibility and may thus decrease the power to detect biological responses and obscure interpretation.

**Objective** Our aim was to develop procedures to address and correct for within- and between-batch variability in processing multiple-batch untargeted LC-MS metabolomics data to increase their quality.

**Methods** Algorithms were developed for: (i) alignment and merging of features that are systematically misaligned between batches, through aggregating feature presence/missingness on batch level and combining similar features

orthogonally present between batches; and (ii) within-batch drift correction using a cluster-based approach that allows multiple drift patterns within batch. Furthermore, a heuristic criterion was developed for the feature-wise choice of reference-based or population-based between-batch normalisation.

**Results** In authentic data, between-batch alignment resulted in picking 15 % more features and deconvoluting 15 % of features previously erroneously aligned. Within-batch correction provided a decrease in median quality control feature coefficient of variation from 20.5 to 15.1 %. Algorithms are open source and available as an R package ('batchCorr').

**Conclusions** The developed procedures provide unbiased measures of improved data quality, with implications for improved data analysis. Although developed for LC-MS based metabolomics, these methods are generic and can be applied to other data suffering from similar limitations.

**Keywords** Metabolomics · LC-MS · Data correction · Batch alignment · Drift correction

**Electronic supplementary material** The online version of this article (doi:10.1007/s11306-016-1124-4) contains supplementary material, which is available to authorized users.

✉ Carl Brunius  
carl.brunius@slu.se

<sup>1</sup> Department of Food Science, Uppsala BioCenter, Swedish University of Agricultural Sciences, Box 7051, 750 07 Uppsala, Sweden

<sup>2</sup> Unit of Nutritional Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Box 210, 171 77 Stockholm, Sweden

<sup>3</sup> Present Address: Department of Biology and Biological Engineering, Chalmers University of Technology, 412 96 Göteborg, Sweden

## 1 Introduction

Untargeted metabolomics aims to profile the global metabolome, i.e. the (semi-)quantitative collection of low molecular weight metabolites within a biological system, usually in biofluids such as urine, serum, plasma or tissue/cellular extracts (Shulaev 2006; Patti et al. 2012; Vinayavekhin and Saghatelian 2010; Yin and Xu 2014; Alonso et al. 2015). Metabolomics thus finds its place downstream of genomics and proteomics and represents the omics technique closest to phenotype, through the interactions of the previous omics levels with the exposome

(Scalbert et al. 2014; Rappaport et al. 2014). Over the past decade, it has become an increasingly used tool in biological and medical research through possibilities offered for predictive biomarker discovery, elucidation of metabolic pathway alterations and disease aetiology and reflection of demography, lifestyle and exposures (Dunn et al. 2011; Matsuda et al. 2009; Bajad and Shulaev 2011; Beckmann et al. 2013; Dunn et al. 2012; Rappaport et al. 2014). Among the different techniques employed in metabolomics, untargeted liquid chromatography-mass spectrometry (LC-MS) is extensively used due to its high sensitivity, simple sample preparation and broad coverage of metabolites (Theodoridis et al. 2008; Fernandez-Albert et al. 2014; Bajad and Shulaev 2011). Until recently, mass spectrometric techniques were not sufficiently reproducible for large-scale untargeted metabolomics studies involving thousands of samples. However, advances in instrumentation, experimental protocols and data processing methods now permit the use of LC-MS in large-scale untargeted studies with thousands of samples for analysis (Ganna et al. 2014; Drohan et al. 2014).

Data from large-scale LC-MS based metabolomics experiments are generally collected over long periods and analysed in multiple batches. The data collected are affected by systematic and random variability in signal sensitivity, mass accuracy ( $m/z$ ) and retention times ( $rt$ ) between samples both within and between batches. This variability gives rise to critical challenges regarding information loss and data processing (Dunn et al. 2011, 2012).

Within- and between-batch variations in signal intensity (Warrack et al. 2009; Sysi-Aho et al. 2007; Dunn et al. 2011) contribute to noise in the data and therefore have a negative impact on statistical analysis and consequently on the discovery and accurate quantification of metabolites of interest (Veselkov et al. 2011; Wang et al. 2013). Shifts in  $m/z$  and  $rt$  of molecular features between analytical runs result in different extracted spectrum patterns for a single metabolite across samples, with potential misalignment as a consequence. Such drifts could therefore severely affect subsequent statistical analysis and further metabolite identification (America et al. 2006; Nordström et al. 2006; Lange et al. 2008; Zhang et al. 2015). In a recent review (Smith et al. 2015), the state-of-the-art on peak alignment algorithms is well summarised. Unfortunately, current algorithms still suffer from shortcomings, especially regarding between-batch misalignment, since  $m/z$  and  $rt$  shifts are generally much larger between batches than within. Moreover, to the best of our knowledge there are no available methods to specifically address systematic misalignment across multiple batches. Improved algorithms are thus urgently needed.

Different approaches for signal intensity drift management are available. A common approach is to include

internal standards (Bijlsma et al. 2006; Sysi-Aho et al. 2007), but this may not be feasible for untargeted metabolomics studies since available internal standards only represent a limited number of metabolites and signal intensity fluctuations may differ between various metabolite classes (Ejigu et al. 2013; Dunn et al. 2012; Vinayavekhin and Saghatelian 2010). In large-scale untargeted metabolomics studies, the most simple normalisation methods are based on total intensity or intensity of the most stable features. However, these methods are questionable since they assume similar intensity shifts for all features between samples and consequently perform less well than feature-based normalisation techniques (Kamleh et al. 2012). Slightly more advanced methods include e.g. quantile normalisation techniques, which are based on the assumption of similarity of signal intensity distributions, rather than the intensities themselves (Kohl et al. 2012; Lee et al. 2012). However, these methods do not take into account specific feature drift patterns or different signal intensity distributions between different sample classes (e.g. case-control). More recently, quality control (QC) sample strategies have been commonly applied in signal drift management (Kirwan et al. 2013; Dunn et al. 2011, 2012; Kamleh et al. 2012). QC samples have a matrix composition similar to that of the biological samples to be studied, normally achieved by pooling aliquots of the study samples. These QC samples are then injected randomly or regularly within batches to evaluate the LC-MS system and data pretreatment performance, followed by algorithms aiming to discard noisy features or to reduce sample-to-sample or batch-to-batch variations in signal intensity (Dunn 2012; Nezami Ranjbar et al. 2013; Fernandez-Albert et al. 2014; Dunn et al. 2011; Kamleh et al. 2012; Kirwan et al. 2013).

In the present work we introduce two new approaches for overcoming the above-mentioned obstacles regarding processing multiple batch LC-MS metabolomics data, i.e. between-batch feature alignment and within-batch cluster-based drift correction. We also introduce a heuristic suitability criterion to aid in the choice of reference-based or population-based between-batch signal intensity normalisation per feature. Although these approaches are designed for untargeted LC-MS metabolomics, they can be extended to other areas of chemical analysis, such as GC-MS or LC-MS, and for purposes other than metabolomics where signal intensity drift and alignment issues may occur.

## 2 Materials and methods

Throughout this article, the term ‘feature’ refers to a mass spectral peak, i.e. a molecular entity with a unique  $m/z$  and retention time as measured by an LC-MS instrument, such

as a metabolite ion, isotope, adduct, fragment or random noise.

## 2.1 Data set

Fasting plasma samples with heparin as anticoagulant originating from a type 2 diabetes (T2D) case-control study nested within the Northern Sweden Health and Disease Study Cohort (Norberg et al. 2010) were obtained from the Medical Biobank in Umeå (Hallmans et al. 2003). The study was approved by the regional ethical review board in Uppsala (Dnr 2011/228). The samples ( $n = 503$ ) were drawn in 1991–2005 from men and women who later developed T2D, for whom previously unfrozen plasma samples were available, and from individually matched controls who remained free of diabetes until the end of follow-up. Additional fasting plasma samples ( $n = 187$ ) taken 10 years later from controls were used to assess the long-term stability of potential biomarkers and repeat samples from cases ( $n = 187$ ) were analysed to assess potential changes in metabolites related to the risk of developing of T2D. Instrument analyses were performed with approximately 250 samples per batch (including QC and reference samples) in eight batches over 6 months. In the instrumental analysis protocol, two independent biological sample types were used to monitor the stability and functionality of the system throughout all analyses. These were: batch-specific quality control samples (QCs), i.e. pooled plasma samples of all biological samples within batch; and long-term reference samples, i.e. pooled plasma samples of healthy people stored and offered by the Institute of Public Health and Clinical Nutrition, Kuopio, Finland, consistently used throughout all batches. The reference samples were thus not of the same biological origin as the samples and QCs and, unlike the QCs, were therefore not directly representative of the sample population. The QCs and reference samples were injected at the beginning and end and as every 14th injection throughout each batch sequence, and together constituted approximately 16 % of analytical samples. The LC-MS data used were taken from three of the eight batches, selected randomly, and constituted a subgroup of quality monitoring samples, including 48 QCs and 42 reference samples.

LC-MS based metabolomics was conducted in collaboration with the metabolomics platform at the University of Eastern Finland. Preparation of fasting plasma samples for metabolite profiling followed the procedure described by Hanhineva et al. (2015b). In brief, 90  $\mu\text{L}$  sample was mixed with 360  $\mu\text{L}$  acetonitrile, incubated in an ice bath for 15 min and then centrifuged at 1200 g through 0.2  $\mu\text{m}$  polytetrafluoroethylene filters. After 5 min, clear, de-proteinated filtrate was collected for analysis. Plasma samples were analysed by ultra-high performance liquid

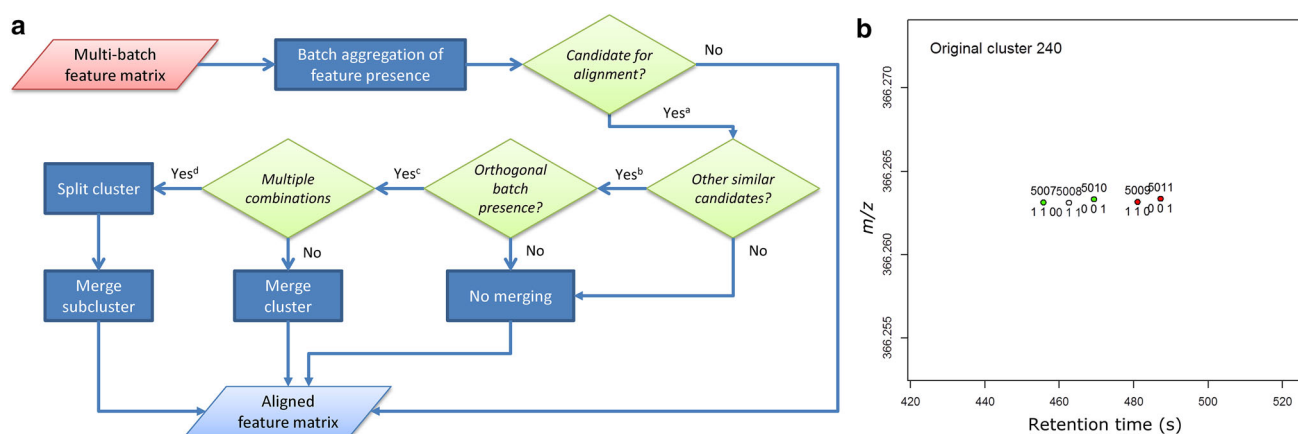
chromatography quadrupole time-of-flight mass spectrometry (UHPLC-qTOF-MS, Agilent Technologies). The system consisted of a 1290 LC system, a Jetstream electrospray ionisation source and a 6540UHD accurate mass qTOF spectrometer operating in positive ionisation mode. The procedure for sample analysis was as described in detail by Hanhineva et al. (2015b), with modification. In brief, 4  $\mu\text{L}$  of the sample solution were injected on the column (Zorbax Eclipse XDB-C18,  $2.1 \times 100$  mm, 1.8  $\mu\text{m}$ ) operating at 50  $^{\circ}\text{C}$ . The mobile phase was delivered in a reversed-phase gradient elution at 0.4 mL/min, using water (eluent A) and methanol (eluent B), both containing 0.1 % formic acid. The following gradient profile was used: 2/100 % B (0–10 min), 100 % B (10–14.5 min), 100/2 % B (14.5 min), 2 % B (14.5–16.5 min). The MS conditions were set up as previously described (Hanhineva et al. 2015b) and the instrument scanned from 20 to 1600 m/z. Data were collected in centroid mode at an acquisition rate of 1.67 spectra/s with an abundance threshold of 150.

Instrument data were exported to ‘xml’ file format and processed in the R open source environment (v 3.2.0; R core team 2016) using the XCMS package (Smith et al. 2006; Tautenhahn et al. 2008). XCMS peak picking parameters (prefilter, peakwidth, mzdif, snthresh) were obtained using the IPO R package (Libiseller et al. 2015). Final peak picking parameters were: prefilter = c(3,440), peakwidth = c(5,76), snthresh = 6, mzdif = 0.0045, ppm = 15. Initial alignment (bw = 15, minfrac = 0.75, minsamp = 1, mzwid = 0.015) and retention time correction (standard loess, family = “s”, span = 0.2) were then applied. For the final alignment, bw was set to the largest observed retention time deviation from visual inspection of XCMS retention time correction plots obtained within batch. Consequently, final alignment was applied with parameters: bw = 1, mzwid = 0.015, minfrac = 0.75, and was followed by filling in missing peaks (method = ‘chrom’).

## 2.2 Feature alignment between batches

Alignment of features systematically misaligned between batches was performed in a multistep algorithm (Fig. 1a). First, to investigate systematic missingness and filter out random noise inherent in individual samples, feature missingness was aggregated on batch level. This was done by batch-wise per-feature calculation of the proportion of missingness among the reference samples and flagging batch absence for those features satisfying the criterion (Eq. 1):

$$proportion_{NA,feature,batch} = \frac{nSamples_{NA,feature,batch}}{nSamples_{Total,batch}} > 80\% \quad (1)$$



**Fig. 1** Proposed algorithm for between-batch feature alignment. **a** Flowchart for alignment of features systematically misaligned between batches. <sup>a</sup>A feature is considered a potential candidate for alignment if  $0 < \text{total batch presence} < \text{number of batches}$ . <sup>b</sup>Alignment candidates are considered similar if having  $m/z$  and  $rt$  within user-defined tolerance. <sup>c</sup>Candidates are considered for alignment and subsequently clustered if not mutually present in the same batch (i.e. presence vector orthogonality). <sup>d</sup>Clusters containing multiple possible alignments are recursively subdivided into sub-clusters. **b** Deconvolution of multiple batch alignment candidates within a selected cluster. During peak picking, five molecular features (numbers 5007–5011) were detected within the specified  $m/z$  and  $rt$  tolerance, in samples from three analytical batches. Feature presence/

absence was aggregated on batch level and marked as absent if missing from  $>80\%$  of QC samples per batch. Presence (1) or absence (0) is noted in the figure for the five features in vector format, where each position in the vector corresponds to presence/absence in the three batches. Candidates for alignment (i.e. the four colour-filled features) were identified through (i) proximity in the  $m/z$  and  $rt$  domains (i.e. closeness) and (ii) orthogonality of presence vectors (i.e. two alignment candidates cannot be present in the same batch). Note the unfilled feature number 5008, which due to the second criterion is excluded as a possible alignment candidate. Multiple alignment candidates were sub-clustered (different colour-filled features) through a recursive deconvolution algorithm

where  $NA$  denotes missing value and  $nSamples$  denotes the number of samples within a batch. The  $80\%$  limit was chosen as an extension of the  $80\%$  rule often employed in metabolomics (Smilde et al. 2005). Batch presence was similarly flagged as non-absence. Candidates for batch alignment were those features where the sum of presence flags exceeded zero (representing being missing from all batches, thus with no possibility for alignment) and lower than the total number of batches (representing being present in all batches, thus with no possibility for alignment).

For each candidate, correspondence with other candidates (“events”) was investigated through distance, i.e. within a user-defined box bounded by largest allowed absolute  $m/z$  and  $rt$  differences under the constraint of batch presence being orthogonal between features, i.e. ensuring that two features present in the same batch cannot be aligned. The boundary value for  $m/z$  (0.002 Da) was set according to instrument resolution and  $rt$  (15 s) was determined from maximum retention drift between batches obtained from XCMS.

All distinct events thus consisted of two alignment candidates. Events which shared common alignment candidates were then clustered. In the case where all such cluster candidates were mutually orthogonal, correspondence was assumed and alignment candidates would then be merged. However, in some of these clusters multiple alignment combinations were possible (Fig. 1b; All

coloured features), indicating correspondence to more than one underlying feature. Multiple alignment candidates were in that case disentangled into their respective correspondences through a recursive sub-clustering algorithm: The largest distances per cluster were iteratively removed until single possible alignment candidates (sub-cluster, i.e. unique correspondence) could be identified. All other possible alignment candidates per cluster, including those previously removed, then underwent the same recursive algorithm until no further sub-clustering could be achieved (Fig. 1b; Different colours for different sub-clusters, i.e. unique correspondences).

### 2.3 Cluster-based within-batch drift correction

The multi-batch data were separated into batch-specific subsets and within-batch drift correction was performed separately on each of these subsets in an algorithm consisting of four distinct steps (Fig. 2a): Clustering of features; drift modelling per cluster; drift correction per cluster; and removal of individual features with poor reproducibility after cluster-based drift correction.

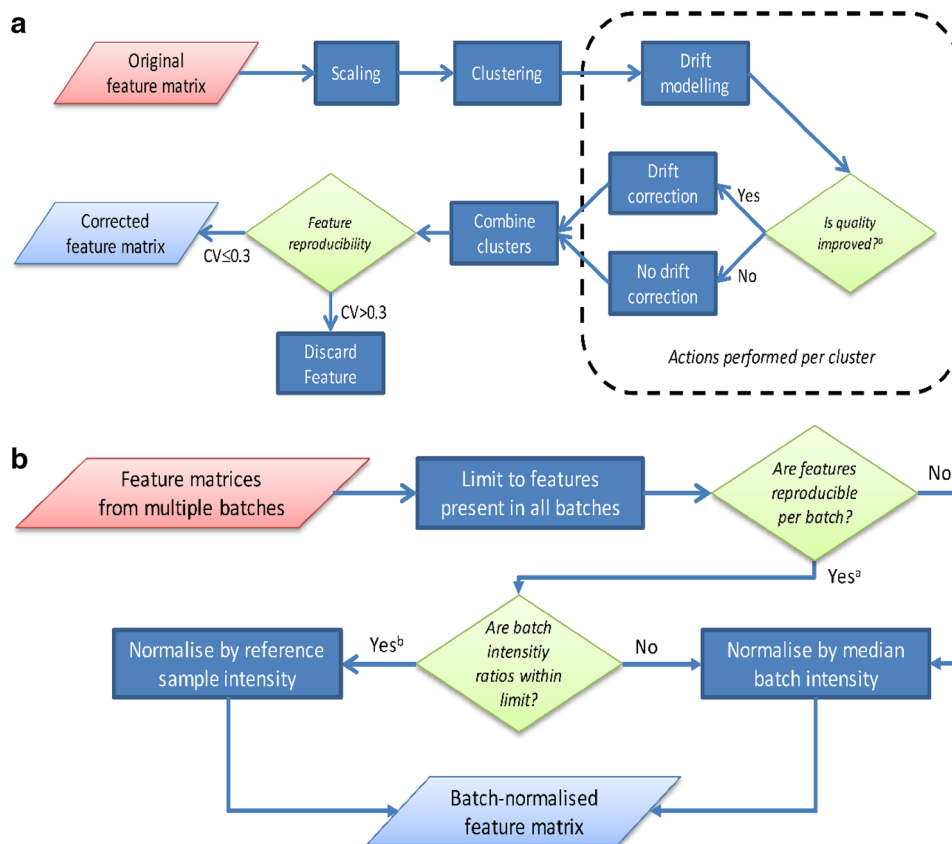
To facilitate clustering of variables, data were first scaled by standard deviation but not centred, under the assumption of predominantly multiplicative rather than additive error terms in instrumental chemical analysis. Clustering of variables was performed under the

assumption that variables with similar drift pattern, in addition to being strongly correlated, are characterised by small Euclidean distances when seen as coordinates in the multivariate sample (or observation) space as opposed to viewing samples as observations in the variable space, which is normally performed in multivariate statistical modelling. A visualisation of distinct drift patterns using lower-dimensionality synthetic data is available as supplementary material (Suppl. Table 1, Figs. 1 and 2).

Clustering of variables in the observation space was achieved by employing the “mclust” algorithm, which utilises a Bayesian approach to determine the type (i.e. geometrical constraint) and optimal number of clusters inherent in the data (Fraley and Raftery 2002; Fraley et al. 2012). This algorithm was chosen to decrease operator bias in the clustering operation. First, a wide range of geometrical constraints in clusters are available, together with the ability to specify a range in potential numbers of cluster to

examine. In the mclust algorithm, final cluster parameters are automatically chosen from the Bayesian Information Criterion (BIC) values from all combinations of user-supplied parameter values (cluster type and number) and a BIC plot is optionally produced for a visual overview of clustering performance for the available parameter combinations. It should be noted, however, that for large, multidimensional data, mclust is a computationally expensive algorithm. Initial testing revealed that restricting the model type to ‘VVE’, i.e. ellipsoidal clusters with equal orientation and number of clusters from 1 to 52 in steps of three, provided a good balance between high quality, unbiased clustering and computational cost. Applying these parameter settings, clustering of 12–18 QC samples × 11 815 features resulted in 25–28 clusters and required approximately 12–18 min in the present case.

After clustering, scaled variables belonging to the same cluster were pooled together and a cubic spline regression



**Fig. 2** Proposed algorithms for within- and between-batch signal intensity drift corrections. **a** Flowchart of cluster-based within-batch intensity drift correction. <sup>a</sup>In this example, cluster quality is considered to be improved if  $rmsd(Ref)_{with\ correction} < rmsd(Ref)_{without\ correction}$ , where ‘rmsd’ denotes root mean squared distance from the cluster centre point. *Ref* denotes long-term reference samples not used for within-batch intensity drift modelling (see Materials and methods section). **b** Flowchart of Between-batch intensity normalisation algorithm. <sup>a</sup>Features are considered reproducible if long-term

reference sample intensity per batch  $CV \leq 30\%$ . <sup>b</sup>Reference sample average feature intensity ratios between batches are considered within limit if not deviating from corresponding average feature intensity ratios by more than a fold change of five. For such features passing both criteria, batches are normalised by average reference sample intensity. For other features, long-term reference samples are not considered sufficiently representative of the sample population and features are thus normalised by median batch intensity

was applied on the pooled cluster data vs injection order to obtain a drift function. The algorithm performs this drift calculation separately for all the clusters and optionally produces plots of the clusters and their correction functions (Fig. 2a, Suppl. Fig. 3). Finally, cluster-based drift correction was achieved by calculation of cluster-wise correction factors (Eq. 2) applied specifically for each injection:

$$\text{correctionFactor}_{c,n} = \frac{\text{driftValue}_{c,1}}{\text{driftValue}_{c,n}} \quad (2)$$

where *correctionFactor* are the cluster-based correction factors derived from the corresponding drift function values *driftValue* for cluster *c* at injection *n*.

These correction factors are derived from the ratio between drift function values at the reference point (i.e. first injection) and at all subsequent injections obtained from the cluster drift function. By multiplying these correction factors to the original unscaled variable data of the cluster, intensity drift was thus normalised to the reference level at the first injection (Suppl. Fig. 3).

Drift correction per cluster was performed only if providing an unbiased measure of increased quality of data measured on non-QC reference samples. This was assessed by cluster-wise evaluation of the root-mean-squared distance (rmsd) from the centre point of the long-term reference samples with and without correction. This provided an unbiased measure since: (i) the long-term reference samples were of different biological origin than the QC samples and (ii) they were not included in drift modelling and correction. Correction per cluster was thus only performed if the rmsd was reduced after drift correction. After drift correction, individual features were removed batch-wise from the subset if QC sample  $CV_{\text{After correction}} > 30\%$ .

## 2.4 Between-batch normalisation

Between-batch normalisation was performed through an iterative process (Fig. 2b). Aggregated batch data after within-batch drift correction and alignment were first limited to common features, i.e. those features common between all batches that were not excluded after cluster-based drift correction per batch. Normalisation was then achieved using either of two standard approaches: Normalisation by reference sample intensity or population-based (median) normalisation. However, to aid in the feature-wise choice between normalisation methods, a heuristic was developed and applied to test for suitability of batch normalisation by reference sample intensity. Normalisation by average reference sample intensity per batch and feature was performed if satisfying the following dual criterion:

$$CV_{\text{Ref}, \text{batch}} < 0.3 \{ \text{for all batches} \} \quad (3)$$

$$\log \left( \frac{\text{Feature Intensity Ratio}_{i,j}}{\text{Average Feature Intensity Ratio}_{i,j}} \right) < \log 5 \{ \text{for all batches } i, j \} \quad (4)$$

where *Feature Intensity Ratio*<sub>*i,j*</sub> is the ratio of average reference sample intensity for a specific feature measured in batches *i* and *j*, respectively and *Average Feature Intensity Ratio*<sub>*i,j*</sub> is the ratio for the average intensity of all features within the batches *i* and *j*, respectively. The use of log-transformation on each side of Eq. 4 is to provide equidistant fold changes for batches *i* and *j*. A fold change limit of 5 was used for this data, but the limit can be user-defined (see Results and discussion). When the criterion was not met, the reference samples were not considered sufficiently representative of the sample population or otherwise inadequate for normalisation, in which case batches were normalised by median intensity of batch sample populations under the assumption of similar population distributions between batches (data not shown).

## 2.5 Computer hardware and software

Algorithms were developed in the open source statistical software environment R v 3.2.2 (R Core Team 2016) and depended on the following non-base packages: “mclust” v 5.0.1 (Fraley and Raftery 2002; Fraley et al. 2012), “reshape” v 0.8.5 (Wickham 2007) and “XCMS” v 1.44.0 (Smith et al. 2006; Tautenhahn et al. 2008). All calculations were performed on a HP Elitebook with an Intel i7-3687U processor operating under Windows 7. Functions and data sets are available as an R package (*batchCorr*) and, together with example data and script, are freely available from <https://gitlab.com/CarlBrunius/batchCorr>.

## 3 Results and discussion

Multi-batch high-resolution LC-MS data present challenges in terms of signal intensity drift and feature misalignment from instrument deviations in the *m/z* and *rt* domains. These deviations have contributions both from within- and between-batch irregularities, with the latter generally expected to make a greater contribution. Data sets should consequently be first examined for systematic feature misalignment between batches to avoid loss of data integrity, the type of which depends on the data pipeline employed. In the case of forced integration data filling (such as the *fillPeaks* algorithm in the XCMS package) and data imputation, there is the obvious risk of splitting one true informative feature into two or more with either less or wrong information (Suppl. Fig. 4). Data filling creates

artificially batch-specific features from the systematic differences in presence/missingness, i.e. erroneous information, whereas imputation in the best of cases results in two (or more) identical features, although most imputation techniques would struggle to achieve accurate imputations with this type of batch-specific systematic missingness, thus increasing noise in the variables and consequently in following statistical analyses, i.e. providing/introducing less and/or wrong information.

Available options for feature alignment (Smith et al. 2015) do not distinguish within-batch from between-batch deviations and are therefore best suited for within-batch alignment of features. To investigate and correct for systematic batch structures, presence/missingness was aggregated per batch to filter out spurious random noise or erroneous misalignment of individual samples. Batch absence was decided employing an 80 % threshold (Bijlsma et al. 2006), but robust results were in fact obtained between threshold settings of 60–85 % (data not shown).

To be considered for alignment, features should be sufficiently close in the  $m/z$  and the  $rt$  domains and also have orthogonal feature vectors. Using features from Fig. 1b as an example: Feature 5007, flagged as [1 1 0] would be present in batches 1 and 2, but not in batch 3, whereas feature 5010, flagged as [0 0 1] would be orthogonally present and therefore a possible alignment candidate. In the majority of cases, alignment candidates were only involved with one other candidate. However, multiple alignment combinations within the same  $m/z$ – $rt$  box were also observed (Fig. 1b; All coloured features), possibly due to extreme similarity in retention time of e.g. stereoisomers. Several approaches for disentanglement of such multiple alignment candidates are possible. The naïve option to choose the alignment candidates with the shortest distance in the  $m/z$ – $rt$  box resulted in apparent misalignments (data not shown). A recursive sub-clustering algorithm was therefore developed to identify all unique correspondences (Fig. 1b; different colours per correspondence). The algorithm also optionally produces plots of alignment events and sub-clusters. The effectiveness of the clustering and sub-clustering algorithms was confirmed by visual inspection of clustering and sub-clustering results.

This algorithm for systematic batch misalignment is easily integrated with available sample-based alignment methods. Within an analytical batch,  $m/z$  and  $rt$  can be expected to remain within small tolerances relative to between-batch variation, especially in the  $rt$  dimension. For metabolomics data spanning several batches, low tolerance settings thus increase the risk of misalignment, whereas high tolerance risks binning unrelated features. Peak alignment in workflows not addressing batch alignment thus necessitates increasingly higher tolerance settings with the number of batches being combined. In a workflow

employing batch-orthogonal alignment, between-sample alignment can instead be optimised per batch, thus with narrower  $m/z$  and  $rt$  setting, and later combined using batch-aggregated data (Suppl. Fig. 4). For comparison, using the current dataset and a  $bw = 15$  setting, this corresponded to picking approximately 9800 features (Suppl. Fig. 5). Using settings optimised for batch-specific peak picking ( $bw = 1$ ) and systematic batch alignment, approximately 11,300 were instead picked. The advantage was two-fold: (i) 1500 features previously aligned were deconvoluted, contributing to noise reduction in 1500 of the previously available features; and (ii) 1500 additional features were added to the dataset, increasing the information content. It should be noted that by decreasing bandwidth setting alone, there is no effective means to distinguish between true feature deconvolution and artificially splitting features between batches. In fact, without the batch alignment algorithm, a  $bw = 1$  setting picked approximately 11,800 features, of which approximately 500 thus resulted from artificially splitting true features between batches (Suppl. Fig. 5).

For the within-batch signal intensity drift correction algorithm, clustering was automated to provide an unbiased trade-off between drift modelling detail and power through unbiased decision on optimum number of clusters through Bayesian clustering (Fraley and Raftery 2002; Fraley et al. 2012). It was observed that for multivariate, authentic data, model restrictions were not required to achieve reproducible clustering. Restrictions were however imposed to reduce computing time, without any apparent loss of information content. It should be noted that clustering could become computationally more efficient by parallelising cluster model calculations, which remains a point to consider in future algorithm development. Moreover, to increase computational efficiency, other clustering methods could also be considered. However, care should be taken to provide automated, unbiased estimations of optimal number of clusters and distance/clustering functions (Rokach 2009). It should be noted that a major limitation of the clustering is its poor capability for managing missing values, since the clustering is based on Euclidean distances in the multivariate observation space. If values are lacking, then these distances are effectively incalculable. Care must therefore be taken to provide full data matrices, i.e. peak tables without any missing data. This can be achieved either by forced integration between consensus peak limits such as performed by the ‘fillPeaks’ function of the XCMS package, or by imputation.

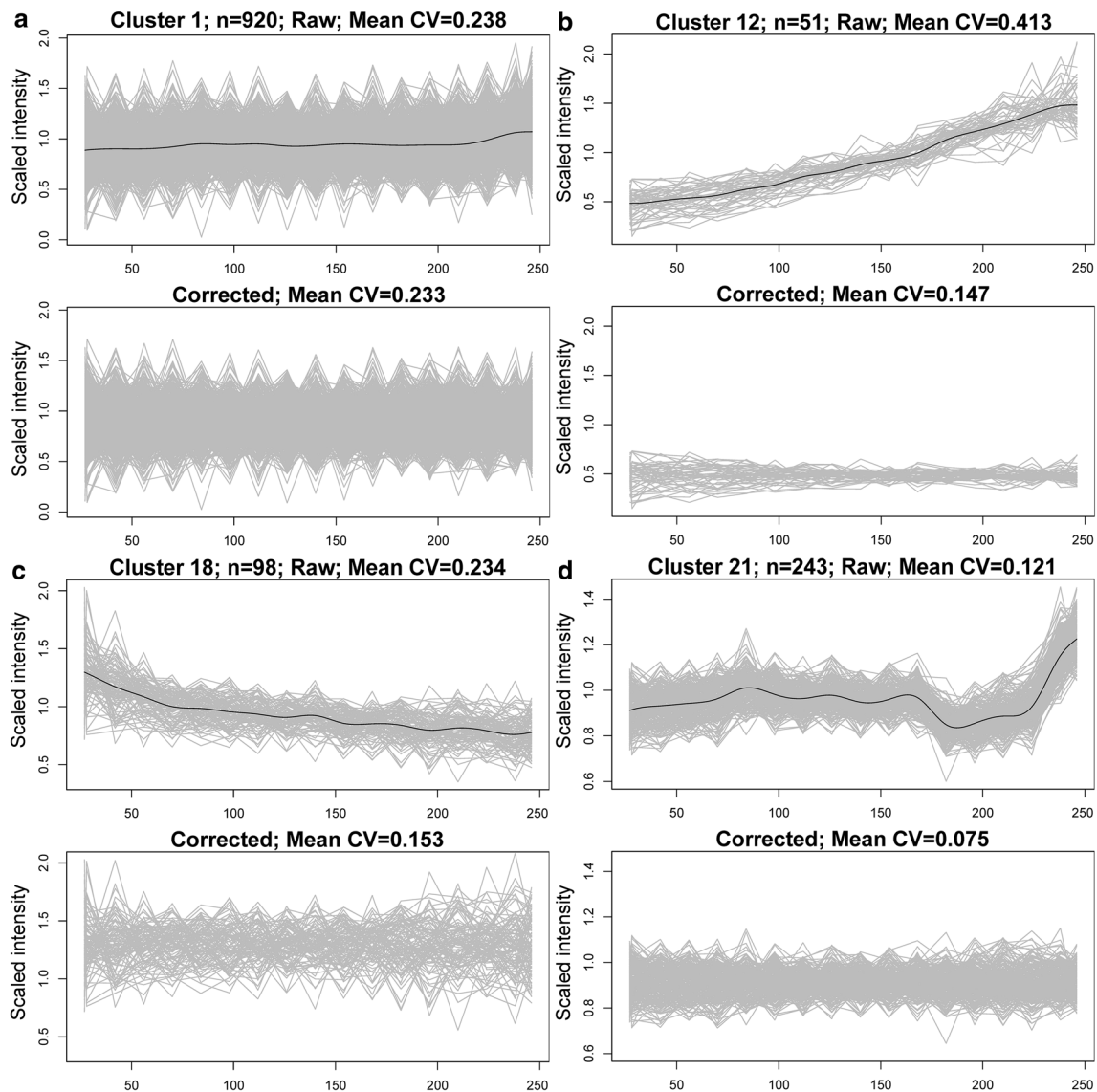
Drift modelling was performed by cubic spline interpolation (Fig. 3). Smoothing functions are often sensitive to parameter settings, but due to the high number of data points per cluster compared with feature-wise interpolation, smoothing was highly reproducible for a wide range

of parameter settings. Similar performance was observed using both local regression (LOESS) and cubic spline regression, but we found that cubic splines tended to be more insensitive to parameter fluctuations and thus less sensitive to operator bias.

When observing the quality control samples in a principal component analysis (PCA), the intensity drift easily observed in the raw data from batch B (Fig. 4a) was drastically reduced when applying the cluster-based drift functions for signal correction (Fig. 4b). However, cluster-wise drift correction was performed only if it increased the quality of the data and such fitness estimation needs to rely on examination of quality monitoring samples not included

in the modelling. In the present case, quality improvement was assessed as decreased Euclidean distance in the multivariate feature space between long-term reference samples (Fig. 4c).

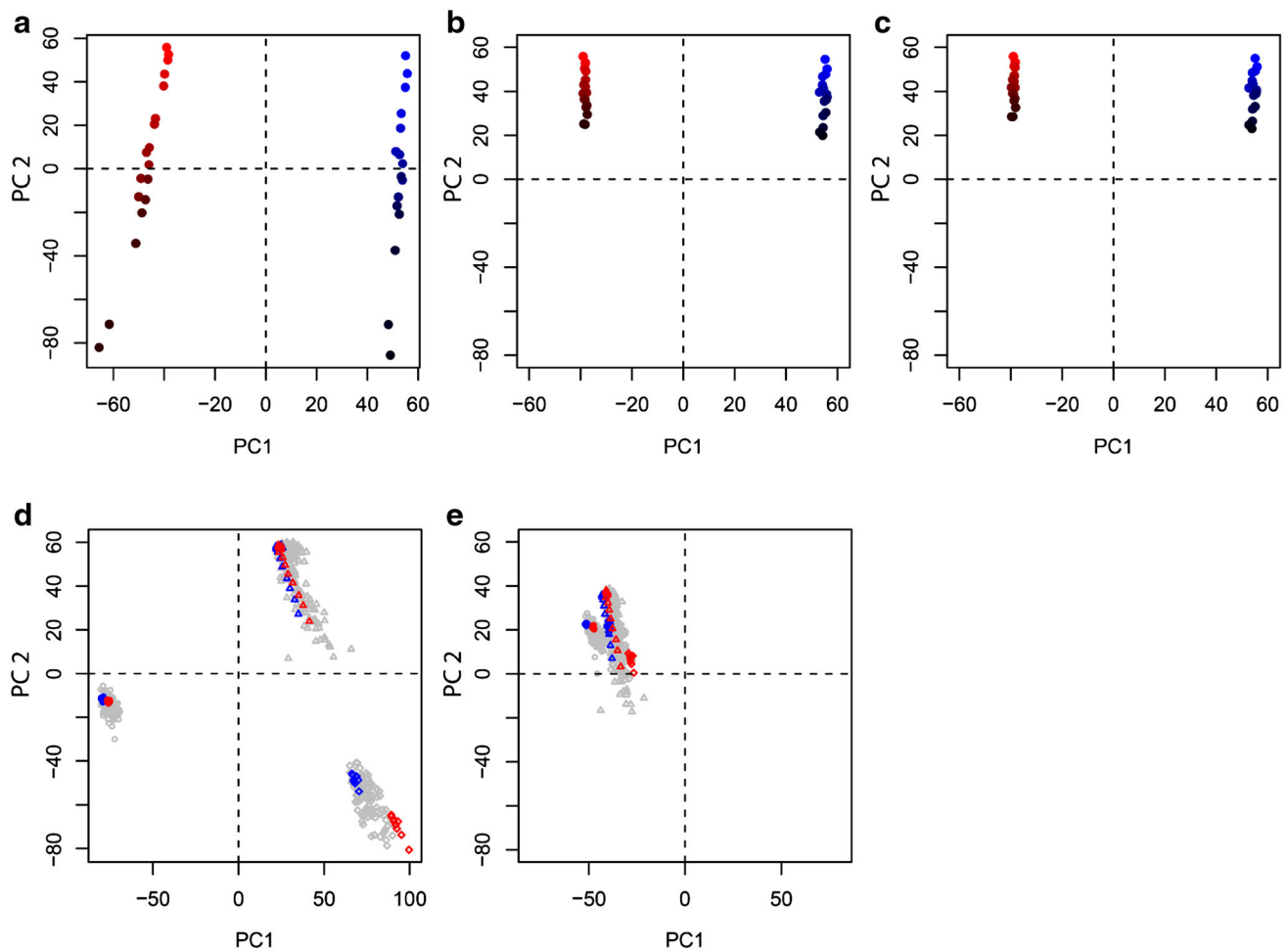
Cluster 18 (Fig. 3) constituted an interesting example of this principle: Calculation of within-cluster drift provided a decrease in QC CV from 23 to 15 %. However, since the drift correction did not result in an increase in unbiasedly assessed quality of independent reference samples, the drift correction was not applied (Table 1). It should also be noted that the algorithm can easily be adapted to suit other schemes for fitness estimations and/or different experimental protocols, such as the use of duplicate (or multiple)



**Fig. 3** Authentic batch B QC features separated into different sized clusters ( $n = 51$ – $920$  features per cluster in this subset) representing different intensity drift behaviours. **a** Cluster 1, similarly to other large clusters (not shown), closely followed the general within-batch intensity drift, which for this batch was minor. Among the other

clusters (**b–d**), several distinctly diverse drift patterns were readily discernible and CV was considerably reduced in those clusters. For each cluster, the upper graph shows the scaled features in grey and the cluster drift function in black. The lower half shows the same features on the same y-scale after application of cluster-based drift correction





**Fig. 4** PCA score plots for performance of within-batch and between-batch drift correction. For within-batch drift correction (**a–c**) *red circles* represent within-batch QC samples and blue circles represent the long-term reference samples. *Color* is scaled corresponding to injection order. The drift observed in the raw, uncorrected data from batch B (**a**), is drastically reduced using cluster-based intensity drift correction either not employing (**b**) or employing (**c**) the fitness criterion of improved reference sample homogeneity

and projected on the same scale as the uncorrected data. For between-batch drift correction (**d–e**), batches B, F and H are presented as *circles*, *triangles* and *squares*, respectively. *Red* represents within-batch QC samples, *blue* the long-term reference samples and *grey* the actual biological samples. The batch effect, clearly observed as the main determinant of variance prior to normalisation (**d**) is drastically reduced using the mixed normalisation procedure (**e**) when projected on the same scale

injections of multiple samples or random subsampling among QC samples. Using multiple samples, fitness can e.g. be assessed through Euclidean distances with one-tailed paired tests of  $H1 < H0$  (populations of distances before/after correction). In the latter case, QCs would be subsampled into two groups, for drift modelling and fitness estimation, respectively.

A final quality control of features was performed by removing individual features with  $CV > 30\%$  in QC samples after drift correction (Fig. 5). A notable effect of the clustering algorithm was the fact that features with poorer reproducibility, as indicated by final QC feature  $CV > 30\%$ , were in general clustered together (Table 1; clusters 3, 8, 9 and 14), as were features with common, highly reproducible drift patterns. It should also be noted

that the correction algorithm may not be relevant for these clusters. However, in order to have a generalisable method, clustering is anyway applied if resulting in unbiasedly assessed increased data quality. According to the criteria, correction led to quality improvement in clusters 3, 8 and 14. However, the quality improvement was not so large as to warrant inclusion of most of these features in the final peak table. The final result is thus a combination of multiple sub-algorithms. In the three batches of the authentic data set, such clusters represented 13–19% of the total number of features. However, some clusters (Table 1, clusters 3 and 14) consisted of both well and poorly behaving features.

Median QC feature CV for the three batches decreased from 20.5 to 18.7% after drift correction and further to

**Table 1** Characteristics of clusters automatically identified by the cluster-based within-batch drift correction algorithm based on quality control (QC) samples from batch B (pooled human plasma samples; see Materials and methods for details) in reverse phase, positive electrospray ionisation UHPLC-QTOF

Cluster	Action <sup>a</sup>	Number of features		Mean QC feature CV (%)		
		Before <sup>b</sup>	After <sup>c</sup>	Before <sup>b</sup>	Correction <sup>d</sup>	After <sup>c</sup>
1	Corrected	920	920	24	23	23
2	Corrected	1177	1177	15	14	14
3	Corrected	715	7	36	35	29
4	Corrected	354	354	9	6	6
5	Corrected	860	860	13	11	11
6	Corrected	1137	1137	21	20	20
7	Corrected	771	771	10	9	9
8	Corrected	305	0	63	63	NA
9	No action	289	0	130	130	NA
10	Corrected	888	691	29	28	28
11	Corrected	309	309	26	25	25
12	Corrected	51	50	41	15	14
13	Corrected	198	198	21	12	12
14	Corrected	530	0	46	45	NA
15	Corrected	115	115	10	4	4
16	Corrected	154	154	30	16	16
17	Corrected	65	65	24	7	7
18	No action	98	83	23	15	21
19	Corrected	214	214	16	8	8
20	Corrected	107	107	12	8	8
21	Corrected	243	243	12	8	8
22	Corrected	1390	1390	18	17	17
23	Corrected	105	105	7	4	4
24	Corrected	221	221	8	6	6
25	Corrected	62	62	6	3	3

<sup>a</sup> Action performed on cluster: corrected if root-mean-squared distance (rmsd) of reference samples is decreased after applied correction, no action otherwise

<sup>b</sup> Values pertaining to original QC features

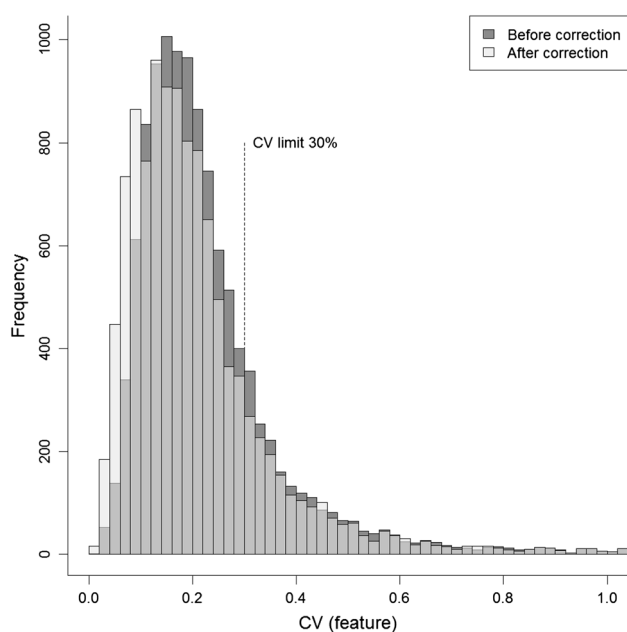
<sup>c</sup> Values pertaining to QC features after entire within-batch correction procedure (Fig. 5)

<sup>d</sup> CV of drift-corrected clusters

15.1 % after feature removal. In many protocols, a CV limit of 20 % is applied, and also recommended in the FDA guidelines for bioanalytical method validation (FDA 2001). For the three batches in the authentic data set, a limit of 20 % resulted in 36–60 % of features being discarded. We advocate that a more accommodating setting of 30 % be applied in exploratory, untargeted metabolomics analysis, which in the present case resulted in 18–32 % of the total number of features in the authentic data batches being discarded (Suppl. Table 2). Allowing a less restrictive CV limit admits additional noise into the variables and subsequent statistics, but also more variables of potential interest into the statistical analyses. Noisy, uninformative variables could then be excluded from the data set in a later step, using e.g. statistical methods incorporating unbiased variable selection (Hanhineva et al. 2015a; Buck et al. 2016).

In addition to drift correction and alignment, batch data are typically intensity-normalised, although in some cases this may not be required, e.g. when employing fold-changes between matched samples within batch as input data for statistical modelling (Jonsson et al. 2015). Similarly to within-batch intensity drift correction, unbiased fitness estimation of normalisation performance should ideally be carried out, although a thorough review of the literature could not reveal established methods for this practice.

For the authentic data, feature-wise between-batch normalisation using either long-term reference samples (with the caveats described above) or sample population was decided using a heuristic dual criterion quality indicator (Fig. 1c), where the first criterion (Eq. 3) assessed the precision of the reference sample intensity for the specific feature and the second criterion (Eq. 4) was a proxy for assessment of accuracy, under the assumption that large



**Fig. 5** CV distribution among QC features before (*dark*) and after (*light*) cluster-based within-batch intensity drift correction of batch B of authentic LC-MS metabolomics data. Before correction, 8996 of originally 11,298 QC features (79.6 %) pass the  $CV \leq 30\%$  criterion. After correction, the corresponding figure is 9233 features (81.7 %)

deviations from the general intensity ratio between batches indicates inaccuracy. With increasing limits of intensity ratio allowance, an increasing proportion of features are normalised by reference sample intensities (Suppl. Fig. 6). When long-term references were not considered representative of the sample population, as indicated by the heuristic, batches were instead normalised by sample population median under the assumption of similar batch population distributions. For visualization purposes, between-batch normalization was performed on the actual study data from 562 samples in 3 analytical batches. When observed in a PCA, samples were initially observed to cluster according to batch (Fig. 4d), whereas these systematic differences were removed after normalization (Fig. 4e).

The algorithms developed for within- and between-batch correction are available as an R package (*'batch-Corr'*), which allows ease of implementation. These algorithms can be used either alone or in combination to suit any particular analytical situation. For example, within-batch correction without alignment or normalisation can be applied if samples are analysed within only one batch. Moreover, in the case of multiple batches, these algorithms can easily be chosen at will, combined with other drift correction and/or normalisation procedures and incorporated into a customised workflow. The internal application

order of the batch correction algorithms developed also leaves freedom of choice as to whether to perform drift correction or alignment first. In the present study, alignment was performed on an entire dataset containing all features, rather than batch-specific subsets with non-similar features present. Moreover, clustering of variables in within-batch drift correction was improved through the removal of noisy non-relevant features by the alignment procedure.

## 4 Conclusions

An approach including multiple algorithms for within- and between-batch correction was developed to overcome some of the measurement errors in LC-MS metabolomics data and thereby improve the quality of data used for statistical analysis. Alignment of peaks systematically misaligned between batches improved the quality of the dataset by merging features otherwise split between batches. This was achieved by aggregating presence/missingness on batch level and combining similar features orthogonally present between batches. Signal intensity drift correction by clustering of features in the observation space increased within-batch data quality by allowing for multiple drift patterns within the same batch. It also minimised the risk of overfitting (e.g. modelling of noise in individual features) by adding statistical strength of multiple features to the individual cluster regressions. Between-batch correction strategies must correspond to the experimental setup at hand. Long-term reference or QC samples are not necessarily representative of the sample population and normalisation of such features can easily introduce severe batch bias. A heuristic indicator was developed to assess the suitability per feature to utilise different normalisation techniques, i.e. reference-based or population-based between-batch normalisation. Care should be taken to employ unbiased measures for quality improvement using data correction techniques to avoid overfitting and introducing bias.

**Acknowledgments** The Swedish University of Agricultural Science (quality grant for young research leaders), the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS) and the Swedish Research Council—Medicine are acknowledged for funding. We thank Dr. Kati Hanhineva for instrumental analysis of plasma samples at the University of Eastern Finland LC-MS Metabolomics Center. We also thank Johan Rosén and Marco Kunzelmann at the Swedish National Food Agency for beta testing the within-batch signal intensity drift correction algorithm.

### Compliance with ethical standards

**Conflict of interest** Carl Brunius, Lin Shi and Rikard Landberg declare that they have no conflicts of interest.

**Informed consent** Informed consent was obtained from all individual participants included in the prospective cohort study from which biological samples analyzed in this study originated.

**Research involving human participants** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The study was approved by the regional ethical committee in Uppsala.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Alonso, A., Marsal, S., & Julià, A. (2015). Analytical methods in untargeted metabolomics: State of the art in 2015. *Frontiers in Bioengineering and Biotechnology*, *3*, 1–20.
- America, A. H. P., Cordewener, J. H. G., van Geffen, M. H. A., Lommen, A., Vissers, J. P. C., Bino, R. J., et al. (2006). Alignment and statistical difference analysis of complex peptide data sets generated by multidimensional LC-MS. *Proteomics*, *6*, 641–653.
- Bajad, S., & Shulaev, V. (2011). LC-MS-based metabolomics. *Methods in Molecular Biology*, *708*, 213–228.
- Beckmann, M., Lloyd, A. J., Haldar, S., Favé, G., Seal, C. J., Brandt, K., et al. (2013). Dietary exposure biomarker-lead discovery based on metabolomics analysis of urine samples. *The Proceedings of the Nutrition Society*, *72*(3), 352–361.
- Bijlsma, S., Bobeldijk, I., Verheij, E. R., Ramaker, R., Kochhar, S., Macdonald, I. A., et al. (2006). Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Analytical Chemistry*, *78*, 567–574.
- Buck, M., Nilsson, L. K. J., Brunius, C., Dabire, R. K., Hopkins, R., & Terenius, O. (2016). Bacterial associations reveal spatial population dynamics in *Anopheles gambiae* mosquitoes. *Scientific Reports*. doi:10.1038/srep22806.
- Drogan, D., Dunn, W. B., Lin, W., Buijsse, B., Schulze, M. B., Langenberg, C., et al. (2014). Untargeted metabolic profiling identifies altered serum metabolites of type 2 diabetes mellitus in a prospective, nested case-control study. *Clinical Chemistry*, *61*, 487–497.
- Dunn, W. B. (2012). Diabetes-the role of metabolomics in the discovery of new mechanisms and novel biomarkers. *Current Cardiovascular Risk Reports*, *7*(1), 25–32.
- Dunn, W. B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., et al. (2011). Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols*, *6*, 1060–1083.
- Dunn, W. B., Wilson, I. D., Nicholls, A. W., & Broadhurst, D. (2012). The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis*, *4*, 2249–2264.
- Ejigu, B. A., Valkenburg, D., Baggerman, G., Vanaerschot, M., Witters, E., Dujardin, J.-C., et al. (2013). Evaluation of normalization methods to pave the way towards large-scale LC-MS-based metabolomics profiling experiments. *Omics: A Journal of Integrative Biology*, *17*(9), 473–485.
- FDA. (2001). *Guidance for industry: Bioanalytical method validation*. US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research
- Fernandez-Albert, F., Llorach, R., Garcia-Aloy, M., Ziyatdinov, A., Andres-Lacueva, C., & Perera, A. (2014). Intensity drift removal in LC/MS metabolomics by common variance compensation. *Bioinformatics*, *30*(20), 2899–2905.
- Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*(458), 611–631.
- Fraley, C., Raftery, A., Murphy, T., & Scrucca, L. (2012). *mclust Version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation*. Seattle: University of Washington.
- Ganna, A., Salihovic, S., Sundström, J., Broeckling, C. D., Hedman, Å. K., Magnusson, P. K. E., et al. (2014). Large-scale metabolomic profiling identifies novel biomarkers for incident coronary heart disease. *PLoS Genetics*, *10*(12), e1004801.
- Hallmans, G., Agren, A., Johansson, G., Johansson, A., Stegmayr, B., Jansson, J.-H., et al. (2003). Cardiovascular disease and diabetes in the Northern Sweden Health and Disease Study Cohort-evaluation of risk factors and their interactions. *Scandinavian Journal of Public Health*, *31*(61), 18–24.
- Hanhineva, K., Brunius, C., Andersson, A., Marklund, M., Juvonen, R., Keski-Rahkonen, P., et al. (2015a). Discovery of urinary biomarkers of whole grain rye intake in free-living subjects using nontargeted LC-MS metabolite profiling. *Molecular Nutrition & Food Research*, *59*, 2315–2325.
- Hanhineva, K., Lankinen, M. A., Pedret, A., Schwab, U., Kolehmainen, M., Paananen, J., et al. (2015b). Nontargeted metabolite profiling discriminates diet-specific biomarkers for consumption of whole grains, fatty fish, and bilberries in a randomized controlled trial. *Journal of Nutrition*, *145*(1), 7–17.
- Jonsson, P., Wuolikainen, A., Thysell, E., Chorell, E., Stattin, P., Wikström, P., et al. (2015). Constrained randomization and multivariate effect projections improve information extraction and biomarker pattern discovery in metabolomics studies involving dependent samples. *Metabolomics*, *11*(6), 1667–1678.
- Kamleh, M. A., Ebbels, T. M. D., Spagou, K., Masson, P., & Want, E. J. (2012). Optimizing the use of quality control samples for signal drift correction in large-scale urine metabolic profiling studies. *Analytical Chemistry*, *84*(6), 2670–2677.
- Kirwan, J. A., Broadhurst, D. I., Davidson, R. L., & Viant, M. R. (2013). Characterising and correcting batch variation in an automated direct infusion mass spectrometry (DIMS) metabolomics workflow. *Analytical and Bioanalytical Chemistry*, *405*, 5147–5157.
- Kohl, S. M., Klein, M. S., Hochrein, J., Oefner, P. J., Spang, R., & Gronwald, W. (2012). State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics*, *8*, 146–160.
- Lange, E., Tautenhahn, R., Neumann, S., & Gröpl, C. (2008). Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics*, *9*, 1–19.
- Lee, J., Park, J., Lim, M., Seong, S. J., Seo, J. J., Park, S. M., et al. (2012). Quantile normalization approach for liquid chromatography-mass spectrometry-based metabolomic data from healthy human volunteers. *Analytical Sciences*, *28*(8), 801–805.
- Libiseller, G., Dvorzak, M., Kleb, U., Gander, E., Eisenberg, T., Madeo, F., et al. (2015). IPO: A tool for automated optimization of XCMS parameters. *BMC Bioinformatics*, *16*(1), 1–10.
- Matsuda, F., Yonekura-Sakakibara, K., Niida, R., Kuromori, T., Shinozaki, K., & Saito, K. (2009). MS/MS spectral tag-based annotation of non-targeted profile of plant secondary

- metabolites. *The Plant Journal : For Cell and Molecular Biology*, 57(3), 555–577.
- Nezami Ranjbar, M. R., Zhao, Y., Tadesse, M. G., Wang, Y., & Resson, H. W. (2013). Gaussian process regression model for normalization of LC-MS data using scan-level information. *Proteome Science*, 11, 1–13.
- Norberg, M., Wall, S., Boman, K., & Weinehall, L. (2010). The Västerbotten Intervention Programme: background, design and implications. *Global Health Action*, 3, 1–15.
- Nordström, A., O'Maille, G., Qin, C., & Siuzdak, G. (2006). Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: Quantitative analysis of endogenous and exogenous metabolites in human serum. *Analytical Chemistry*, 78(10), 3289–3295.
- Patti, G. J., Yanes, O., & Siuzdak, G. (2012). Metabolomics: The apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*, 13(4), 263–269.
- R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rappaport, S. M., Barupal, D. K., Wishart, D., Vineis, P., & Scalbert, A. (2014). The blood exposome and its role in discovering causes of disease. *Environmental Health Perspectives*, 122(8), 769–774.
- Rokach, L. (2009). A survey of Clustering Algorithms. *Data mining and knowledge discovery handbook* (pp. 269–298). Boston: Springer.
- Scalbert, A., Brennan, L., Manach, C., Andres-Lacueva, C., Dragsted, L. O., Draper, J., et al. (2014). The food metabolome: A window over dietary exposure. *The American Journal of Clinical Nutrition*, 99(6), 1286–1308.
- Shulaev, V. (2006). Metabolomics technology and bioinformatics. *Briefings in Bioinformatics*, 7(2), 128–139.
- Smilde, A. K., Van Der Werf, M. J., Bijlsma, S., Der Werff-Van, Van, Der Vat, B. J. C., & Jellema, R. H. (2005). Fusion of mass spectrometry-based metabolomics data. *Analytical Chemistry*, 77(20), 6729–6736.
- Smith, R., Ventura, D., & Prince, J. T. (2015). LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Briefings in Bioinformatics*, 16(1), 104–117.
- Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3), 779–787.
- Sysi-Aho, M., Katajamaa, M., Yetukuri, L., & Orešič, M. (2007). Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC Bioinformatics*, 8, 93.
- Tautenhahn, R., Böttcher, C., & Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*, 9, 504.
- Theodoridis, G., Gika, H. G., & Wilson, I. D. (2008). LC-MS-based methodology for global metabolite profiling in metabonomics/metabolomics. *TrAC Trends in Analytical Chemistry*, 27(3), 251–260.
- Veselkov, K. A., Vingara, L. K., Masson, P., Robinette, S. L., Want, E., Li, J. V., et al. (2011). Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Analytical Chemistry*, 83(15), 5864–5872.
- Vinayavekhin, N., & Saghatelian, A. (2010). Untargeted metabolomics. *Current Protocols in Molecular Biology*. Hoboken, NJ: Wiley.
- Wang, J., Li, Z., Chen, J., Zhao, H., Luo, L., Chen, C., et al. (2013). Metabolomic identification of diagnostic plasma biomarkers in humans with chronic heart failure. *Molecular BioSystems*, 9(11), 2618–2626.
- Warrack, B. M., Hnatyshyn, S., Ott, K. H., Reily, M. D., Sanders, M., Zhang, H., et al. (2009). Normalization strategies for metabolomic analysis of urine samples. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, 877(5–6), 547–552.
- Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12), 1–20.
- Yin, P., & Xu, G. (2014). Current state-of-the-art of nontargeted metabolomics based on liquid chromatography–mass spectrometry with special emphasis in clinical applications. *Journal of Chromatography A*, 1374, 1–13.
- Zhang, W., Lei, Z., Huhman, D., Sumner, L. W., & Zhao, P. X. (2015). MET-XAlign: A metabolite cross-alignment tool for LC/MS-based comparative metabolomics. *Analytical Chemistry*, 87(18), 9114–9119.