Published in final edited form as:

Anal Chem. 2015 April 7; 87(7): 3606–3615. doi:10.1021/ac502439y.

Statistical methods for handling unwanted variation in metabolomics data

Alysha M. De Livera[†], **Marko Sysi-Aho**[‡,¶], **Laurent Jacob**[§], **Johann A. Gagnon-Bartsch**[||], **Sandra Castillo**[¶], **Julie A Simpson**[†], and **Terence P. Speed**[||,⊥,#]

Biostatistics Unit, Centre for Epidemiology and Biostatistics, University of Melbourne, VIC 3800, Australia, Zora Biosciences Oy, FIN-02150 Espoo, Finland, VTT Technical Research Centre of Finland, Finland, Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, CNRS, INRA, UMR5558, Villeurbanne, France, Department of Statistics, University of California, Berkeley, USA, Bioinformatics Division, Walter and Eliza Hall Institute, and Department of Mathematics and Statistics, University of Melbourne, VIC 3800, Australia

Alysha M. De Livera: alyshad@unimelb.edu.au

Abstract

Metabolomics experiments are inevitably subject to a component of unwanted variation, due to factors such as batch effects, long runs of samples, and confounding biological variation. Although the removal of this unwanted variation is a vital step in the analysis of metabolomics data, it is considered a gray area in which there is a recognised need to develop a better understanding of the procedures and statistical methods required to achieve statistically relevant optimal biological outcomes. In this paper, we discuss the causes of unwanted variation in metabolomics experiments, review commonly used metabolomics approaches for handling this unwanted variation, and present a statistical approach for the removal of unwanted variation to obtain normalized metabolomics data. The advantages and performance of the approach relative to several widely-used metabolomics normalization approaches are illustrated through two metabolomics studies, and recommendations are provided for choosing and assessing the most suitable normalization method for a given metabolomics experiment. Software for the approach is made freely available online.

Introduction

In analytical biochemistry, metabolomics is becoming an increasingly popular discipline, with its applications expanding to diverse research fields in the life sciences.[1] The study of metabolites and their responses to factors of interest such as physiological, environmental

Correspondence to: Alysha M. De Livera, alyshad@unimelb.edu.au.

[†]Biostatistics Unit, Centre for Epidemiology and Biostatistics, University of Melbourne, VIC 3800, Australia

[‡]Zora Biosciences Oy, FIN-02150 Espoo, Finland

[¶]VTT Technical Research Centre of Finland, Finland

[§]Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, CNRS, INRA, UMR5558, Villeurbanne, France

[||]Department of Statistics, University of California, Berkeley, USA

[⊥]Bioinformatics Division, Walter and Eliza Hall Institute

[#]Department of Mathematics and Statistics, University of Melbourne, VIC 3800, Australia

This material is available free of charge via the Internet at http://pubs.acs.org/.

and genetic conditions, allow for biological researchers to answer a range of sought-after scientific questions.[2] Typical aims in the statistical analysis of metabolomics data include the identification and quantification of metabolites, the discovery of differentially abundant metabolites between factors of interest (also known as “groups”), classification, clustering, and correlation analysis.[3]

In metabolomics experiments, the biological variation of interest is inevitably confounded with unwanted variation, often due to both the unwanted experimental and unwanted biological variability. Understanding the causes of unwanted variation in a given metabolomics experiment and the removal of this unwanted variation can pose a challenging task. This is further complicated by the fact that the unwanted variation can be unmeasurable, making it difficult to quantify the unwanted variation component. For example, a researcher may be interested in identifying metabolites present in urine which differentiate between certain disease types. In this situation, the varying concentration levels in metabolites indicating the amount of water they have had prior to obtaining urine samples, becomes unwanted variation in the biological samples (unwanted biological variation). Several practical examples of unwanted variation found in recent metabolomics literature are summarized in Figure 1.

In order to make inferences about the biological factors of interest, the overall unwanted variation component (as indicated in red in Figure 1, with unmeasurable examples of unwanted variation shown in italics) must either be accommodated appropriately in a statistical model which answers the research question or removed prior to further statistical analysis, ascertaining that the biological variation of interest are not affected nor removed. This is necessary to reduce the problems of falsely identifying differentially abundant metabolites, failing to identify truly differentially abundant metabolites, having spurious correlations between metabolites, artificial clustering and poor classification.

The metabolomics literature refers to the process of removing unwanted variation by various terms such as *signal drift correction*,[4] *batch effect removal*,[5] *scaling*,[6] and *removal of matrix effects*,[7] mostly referring to specific rather than overall unwanted variation encountered in an experiment, often handled separately in multiple steps.[8] The overall removal of unwanted variation (referred to as *normalization*) has been considered *a gray area in which there is a distinct need to develop a greater understanding of when, why, and how* [9] in order to achieve optimal biological outcomes. It has also been shown that the statistical results such as those obtained by identifying differentially abundant metabolites can vary depending on the method chosen for removing unwanted variation.[10] In this paper, we attempt to discuss these matters in detail.

The paper is organised as follows: In the next two sections, we review commonly used metabolomics approaches which have attempted to remove unwanted variation as explained above, and describe ways of choosing and assessing the effectiveness of a normalization method in a given metabolomics experiment. We then present a statistical approach for the overall removal of unwanted variation to obtain normalized metabolomics data. The advantages and performance of the proposed normalization approach relative to several widely-used metabolomics approaches are then illustrated with two metabolomics studies. A summary of the findings is presented in the final section.

A brief review of commonly used approaches

Use of scaling factors

Several well-known methods for removing unwanted variation in metabolomics data involve the use of various scaling factors. A scaling factor refers to a sample-specific constant which assigns an appropriate weight to each sample attempting to make them comparable. Normalizing by the median[11] or by the sum of squares[12] are two of the most commonly used methods, where each sample is scaled such that the median or the sum of squares of all abundances in a sample equals one respectively. Other similar methods include normalization to the total ion current[13] and normalization by unit norm.[14] The scaling methods are not applicable in general to most metabolomics experiments, as they rely on the *self-averaging* property.[15] It is assumed that an increase in the abundances of a group of metabolites in response to a perturbation is balanced by a decrease in abundances of metabolites in another group - an assumption which does not hold in many practical applications.[15,16] For instance, in a recent study involving obese and lean mice, the authors showed that adjusting individual liver lipid profiles of these mice using total signal incorrectly implies that there is a decrease in the levels of phospholipids in obese mice relative to the lean to balance the increased amount of triacylglycerols.[15]

Use of quality control samples

Certain forms of unwanted variation, such as the drift in signal over time and batch effect removal may also be handled using quality control samples.[4,17–20] These samples are composed of identical amounts of metabolites which are supposedly representative of those of the biological samples. Two types of quality control samples are being used in metabolomics studies: pooled biological quality control samples where each sample is a mixture of small aliquots of each biological sample present in the study, or externally purchased quality control samples where each sample is a mixture of small aliquots of multiple commercial samples. The applicability of the normalization methods based on quality control samples can be limited by practical considerations. Externally purchased quality control samples often do not have the same composition as the biological samples, posing challenges in the peak alignment process which can lead to a considerable amount of spuriously missing metabolite abundances. Although the pooled biological quality control samples are the closest to the composition of the biological samples, in many situations it is not possible to consume aliquots of preserved biological samples in order to prepare pooled quality control samples, and in large-scale studies where sample collection is not completed before sample preparation begins, preparing pooled quality control samples may not be possible.[4]

Use of internal standards

An alternative is to use internal standards which are known metabolites added to each biological sample before extraction. The simplest of such methods is the single internal standard (SIS) method, where a normalized data matrix is obtained by subtracting the log metabolite abundance of a single internal standard from the log abundances of the metabolites in each sample.[21] The variation captured by an internal standard however, depends on its own chemical properties,[22] and includes other sources of variation such as those which arise from chromatographical separation and ion suppression.[15,16] A slightly modified approach is to choose different internal standards according to the proximity of retention times to certain metabolite classes,[22] although retention time does not necessarily describe all chemical properties leading to unwanted variation.[15] The use of a single internal standard can lead to highly variable normalized values which depend on the compound that is used as the internal standard.[23] Consequently, recent literature has demonstrated that the use of a single internal standard is inadequate for removing unwanted variation, and has suggested the use of multiple internal standards in doing so[15,16,23]- a practice we support. Similar to the SIS method, a normalized data matrix can be obtained by subtracting the average of the multiple internal standards (AIS) on a log scale from the log abundances of the metabolites in each samples. More complex methods which use multiple internal standards include the NOMIS (Normalization using optimal selection of multiple internal standards) method,[15] where an optimal combination of multiple internal standards is selected using multiple linear regression, and the CCMN (Cross-contribution compensating multiple standard normalization) method,[16] where it is argued that the unwanted variation implied by the internal standard is influenced by contamination from the rest of the metabolites and this concept referred to as cross contribution is allowed for. The NOMIS method can be used in both supervised and unsupervised methods. When only one internal standard is available and the variation in the metabolite abundances highly correlate with the unwanted variation implied by the internal standard, it reduces to the SIS method. The CCMN method assumes that the factors of interest are known when adjusting the data, hence cannot be used in unsupervised methods.

Use of quality control metabolites

It was shown recently that in addition to internal standards, certain metabolites present in the biological samples which have been exposed to the unwanted variation but are unassociated with the factors of interest, may also be used as quality control metabolites.[23] A key advantage of doing so, is that it allows for unwanted biological variation to be accommodated, while retaining the essential biological variation of interest. In experiments exploring metabolomics changes in urine for example, unwanted variation in the form of varying concentration levels can be handled by using the concentration level information available in the quality control metabolites which are present in the urine samples. The RUV-2 (Remove unwanted variation-2) method[24] which uses quality control metabolites has been shown to perform very well for identifying differentially abundant metabolites, and was successfully applied to accommodate both unobserved and observed variation, to situations where the quality control samples are not available, and to systematically integrate data from different sources on the same quantities (e.g., interlaboratory studies).[23] However, the RUV-2 method which is based on a linear model designed for identifying differentially abundant metabolites, requires factors of interest to be known. These factors of interest are included in the linear model along with the factors of unwanted variation.[23,24] Hence, RUV-2 cannot be used prior to unsupervised analyses such as correlation, principal component or hierarchical cluster analysis. In this paper, using a variation of the RUV-2 method, denoted by *RUV-random* [25], we present an approach which is applicable in both supervised and unsupervised scenarios. We introduce methods for choosing parameters in RUV-random, and discuss how the estimation could be improved for exploratory clustering purposes. The differences between the approach presented here and those of Jacob et al[25] are described more fully in the proposed approach section. A summary indicating the applicability of the existing and proposed normalizing methods for removing unwanted variation in metabolomics data is given in Table 1.

Assessing the effectiveness of a normalization approach

As the statistical results of a metabolomics experiment can vary depending on the chosen normalization method,[10] assessing the effectiveness of a normalization method should form an integral part in the statistical analysis. In doing so, changes to both the variability and the bias[23] need to be explored. Closer replicates or smaller coefficients of variation do not guarantee that all variation of interest has remained and only unwanted variation has been removed. Further, an increase in the number of differentially abundant metabolites found after normalizing, does not necessarily imply the success of a normalization approach. For example, if the unwanted variation is correlated with the biological variation of interest, the removal of unwanted variation component may lead to a decrease in the number of differentially abundant metabolites. Hence, instead, we recommend using the following strategies (see Step 3 in Figure 1) for assessing whether the biological variation of interest is retained in the normalized data matrix.

The use of positive and negative control metabolites

In every experiment, there are facts that are known a priori, and these must be used to determine whether a normalization method has improved the analysis. For example, the biological background to an experiment (e.g., from previous literature or through the design of the experiment) can provide insight into metabolites that are known beforehand to be associated or unassociated with the biological factors of interest. We refer to such metabolites as our *positive control* and *negative control* metabolites respectively, distinguishing these from our *quality control* metabolites defined earlier. Following an appropriate normalization method, differentially abundant metabolites can be identified using a linear model fitted to each metabolite.[3] Subsequently, metabolites can be ranked using a suitable criterion such as the fold change or the t-statistic. If the biological variation of interest has not been removed with the removal of unwanted variation during the process of normalizing, we expect to find the positive control metabolites which are associated with the factors of interest to be near the top of the ranking, and do not expect for the negative control metabolites to appear near the top of this ranking.

The distribution of the p-values

Recent literature has discussed the usefulness of examining the distribution of p-values obtained from a differentially abundant analysis.[26] If there are no differentially abundant metabolites present in the dataset, the distribution of p-values should be uniform between zero and one. Hence, with the presence of some differentially abundant metabolites, a histogram of p-values should be uniformly distributed but with a peak close to zero. We will see in an application of this paper that when the data have not been normalized appropriately to remove any unwanted variation, the distribution of histogram can be far from ideal.

Visualization of the wanted and unwanted variation components

Together with the above mentioned assessments, visualization of the normalized data and the removed unwanted variation component can aid in determining whether a normalization method has improved the analysis. A simple and a very useful tool for doing this, is the use of relative log abundance (RLA) plots.[23] To obtain these, we firstly compute the median of each metabolite in the data matrix within (for within-group RLA plot) or across (for across-group RLA plot) the factors of interest, and then subtract this median from each metabolite. Sample-wise boxplots of this centred data matrix, then form the RLA plots. Within-group RLA plots should have a median close to zero and low variation around the median, and can be used to assess the tightness of the replicates. On the contrary, when a substantial proportion of the data matrix contains differentially abundant metabolites, across-group RLA plots may not have a median close to zero, but indicate the grouping structure. Here, we would expect to see low variation within each group. In addition to RLA plots, familiar multivariate techniques such as principal component analysis (PCA) and hierarchical cluster analysis (HCA) may also be used to explore both the biological factors of interest and preferably also the factors of unwanted variation.

Consistency of biological results obtained from complementary analyses

One can also monitor the consistency of results obtained from separate analyses from different metabolomics platforms (e.g., nuclear magnetic resonance spectroscopy, gas chromatography/mass spectrometry, and liquid chromatography/mass spectrometry). A good normalization method should lead to consistent overlapping results from the such independent analyses.[23]

Proposed approach

In the methods we present in this paper, we utilize the concept of quality control metabolites embedded in a linear mixed effects modelling framework. The former has several demonstrated advantages[23] which we have described briefly in the previous section, and the latter has been found to be a desirable alternative to fixed effects modelling in certain high dimensional settings.[27,28]

To describe the methods, we use the following notation:

- $\mathbf{Y}$ is a complete $m \times n$ matrix whose (i, j)th entry is y_{ij}, the log abundance of the jth metabolite in the ith sample, $i = 1, \dots, m$ and $j = 1, \dots, n$, where m is the number of samples and n is the number of metabolites.
- $\mathbf{Y}_\mathrm{c}$ is a $m \times n_c$ matrix containing the log abundances of the quality control metabolites, where n_c is the number of quality control metabolites.
- $\mathbf{X}$ is a $m \times p$ matrix containing for each sample, the p factors of interest.
- $\mathbf{W}$ is a $m \times k$ matrix containing for each sample, the k factors of unwanted variation.

- ε is a $m \times n$ matrix representing the unobserved error component.

We next introduce the linear mixed model given by,

$$\boldsymbol{Y}=\mathbf{X}\beta+\mathbf{W}\alpha+\varepsilon, \quad (1)$$

where, β and α are $p \times n$ and $k \times n$ matrices containing information on the effect of the factors of interest and unwanted variation respectively, on the log metabolite abundances. Further, it is assumed that $\varepsilon_j \overset{iid}{\sim} N(\mathbf{0}, \sigma_e^2 \mathbf{I}_m)$ and $\alpha_j \overset{iid}{\sim} N(\mathbf{0}, \sigma_\alpha^2 \mathbf{I}_k)$ for $j = 1, \ldots, n$.

If **X** and **W** are known, best linear unbiased estimates (BLUE) and predictors (BLUP) for β and α respectively, can be obtained using mixed model equation,[29]

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W}+\lambda\mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\boldsymbol{Y} \\ \mathbf{W}'\boldsymbol{Y} \end{bmatrix}, \quad (2)$$

where $\lambda=\sigma_e^2/\sigma_\alpha^2$.

However, as we discussed in the previous sections, the unwanted variation component **W** is often unobserved in practice. In addition, our goal is to provide a normalization method that is applicable in both supervised methods (where **X** is known) and unsupervised methods (where **X** is unknown). To normalize in cases like this, in what follows, we describe an approach based on RUV-random and its iterative versions introduced by Jacob et al[25] which have had considerable success in removing unwanted variation with applications to gene expression data. The approach presented here differs from Jacob et al in two ways: First, we use maximum likelihood estimation to aid the choice of parameters in the RUV-random method. Second, we improve the estimation using iteration for the purpose of clustering only, using information obtained from RUV-random normalized data, and update the parameters as appropriate. The approach is described as follows:

From Equation (1), for the control metabolites we have

$$\boldsymbol{Y}_c=\mathbf{W}\,\alpha_c+\varepsilon_c, \quad (3)$$

where $\varepsilon_{cj} \overset{iid}{\sim} N(\mathbf{0}, \sigma_e^2 \mathbf{I}_m)$ and $\alpha_{cj} \overset{iid}{\sim} N(\mathbf{0}, \sigma_\alpha^2 \mathbf{I}_k)$ for $j = 1, \ldots, n_c$. Notice that $\mathbf{Y}_c$ has a multivariate normal distribution given by $N(\mathbf{0}, \Sigma)$, where $\sum=\sigma_\alpha^2 \mathbf{W}\mathbf{W}'+\sigma_e^2 \mathbf{I}_m$. Thus, if **W** is known, the maximum likelihood estimators of the unknowns ($\sigma_e^2, \sigma_\alpha^2$) given the observed control metabolite abundances can be obtained by minimizing the quantity

$$\mathscr{L}(\sigma_\alpha^2, \sigma_e^2) n_c \log\left|\sum\right| + \mathrm{Trace}\left(\sum^{-1} \sum_{i=1}^{n_c} \mathbf{y}_{ci}\mathbf{y}_{ci}'\right), \quad (4)$$

where $\sum=\sigma_{\alpha}^{2}\mathbf{W}\mathbf{W}'+\sigma_{e}^{2}\mathbf{I}_{m}$ and $\mathbf{y}_{ci}$ is the *i*th row of $\mathbf{Y}_{c}$, subject to constraints $\sigma_{e}^{2}>0$ and $\sigma_{\alpha}^{2}>0$, using a non-linear optimization algorithm.[30]

Since **W** is unknown, we first estimate it using factor analysis on $\mathbf{Y}_c$.[23] This is the first step used in the RUV-2 method.[24] We let $\hat{\mathbf{W}} = \mathbf{U}\mathbf{D}_k$, where $\mathbf{Y}_c = \mathbf{U}\mathbf{D}\mathbf{V}'$ is the singular value decomposition of $\mathbf{Y}_c$, **U** and **V** are orthonormal matrices, and $\mathbf{D}_k$ is a diagonal matrix with the *k* largest singular values as its *k* first entries and 0 on the rest of the diagonal (the choice of k is discussed under the next subheading). Substituting this $\hat{\mathbf{W}}$ into Equation (4), we now obtain a rough estimate of α from Equation (2) as $\alpha \hat{=} (\hat{\mathbf{W}}'\hat{\mathbf{W}} + \lambda\hat{\mathbf{I}})^{-1}\hat{\mathbf{W}}'\mathbf{Y}$, where $\hat{\lambda}=\hat{\sigma}_{e}^{2}/\hat{\sigma}_{\alpha}^{2}$ assuming that **X** and **W** are not highly correlated. The normalized data matrix $\tilde{\mathbf{Y}}$ is then given by

$$\tilde{\boldsymbol{Y}}=\boldsymbol{Y}-\hat{\mathbf{W}}\hat{\alpha}=\boldsymbol{Y}-\hat{\mathbf{W}}\left(\hat{\mathbf{W}}'\hat{\mathbf{W}}+\hat{\lambda}\mathbf{I}\right)^{-1}\hat{\mathbf{W}}'\boldsymbol{Y}. \quad (5)$$

Judgement in the choice of $\hat{\lambda}$ and $\hat{k}$

The estimation of λ as explained in the above procedure requires a value for *k*. A useful approach for obtaining a suitable range of values for *k* is to look at a plot of the variance explained by the principal components of the control metabolite abundances (e.g., for the two applications in this paper, Figure I in the Supplementary Information shows the proportion of variance against the number of factors). In the work presented in this paper, as well as in our previous work,[23,24] we have found that this visual inspection, together with the criteria explained in the previous section (Step 3 in Figure 1), leads to good values for *k*.

It is important to understand that for a given dataset, there is no "optimal" λ and *k*. For example, even on the same dataset, "optimal" λ and *k* can change in the light of what variation need to be removed from the dataset (see Step 1 in Figure 1), how well the control metabolites capture this unwanted variation (discussed under the next subheading), and the type of analysis a researcher wishes to perform on the dataset (e.g., in the "Applications" section, we will see that different $\hat{W}$, λ and *k* are obtained when the purpose is clustering). Thus, it is necessary that both λ and *k* are chosen using context-specific knowledge, also allowing for model misspecifications and imperfect controls.

Choosing quality control metabolites

The role of quality control metabolites must be well-understood by the users of these RUV methods. It is required that the quality control metabolites be associated with the unwanted variation that needs to be removed or accommodated and be unassociated with the factor of interest. In this sense, multiple internal standards play an important role in removing the unwanted experimental variation shown in Figure 1.

In the absence of sufficient internal standards, possible quality control metabolites need to be found empirically. For supervised analyses where **X** is known, this can be done iteratively. For example, one can use the few available internal standards as the initial set of known quality control metabolites to normalize the data in the first instance, and then identify a further set of empirical quality control metabolites using p-value and/or fold-change rankings. For unsupervised analysis where **X** is unknown, in addition to the internal standards, one may find nearest-neighbour metabolites to the internal standards using an appropriate distance matrix, or simply find those which correlate highly with the internal standard[23] or the average of the internal standards if multiple internal standards are available (an upper percentile of these correlations may be used as a cut-off). Care needs to be taken that the empirical quality control metabolites are not influenced by the factors of interest. A prior exploratory analysis of the quality control metabolites can be helpful in visualising the unwanted variation captured in these quality control metabolites.

Another delicate point to consider is that, at their best, the internal standards will only remove the unwanted experimental variation. If the goal is to remove unwanted biological variation (see Figure 1), one needs to include as quality control metabolites those which are present in the biological samples and are associated with the unwanted biological variation. For example, in a study involving *Leishmania mexicana*- a sandfly transmitted parasitic protozoan, we have previously shown[23] that this can be achieved using prior biological information regarding the experiment. Thus, in removing unwanted biological variation in particular, it is vital to have a good biological knowledge of the research area, the question of interest, background of the experiment, and hence, the use of positive and negative control metabolites.

Improving the estimation for exploratory approaches

For exploratory clustering approaches, the RUV-random normalization method may be improved by incorporating iterative estimation[25] of the components of the overall model given by Equation (1). To do so, we first obtain estimates for the components **W** and α using the RUV-random method. We then iterate the following two steps: (i) Using $\mathbf{Y} - \hat{\mathbf{W}}\hat{\alpha}$, estimate the component $\mathbf{X}\beta$ by using the *k*-means algorithm described by Hartigan et al,[31] (ii) With this estimate for $\mathbf{X}\beta$, obtain a refined estimate for α given the estimates for **W**, λ and *k* using $\boldsymbol{Y} - \hat{\mathbf{X}}\beta$ (In the applications of this paper we carried out 200 iterations). In addition, during the iteration, we obtain updated estimates for **W**, λ and *k* intermittently as appropriate (In the applications of this paper we updated these estimates every 100 iterations).

In order to avoid artificial clustering, we employ several strategies: First, we initialize the iteration procedure by using the *k*-means algorithm on the RUV-random normalized data. Hence, we can reasonably assume that this normalized data matrix predominantly contains the biological variation of interest, which is then captured with the use of the *k*-means algorithm. Second, we carry out exploratory clustering analysis on the RUV-random normalized data prior to iteration. This helps in detecting the number of clusters to be included in the *k*-means algorithm, as well as identifying any peculiar clustering that can occur after iteration. Further, we explore the removed unwanted variation component to gain insight into what has been discarded.

Thus, this iteration procedure should not be considered as a separate method on its own and must be used in conjunction with the RUV-random normalization only to improve the exploratory clustering analyses when required.

Applications

We now use two metabolomics studies to demonstrate the performance of the proposed approach in obtaining a normalized metabolomics data matrix. We assess our approach relative to several commonly used metabolomics normalization approaches, in regards to its applicability in situations where widely-used metabolomics methods are not applicable and the ability to remove unwanted variation while preserving the biological variation of interest. In these applications, when normalizing the datasets using RUV-random, we treated the known biological factors of interest as unknown. The known biological factors of interest were only used for the purpose of comparing the performance across the different normalization methods. Unless otherwise stated, the applications presented in this section were carried out using R software.[32]

Application to a multi-site study

This application involves a comparative study, originally designed to compare the quality of the instrumental performance across four different laboratory sites. The dataset consists of two different metabolite mixtures: a mixture which contains 33 known metabolites (Mix I), and another mixture (Mix II) containing the same metabolites with some at higher concentrations. In Mix II, eleven metabolites were at three-fold and one metabolite was at five-fold concentration relative to Mix I (spiked-in metabolites), while the other twenty-one metabolites remained unchanged. The dataset contained only one internal standard and no quality control samples, which is typical of most practical metabolomics experiments we encounter. Eight replicates from each mixture were run in three different laboratory sites, on four different GC-MS instruments, at three different temperature (7,15, and 25°C) settings. A detailed description of the analytical methods is published elsewhere.[23]

Since only one internal standard is available, without any quality control samples being available, and the self-averaging property clearly does not hold for this dataset, out of the metabolomics normalization methods we have described, only the SIS and RUV methods are applicable to normalize this data (see Table 1). Hence, we compared the performance of the RUV-random method with the SIS normalized data. We chose the $n_c = 9$ metabolites that have a correlation of greater than 0.9 with the internal standard (approximately the 70th percentile in this application) as quality control metabolites, and did not use the knowledge that twenty one metabolites were at constant levels throughout the study. The parameters k and λ for the RUV approach were chosen by following the approach presented in the "Proposed approach" section.

The samples are plotted in the space of the first three principal components of the unadjusted in Figure 2 (a), and it is seen that the predominant variation arises from clear concentration differences, as the samples cluster by mixture type in the space of the first two principal components. Hence, for exploratory clustering analysis, a researcher may wish to retain the unwanted variation in the unadjusted data for manual exploration. Here we demonstrate the use of the RUV-random method for situations where unwanted variation needs to be removed for improved clustering of the biological variation of interest. In the data normalized using SIS and RUV-random without iteration, unwanted variation is still visible in the form of temperature and instruments (see Figure II in the Supplementary Information). Much improved performance is achieved by the improved RUV-random method for clustering as shown by Figure 2 (b).

The variation due to mixture type is explained by the second principal component in the unadjusted and SIS normalized data, while in the RUV-random normalized data, this variation is explained by the first principal component. These conclusions are strengthened by the visual inspection of the within-group RLA plots shown in Figure 3. Here, each data matrix was centred by finding the median of each metabolite within the factor of interest (e.g., within each Mix I and Mix II samples) and subtracting it from each metabolite. If the unwanted variation was negligible, the boxplots of the samples of these centred metabolites should have a median close to zero and low variation around this median. It is seen from Figure 3 that the RUV-random method improved for clustering has succeeded in achieving such stable samples by removing most of the unwanted variation- a considerable improvement relative to the unadjusted, SIS and RUV-random normalized data without iteration.

Since the unwanted variation component has been treated as unobserved in the RUV methods, exploring whether the effects captured in the estimated unwanted variation component are associated with known factors of unwanted variation can also be helpful in assessing the quality of these methods in removing unwanted variation. For instance, Figure III (a) shown in Supplementary Information illustrates the unwanted variation captured by the first column of the matrix $\hat{\mathbf{W}}$. The boxplots of the samples indicate that this consists of both the instrument and temperature effects, and from the analysis of variance these effects were found to be significantly different between the samples with p-values < 10^{-3}. The first three principal components of $\hat{\mathbf{W}}\hat{\alpha}$, that is the estimated unwanted variation component removed by the RUV-random method improved for clustering, are shown in Figure III (b), confirming that the removed unwanted variation consists of known factors, which were treated as unobserved throughout the analysis.

In order to evaluate the performance of normalization methods, in addition to assessing whether the unwanted variation have been removed, it is necessary to ascertain that the biological variation of interest is preserved in the normalized data.[3,23–25,33] Firstly, we used hierarchical cluster analysis on the samples and metabolites separately to assess the tendency of the data to accurately identify the two groups which they should belong to. In doing so, Manhattan distance was used and the variables were partitioned using Ward's minimum variance agglomerative clustering.[34] Here, the samples are expected to cluster by mixture type, while by metabolites the spiked-in and non spiked-in metabolites are expected to cluster separately. Clustering error was quantified as the number of misclassified samples and metabolites. Secondly, we used a linear model fitted to each metabolite to obtain estimates of the log fold changes, in order to assess the ability of the methods to assist in identifying differentially abundant metabolites. The results are summarised in Table 2.

Since the most prominent variation in this dataset is due to mixture type, the number of misclassified samples were zero for all normalization methods. The number of misclassified metabolites varied, and was the lowest for the data normalized by the RUV-random method. Similarly, the mean square error obtained by comparing the true fold changes with those obtained from normalized data was the lowest for the RUV-random method. Supplementary Information Figures IV, V and Table I show that for this analysis no substantial gain was found by varying k and λ from the estimated values.

Application to a multi-cohort study

The dataset is taken from a multi-cohort study designed to explore the associations between lipid levels and those of small polar metabolites with a disease related to metabolic syndrome. This subset of the data consists of serum samples from six independent cohorts of 591 healthy individuals differing in age, body mass index (BMI), and gender. The samples were analyzed by liquid chromatography-mass spectrometry, using the analytical method described by Nygren et al.[35] These were run in six batches, and the running order within each of the six batches was randomized. Five internal standards were used in the experiment, and externally purchased Seronorm samples [1] were run within the six batches. Pre-processing was carried out using MZmine 2 software version 2.10.[36]

Figure 4 (a) shows the first three principal components of the raw data, indicating strong batch variation. In addition, the physiological states of the individuals varied with each cohort. Pairwise relationships between the variables age, BMI, gender, and batch can be visualized using a generalised pairs plot[37] which allows for the simultaneous inspection of both categorical and quantitative information in the data. Figure 4 (b) shows that the individuals whose serum samples were run in batches 3, 6, and 7 have non-overlapping age ranges, and that the BMI is higher in the individuals whose serum samples were run in batch 2. Further, BMI information is missing in the individuals belonging to batches 3 and 6, and gender information was not available for batch 5. Thus, the removal of the unwanted batch/ cohort variation in this study is complicated by the fact that it is confounded with the physiological states of the individuals.

The data was normalized using the methods described earlier to remove unwanted batch variation, while preserving age, gender, and BMI information of the individuals. The five internal standards were used when applying AIS and NOMIS methods. As quality control metabolites for the RUV-random method, we used metabolites which correlated highly with the average of the internal standards. A correlation of 0.6 (approximately the 70th percentile) was used for this purpose. In this application, this corresponded to selecting n_c = 32 quality control metabolites out of a total of n = 129 metabolites. The parameters k = 10 and λ = 0.31 was chosen using the approach described in the ‘Proposed approach’ section. As the variation in age, gender, and BMI were retained in the normalized data matrix, clear clustering in the normalized data was not anticipated in an exploratory analysis. Hence we did not carry out iteration. The use of Seronorm samples in the methods based on quality control samples, however, posed several difficulties. There were no metabolites present in

[1] SeroNorm Human, Sero AS, Norway, www.sero.no

the Seronorm samples which were also present in all of the biological samples in all batches. Thus, the Seronorm and biological samples needed to be pre-processed separately within each batch, which gave rise to a substantial number of missing peaks leading to loss of information.

Figure 5 shows the first two principal components of the data normalized by the methods which performed the best in the analysis of this data. In the data normalized by the AIS, MEDIAN, and NOMIS methods, clustering due to batch effect is clearly seen in the first two principal components and substantial unwanted variation still remains in the RLA plots (see Figures VI in Supplementary Information). The RUV-random method shows no visible unwanted batch variation in either the PCA or the RLA plots, and the NOMIS method seems to perform the next best. Figure VII in Supplementary Information illustrates that the estimated unwanted variation component removed by the RUV-random method consists of batch variation which was treated as unobserved.

In order to assess whether biological variation has been removed in the process of removing the batch variation, we used information from the literature[38] on metabolites which have been found to be differentially abundant between males and females. We found this way, five gender-specific metabolites[38] which overlapped with the list of metabolites observed in our dataset. We used these five gender-specific metabolites and the five internal standards as our positive and negative control metabolites respectively. By fitting a linear model to each metabolite consisting of a design matrix with gender, age and BMI information, we then examined behaviour of these positive control metabolites in each normalized dataset.

Figure 6 (a) shows the volcano plots obtained from this analysis for the unadjusted data and the data normalized by the RUV-random method. The metabolites which are known[38] to be increasing (LPC20:3, LPC20:4, LPC20:5) and decreasing (SM34:2, SM36:2) in males compared to females appeared with high absolute fold changes and low p-values in the RUV-random normalized data. These metabolites are shown in red and blue respectively, along with green points denoting the internal standards which had low fold changes and high p-values in all three variable comparisons (gender, age, and BMI) in the RUV-random normalized data. In contrast, the rankings for these known metabolites obtained using the other normalization methods were less satisfactory (see Figure VIII in the Supplementary Information).

For each comparison, the p-value histograms for unadjusted and RUV-random normalized data are shown in Figure 6 (b). In the presence of some differentially expressed metabolites, we would expect p-value histograms to be uniformly distributed with a peak close to zero containing the differentially expressed metabolites.[3,26] In the unadjusted data, a large proportion of metabolites have p-values close to zero (including the 5 internal standards in two of the comparisons) implying metabolites falsely identified as differentially abundant, while p-value histograms which are closer to the ideal were obtained from the RUV-random normalized data.

Similar to the multi-site study, for this analysis, no substantial gain was found by varying k and λ from the estimated values (see for example, Figures IX and X).

Conclusions

In this paper, we presented a statistical method based on a linear mixed effects model which utilises quality control metabolites to obtain normalized data in typical metabolomics experiments. The approach, which can be applied without relying on particular experimental requirements such as having quality control samples, attempts to accommodate unwanted biological variation without removing the essential biological variation of interest, and captures unwanted variation which is not observed. The unwanted variation component which has been removed by the model may be explored in order to gain insight into any undetected experimental or biological variation. We illustrated the improved performance of the approach relative to several existing widely-used metabolomics normalization methods. In addition, we provided a brief review of the existing metabolomics normalization methods and recommendations for choosing and assessing the effectiveness of a normalization method, with particular emphasis placed on the importance of using positive control metabolites to ascertain that biological variation of interest are not removed along with the removal of unwanted variation. The software for the approaches described in the paper are available online.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Monteiro MS, Carvalho M, Bastos ML, Guedes de Pinho P. Current medicinal chemistry. 2013; 20:257–71. [PubMed: 23210853]
2. Armitage EG, Barbas C. Journal of pharmaceutical and biomedical analysis. 2014; 87:1–11. [PubMed: 24091079]
3. De Livera AM, Olshansky M, Speed TP. Methods in molecular biology (Clifton, NJ). 2013; 1055:291–307.
4. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, Brown M, Knowles JD, Halsall A, Haselden JN, Nicholls AW, Wilson ID, Kell DB, Goodacre R. Nature protocols. 2011; 6:1060–1083. [PubMed: 21720319]
5. Wang S-Y, Kuo C-H, Tseng YJ. Analytical chemistry. 2013:1037–46. [PubMed: 23240878]
6. Craig A, Cloarec O, Holmes E, Nicholson JK, Lindon JC. Analytical chemistry. 2006:2262–7. [PubMed: 16579606]
7. Hall TG, Smukste I, Bresciano KR, Wang Y, McKearn D, Savage RE. Intech Europe. 2012:389–420.
8. Kirwan JA, Weber RJ, Broadhurst DI, Viant MR. Scientific Data. 2014:1.
9. Roessner, U.; Nahid, A.; Chapman, B.; Hunter, A.; Bellgard, M. Comprehensive Biotechnology. 2. Vol. 1. Elsevier B.V; 2011. p. 447-460.
10. Temmerman L, De Livera AM, Bowne J, Sheedy RJ, Callahan DL, Nahid A, De Souza D, Schoofs L, Tull DL, McConville JM, Roessner U, Went-worth JM. Journal of Diabetes & Metabolism. 2012
11. Wang W, Zhou H, Lin H, Roy S, Shaler TA, Hill LR, Norton S, Kumar P, Anderle M, Becker CH. Analytical chemistry. 2003; 75:481848–26.
12. Crawford LR, Morrison JD. Analytical chemistry. 1968; 40:1464–1469.
13. Cairns, Da; Thompson, D.; Perkins, DN.; Stanley, AJ.; Selby, PJ.; Banks, RE. Proteomics. 2008; 8:21–7. [PubMed: 18095358]

14. Scholz M, Gatzek S, Sterling A, Fiehn O, Selbig J. Bioinformatics (Oxford, England). 2004; 20:2447–54.
15. Sysi-Aho M, Katajamaa M, Laxman Y, Oresic M. BMC bioinformatics. 2007; 8:93. [PubMed: 17362505]
16. Redestig H, Fukushima A, Stenlund H, Moritz T, Arita M, Saito K, Kusano M. Analytical chemistry. 2009; 81:7974–7980. [PubMed: 19743813]
17. Gika HG, Macpherson E, Theodoridis Ga, Wilson ID. Journal of chromatography B, Analytical technologies in the biomedical and life sciences. 2008; 871:299–305. [PubMed: 18579458]
18. Zelena E, Dunn WB, Broadhurst D, Francis-McIntyre S, Carroll KM, Begley P, O'Hagan S, Knowles JD, Halsall A, Wilson ID, Kell DB. Analytical chemistry. 2009; 81:1357–64. [PubMed: 19170513]
19. Lai L, Michopoulos F, Gika H, Theodoridis G, Wilkinson RW, Odedra R, Wingate J, Bonner R, Tate S, Wilson ID. Molecular bioSystems. 2010; 6:108–20. [PubMed: 20024072]
20. Kamleh MA, Ebbels TMD, Spagou K, Masson P, Want EJ. Analytical chemistry. 2012; 84:2670–7. [PubMed: 22264131]
21. Gullberg J, Jonsson P, Nordström A, Sjöström M, Moritz T. Analytical biochemistry. 2004; 331:283–95. [PubMed: 15265734]
22. Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, Macdonald I, Van Ommen B, Smilde AK. Analytical chemistry. 2006; 78:567–74. [PubMed: 16408941]
23. De Livera AM, Dias DA, De Souza D, Rupasinghe T, Pyke J, Tull D, Roessner U, McConville M, Speed TP. Analytical chemistry. 2012:10768–76. [PubMed: 23150939]
24. Gagnon-Bartsch JA, Speed TP. Biostatistics. 2012; 13:539–52. [PubMed: 22101192]
25. Jacob L, Gagnon-Bartsch JA, Speed TP. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. 2013
26. Leek JT, Storey JD. PLoS genetics. 2007; 3:1724–1735. [PubMed: 17907809]
27. Listgarten J, Kadie C, Schadt EE, Heckerman D. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107:16465–70. [PubMed: 20810919]
28. Jauhiainen A, Basetti M, Narita M, Narita M, Griffiths J, Tavaré S. Bioinformatics (Oxford, England). 2014:1–7.
29. Henderson, CR. Applications of Linear Models in Animal Breeding. 1984.
30. Nelder JA, Mead R. The computer journal. 1965; 7:308–313.
31. Hartigan J, Wong M. Journal of the Royal Statistical Society Series C. 1979; 28:100–108.
32. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2014.
33. Gagnon-Bartsch, JA.; Jacob, L.; Speed, TP. IMS Monographs. 2014. Removing unwanted variation from high dimensional data with negative controls. Accepted for publication
34. Becker, RA.; Chambers, JM.; Wilks, AR. The new S language. Pacific Grove, Ca: Wadsworth & Brooks; 1988.
35. Nygren H, Seppänen-Laakso T, Castillo S, Hyötyläinen T, Orešič M. Methods in molecular biology (Clifton, NJ). 2011; 708:247–57.
36. Pluskal T, Castillo S, Villar-Briones A, Oresic M. BMC bioinformatics. 2010; 11:395. [PubMed: 20650010]
37. Emerson J, Green W. Journal of Computational and Graphical Statistics. 2013; 22:79–91.
38. Weir JM, Wong G, Barlow CK, Greeve MA, Kowalczyk A, Almasy L, Comuzzie AG, Mahaney MC, Jowett JBM, Shaw J, Curran JE, Blangero J, Meikle PJ. Journal of lipid research. 2013; 54:2898–908. [PubMed: 23868910]

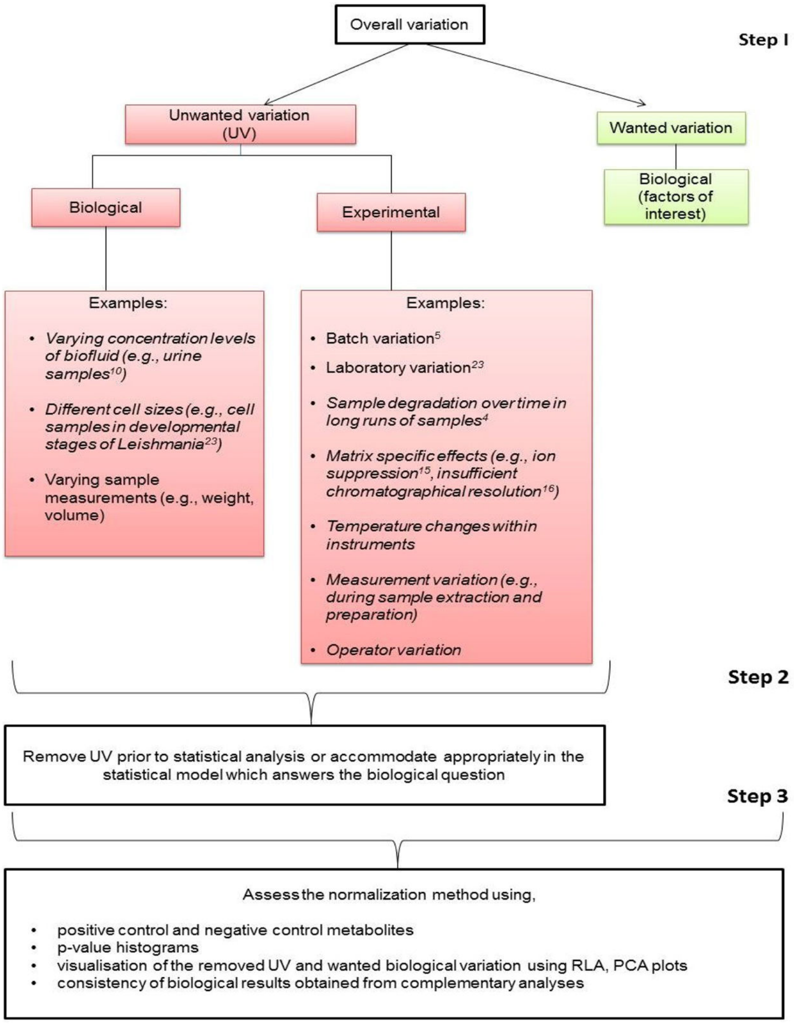

Overall variation
Step I
Unwanted variation (UV)
Wanted variation
Biological
Experimental
Biological (factors of interest)
Examples:
• Varying concentration levels of biofluid (e.g., urine samples[10])
• Different cell sizes (e.g., cell samples in developmental stages of Leishmania[23])
• Varying sample measurements (e.g., weight, volume)
Examples:
• Batch variation[5]
• Laboratory variation[23]
• Sample degradation over time in long runs of samples[4]
• Matrix specific effects (e.g., ion suppression[15], insufficient chromatographical resolution[16])
• Temperature changes within instruments
• Measurement variation (e.g., during sample extraction and preparation)
• Operator variation
Step 2
Remove UV prior to statistical analysis or accommodate appropriately in the statistical model which answers the biological question
Step 3
Assess the normalization method using,
• positive control and negative control metabolites
• p-value histograms
• visualisation of the removed UV and wanted biological variation using RLA, PCA plots
• consistency of biological results obtained from complementary analyses

Figure 1.

A graphical representation of the steps involved in the process of normalizing data from a typical metabolomics experiment. The first step involves identifying overall sources of variation. Here, the unwanted variation component is shown in red, and the unmeasurable unwanted variation examples are shown in italics. The second step involves normalizing (either removing the overall unwanted variation component or accommodating it in an appropriate statistical model). The third step involves assessing the normalizing method.

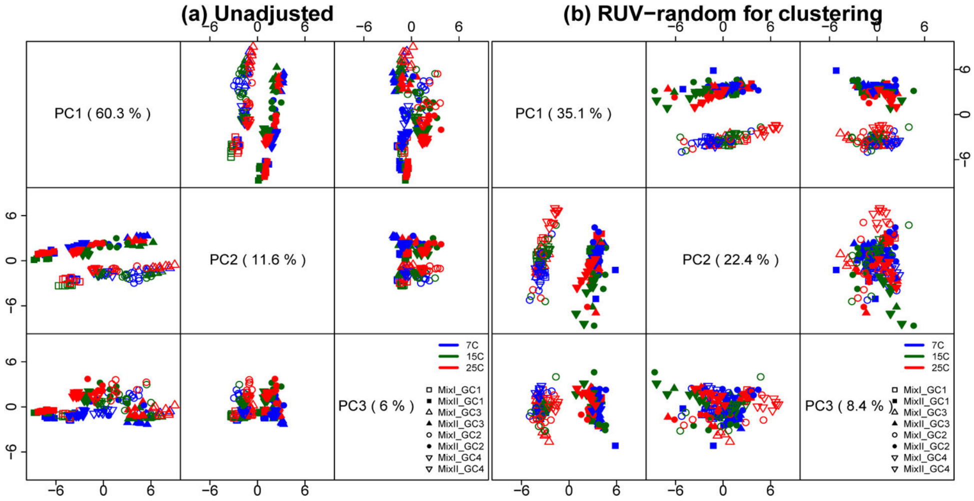


Figure 2.

The first three principal components of the (a) unadjusted data, and (b) the data normalized by RUV-random improved for clustering ($k = 8$, $\lambda = 1.43$). The shapes and colours indicate different instruments and temperatures respectively, and Mix I and Mix II samples are shown by the hollow and solid points respectively.

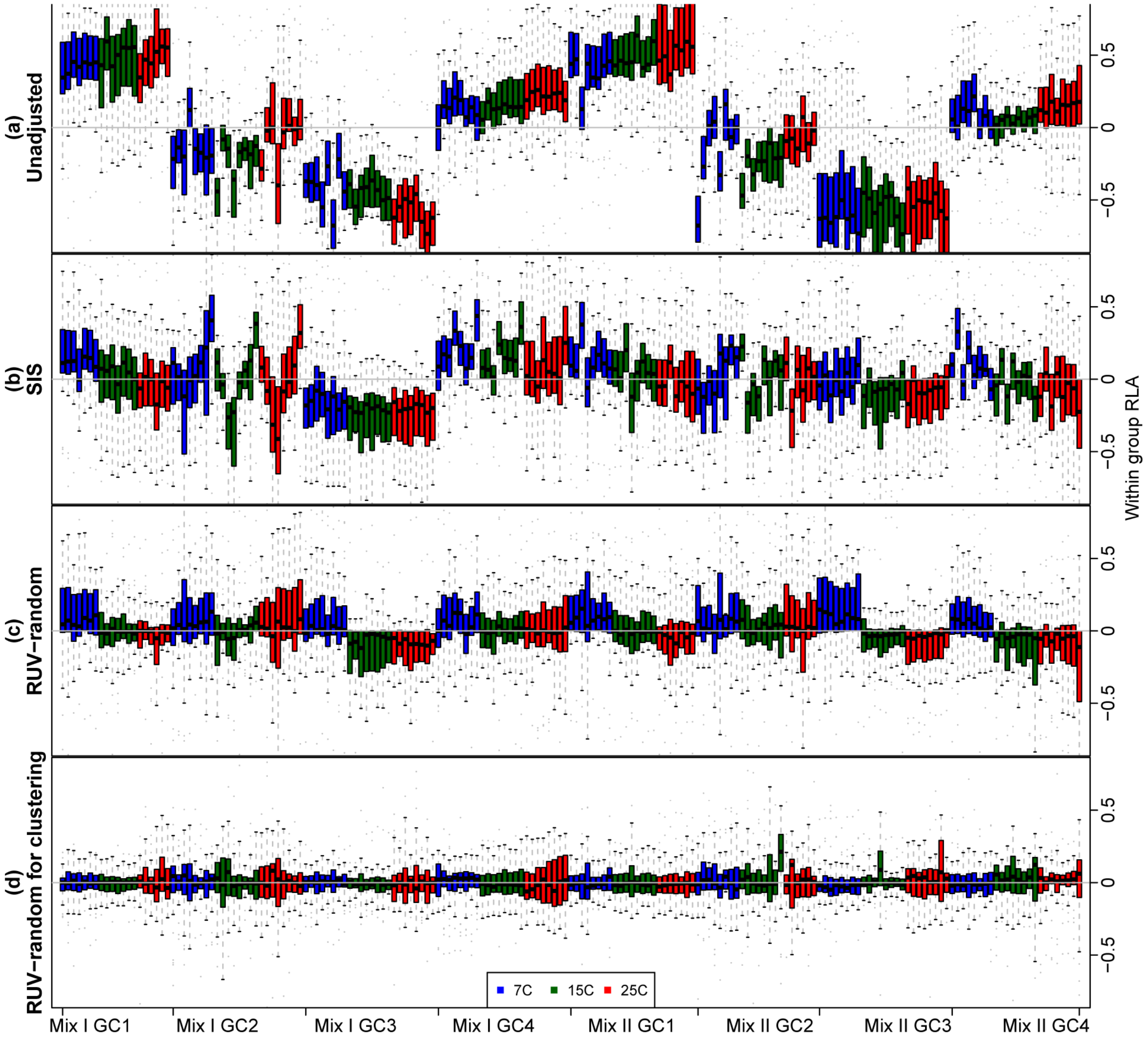
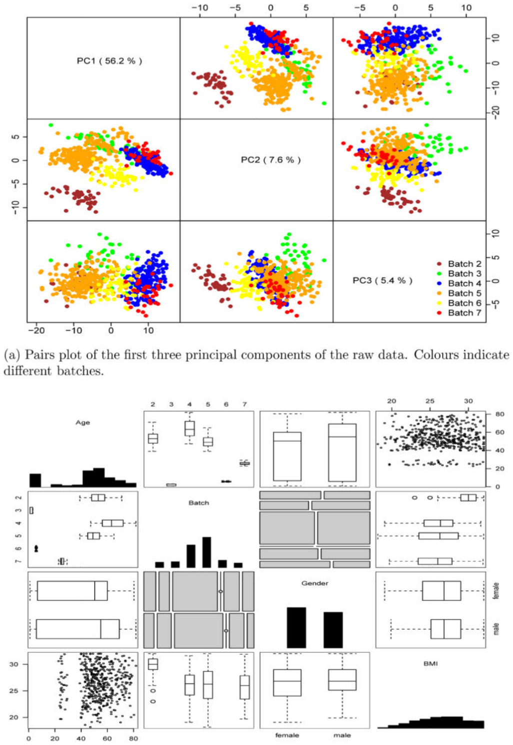


Figure 3.

Within-group RLA plots of the (a) unadjusted data, and the data normalized by the (b) SIS (c) RUV-random ($k = 3$, $\lambda = 0.03$) and (d) RUV-random improved for clustering ($k = 8$, $\lambda = 1.43$). The colours represent different temperatures.



(a) Pairs plot of the first three principal components of the raw data. Colours indicate different batches.

(b) Generalised pairs plot[37] of the variables age, batch, gender and BMI.

Figure 4.

Plots showing (a) the first three principal components and (b) generalised pairs plot[37] of the variables age, batch, gender and BMI (The diagonal panels show the marginal distribution of each variable, and off-diagonal panels display pairwise relationships between the quantitative (age, BMI) and categorical (gender, batch). Scattter plots, boxplots, and mosaic plots are used to represent respectively, the relationship between two quantitative variables, between a categorical and a quantitative variable, and between two categorical variables. In the mosaic plots, areas are proportional to counts.)

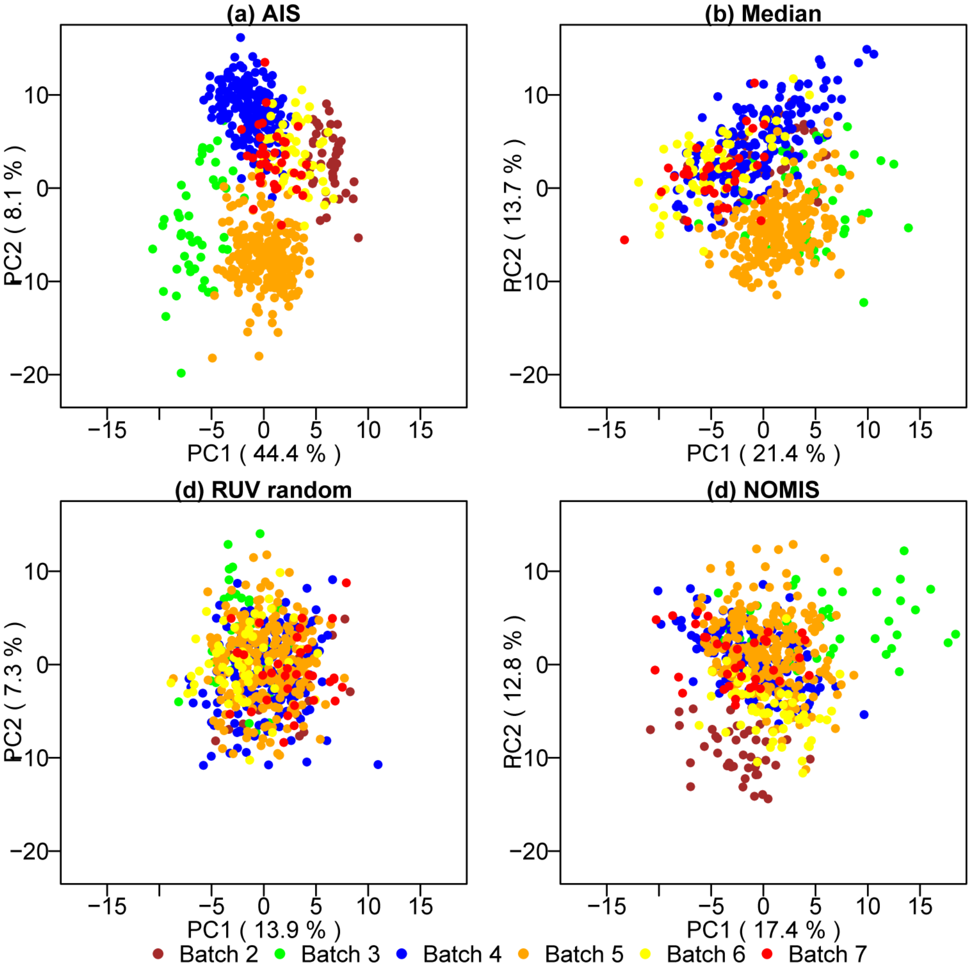


Figure 5.

Figures showing the first two principal components of the unadjusted and normalized data. Colours indicate different batches.

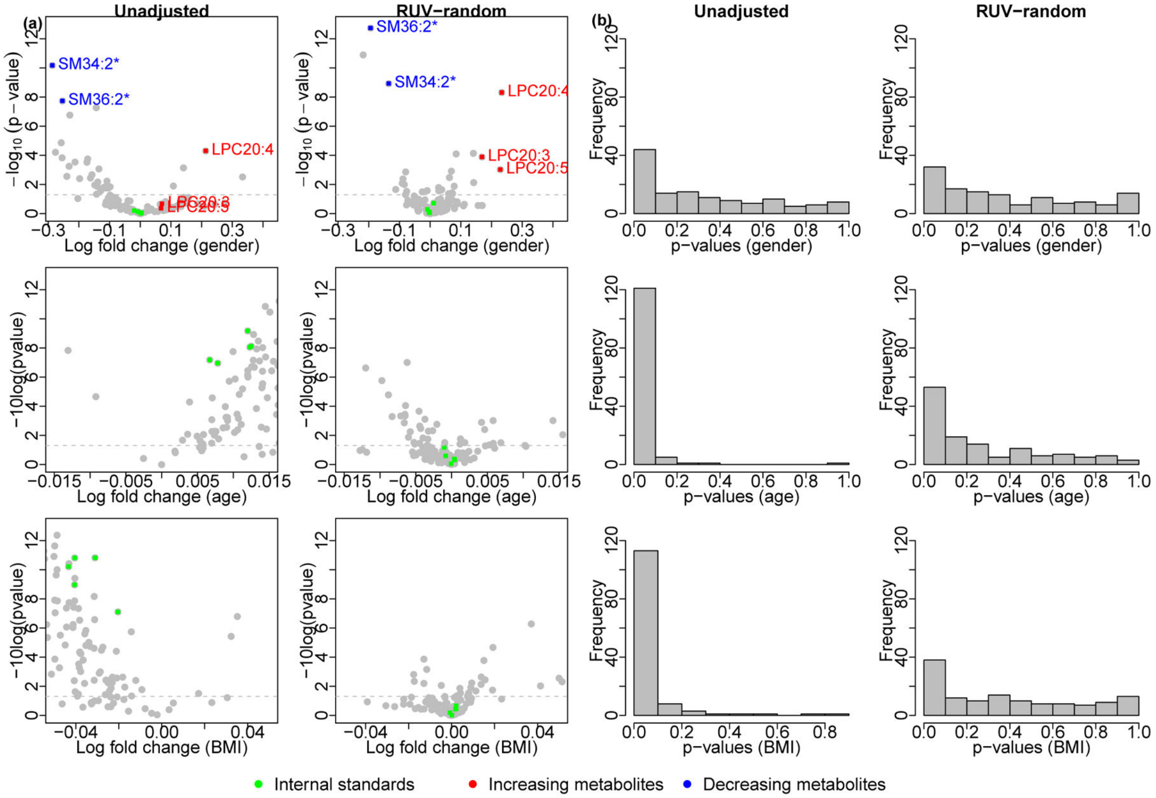


Figure 6.

Figures showing (a) volcano plots, and (b) histograms of p-values for unadjusted and RUV-random normalized data.

Table 1

A summary indicating the applicability of normalizing methods for removing unwanted variation in metabolomics data as described in the section 'A brief review of commonly used approaches'.

Method	**Applicability**
Scaling methods (e.g., normalizing by the median (MEDIAN), total ion signal)	Not suitable when the self-averaging property does not hold Applicable in both supervised and unsupervised methods
Using a single internal standard (SIS) Using internal standards according to retention time	Cannot remove unwanted biological variability Leads to highly variable normalized values Applicable in both supervised and unsupervised methods
Using the average of multiple internal standards (AIS) Normalization using optimal selection of multiple internal standards (NOMIS)	Cannot remove unwanted biological variability Applicable in both supervised and unsupervised methods
Cross-contribution compensating multiple standard normalization (CCMN)	Cannot remove unwanted biological variability Cannot be used in unsupervised methods
Using quality control samples	Cannot remove unwanted biological variability Can lead to spuriously missing metabolites Applicable in both supervised and unsupervised methods
Remove unwanted variation-2 (RUV-2)	Cannot be used in unsupervised methods Attempts to remove overall unwanted variation shown in Figure 1, given suitable controls
Remove unwanted variation-random (RUV-random)	Applicable in both supervised and unsupervised methods Attempts to remove overall unwanted variation shown in Figure 1, given suitable controls
RUV-random improved for clustering	Can only be used for unsupervised exploratory clustering purposes, in conjunction with the RUV-random method Attempts to remove overall unwanted variation shown in Figure 1, given suitable controls

Table 2

The number of misclassified samples and metabolites obtained from the hierarchical cluster analysis, and the mean square error obtained by comparing the true fold changes with the estimated fold changes for unadjusted and normalized data.

	Number of misclassified		**Mean Square Error**
	Samples	**Metabolites**	
Unadjusted	44	13	13.2
SIS	0	12	12.7
RUV-random (k = 3, λ = 0.03)	0	4	11.6
RUV-random improved for clustering (k = 8, λ = 1.43)	0	3	-
RUV2 (k = 6)	-	-	11.4