BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 15 2014, pages 2155–2161
doi:10.1093/bioinformatics/btu175

Gene expression

Advance Access publication April 7, 2014

# Normalization of metabolomics data with applications to correlation maps

Alexandra Jauhiainen[1,*], Basetti Madhu[2], Masako Narita[2], Masashi Narita[2], John Griffiths[2] and Simon Tavaré[2]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, PO Box 281, SE-171 77 Stockholm, Sweden and [2]Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK

Associate Editor: Ziv Bar-Joseph

## ABSTRACT

**Motivation:** In metabolomics, the goal is to identify and measure the concentrations of different metabolites (small molecules) in a cell or a biological system. The metabolites form an important layer in the complex metabolic network, and the interactions between different metabolites are often of interest. It is crucial to perform proper normalization of metabolomics data, but current methods may not be applicable when estimating interactions in the form of correlations between metabolites. We propose a normalization approach based on a mixed model, with simultaneous estimation of a correlation matrix. We also investigate how the common use of a calibration standard in nuclear magnetic resonance (NMR) experiments affects the estimation of correlations.

**Results**: We show with both real and simulated data that our proposed normalization method is robust and has good performance when discovering true correlations between metabolites. The standardization of NMR data is shown in simulation studies to affect our ability to discover true correlations to a small extent. However, comparing standardized and non-standardized real data does not result in any large differences in correlation estimates.

**Availability and implementation:** Source code is freely available at https://sourceforge.net/projects/metabnorm/

**Contact:** alexandra.jauhiainen@ki.se

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

Received on August 30, 2013; revised on February 26, 2014; accepted on March 28, 2014

## 1 INTRODUCTION

In metabolomics, the ultimate goal is to determine quantitatively the levels of all metabolites (small molecules) in a biological sample. At present, it is not possible to measure the complete metabolome, and, depending on the purpose of the experiment, the number of features measured can range from a handful to a couple of thousands. Usually, in *targeted* experiments, a small number of features are measured, and the metabolites are completely identified. In *untargeted* experiments, many more features are quantified, but a majority of them cannot be identified as a specific metabolite. Subsets of the metabolome are sometimes referred to as metabolic profiles (Chung *et al.*, 2003).

Several methods are at our disposal for metabolomic analyses, e.g. nuclear magnetic resonance (NMR) spectroscopy or gas chromatography coupled to mass spectrometry. In this article, we focus on metabolomics data from NMR experiments. When quantifying metabolite levels by using proton NMR, the protons in a molecule give rise to a signal that appears as one or several peaks in a spectrum. Because of spin-quantum effects, peaks split in different ways and may overlap, but it is often possible to assign and quantify peaks that belong to a single metabolite or group of metabolites. Usually, the signal from a metabolite is quantified by calculating the area of a peak that can be uniquely assigned to it, although peak heights are also sometimes used.

To quantify the concentrations of the metabolites, it is common practice to add a known amount of a certain compound to each sample. When the compound is dissolved in the sample itself, it is called an internal standard. In contrast, an external standard would be either a solution of the compound in a capillary or the compound run as a separate sample but with the same experimental setup. The added calibration compound then gives rise to a control peak in the spectrum of the sample. Compounds commonly used as (internal) calibration standards in proton NMR are trimethylsilyl propionate (TSP) and tetramethylsilane.

The following formula is used to calculate the concentration of the metabolites using the calibration standard. Let $C_i$ be the concentration of metabolite $i$, and $C_0$ the (known) concentration of the control compound. Set

$$C_i = C_0 \times \frac{N_0}{N_i} \times \frac{A_i}{A_0}, i = 1, 2, ..., I \quad (1)$$

where $N_0$ is the (known) number of protons in the control molecule, $N_i$ is the (known) number of protons in metabolite $i$, $A_i$ is the measured peak area of metabolite $i$ and $A_0$ is the peak area of the standard. Besides helping with quantification, a calibration standard is useful for chemical shift calibration, which helps in identifying the metabolites by using the chemical shift values in the literature.

However, the use of the standard can be troublesome in some scenarios, as we show in Appendix A in the Supplementary Information. The fact that the standardization is done as a ratio will induce apparent correlations between metabolites, even if none exists in the first place.

*To whom correspondence should be addressed.

Metabolites in the cell participate in enzyme-catalyzed reactions that form complex biochemical networks. There are often interactions between metabolites (affecting their concentrations) when they appear in the same biochemical sub-network, or pathway. If, for instance, there is a rise in the cellular concentration of a metabolite that is the substrate of an enzyme, then the catalyzed reaction will usually proceed more rapidly so that the product of the reaction will also tend to increase in concentration. Thus, the concentrations of substrate and product would tend to be positively correlated. In the same way, negative correlations may occur when an increase in concentration of one metabolite leads to the depletion of the second metabolite. An example might be an enzyme that is subject to allosteric inhibition: a rise in the cellular concentration of the inhibitor metabolite would then tend to reduce the cellular concentration of the enzyme's product. Thus, the inhibitor and the product would be negatively correlated.

Analysis of the positive and negative correlations between metabolites can be performed by preparing a large number (typically 30–50) of apparently identical samples of, for instance, cultured cells. Although the measured concentrations of cellular metabolites in the individual samples will be identical within biological variation, that uniformity is achieved by numerous homeostatic mechanisms that will give rise to positive and negative correlations between metabolite concentrations. This type of analysis is often referred to as metabolite–metabolite correlation analysis (MMCA) (Fiehn and Weckwerth, 2003; Kose *et al.*, 2001; Steuer, 2006). The terms MMCA, correlation analysis and estimation of correlation maps will be used interchangeably in this article.

Commonly, Pearson correlations are used to estimate the interactions between metabolites. The use of mutual information has also been suggested as a dependency measure, as it captures more than linear relationships between metabolites (Numata *et al.*, 2008). However, as metabolomics data generally are quite skewed, even after transformation with a standard as in Equation 1, the Pearson coefficient may not be a good estimate of the correlations (see further discussion below). A transformation of the data may therefore be needed before further analysis. Estimating interactions with mutual information can also benefit from such transformations (Kraskov *et al.*, 2004).

The need for normalization for metabolomics data is crucial, just as with other types of omics data. NMR samples are affected by technical artifacts and may exhibit inflated between-sample variation owing to batch effects. For correlation analysis, these batch effects, together with effects from standardization, will result in large positive correlations, as illustrated in Figure 1.

A few normalization methods for metabolomics data have been suggested in the literature. The NOMIS method (Sysi-Aho *et al.*, 2007) is based on the presence of multiple internal standards in each sample. The optimal combination of standards is selected and used to remove systematic error. Also dependent on multiple internal standards is the CCMN method (Redestig *et al.*, 2009), which is mainly aimed at mass spectrometry-based metabolomics data. A recent addition to the field is the RUV-2 method, which is shown to be powerful for removal of unwanted variability while not being dependent on internal standards (De Livera *et al.*, 2012). However, RUV-2 is not a global normalization method in the sense that it does not produce a complete normalized dataset, but rather a compressed set suitable for detecting differentially abundant metabolites. Hence, RUV-2 is not recommended in connection to classification, clustering problems or MMCA.

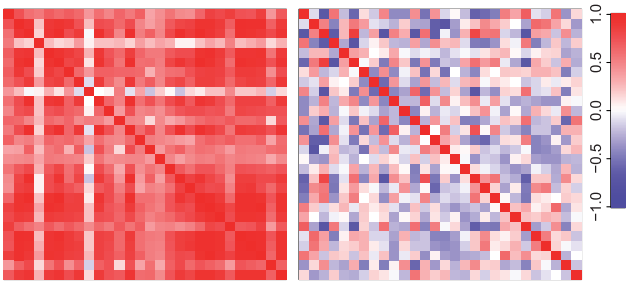


**Fig. 1.** Correlation analysis on a real NMR dataset. The metabolite–metabolite correlations in the left panel are deduced using non-normalized data, whereas the right panel illustrates the correlation estimates when using normalized data. The dataset used is the standardized log-transformed HDF set (see Section 2) and the normalization method is our proposed mixed model (see Section 3). Technical artifacts like batch effects and the use of a calibration standard lead to large positive correlations in the non-normalized data

In situations when these metabolomics normalization methods are not appropriate, normalization methods adapted from the single-channel microarray literature are commonly used. However, the nature of some common single-channel methods, e.g. quantile normalization (Boes and Neuhäuser, 2005; Bolstad *et al.*, 2003), renders them inappropriate to use on the smaller datasets that are usually produced in metabolomics. In quantile normalization, the idea is to give each sample the same distribution over features (e.g. metabolites). This is achieved by sorting the values in each sample, calculating a mean quantile over the samples and substituting the value of the data item in the original dataset with the mean (followed by a re-sort of each sample). This can be problematic for features in the tails of the distribution, as it is possible that a feature could receive the same value across all samples. The result of this is that, as happens with our dataset described in Section 2, some correlations between metabolites cannot be calculated. The median centering normalization also proves to be relatively non-robust if the number of metabolites is small (see discussion below).

In this article, we investigate how the use of a calibration standard for quantitation affects the reliable estimation of correlation maps. We argue that log transformations are reasonable for metabolomics data, and we suggest a global normalization method, suitable for smaller (targeted) metabolomics datasets, that is robust in connection to MMCA. The purpose of this normalization is to remove variation due to sources other than homeostatic changes. The proposed global normalization method is intended to be a complement to existing methods dependent on multiple internal standards and methods that only function in differential expression settings.

## 2 APPROACH

We use two versions of a real NMR dataset for our investigations and to illustrate our methods. The dataset stems from six cohorts of IMR90 human diploid fibroblasts (HDFs). For each

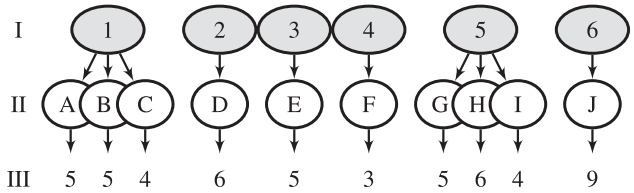


**Fig. 2.** Design for the real NMR dataset. The samples of HDFs originate from six cohorts (level I), which are divided into 10 batches (level II, A-J). From each batch, a number of different samples are cultured (level III). The total number of samples is 52

cohort, a number of batches were extracted, which in turn gave rise to a number of samples (52 in total). The design is illustrated in Figure 2. Cells from each sample were harvested (and further processed), and 10 $\mu$L of 10 mM TSP was added as an internal standard. Proton NMR spectroscopy data were acquired on a 600 MHz Bruker Avance NMR spectrometer by using a water pre-saturation pulse sequence. Preprocessing of the time-domain data included exponential multiplication (line broadening 0.3 Hz), Fourier transformation and zero and first order phase correction as well as $\log_2$ transformation.

In one version of this dataset, which we refer to as *standardized* (std), TSP was used for chemical shift calibration and metabolite quantitation. In total, 28 metabolites were uniquely characterized. In the other version of the dataset, called *raw*, TSP was not used for quantitation. In this set, 26 spectral features could be identified, of which some features are sums of signals for several metabolites. The two datasets are not directly comparable, as, in addition to the issues involving standardization, the same metabolite might not correspond to the same peaks in both sets.

We also use simulated datasets to compare the performance of different normalization techniques. The structure of the variation modeled for the simulated datasets is similar to that in the real dataset as we include artifacts like cohort, batch and sample effects.

## 3 MATERIALS AND METHODS

### 3.1 Transformation of data

In Supplementary Figure S1 (Appendix B in the Supplementary Information), we show the distribution of the samples in the standardized HDF dataset without any normalization or transformation. The data exhibit skewness, with many large outliers. In such situations, the Spearman rank correlation coefficient may give a better estimate of the interactions than the Pearson coefficient. However, not only MMCA but also different types of normalization are performed on the data. These methods are not always suited for highly skewed data, so instead we argue that the data should be transformed to achieve more balanced distributions for each sample. One such transformation is the log transformation, which we use throughout this article.

### 3.2 Raw versus standardized peak areas

When normalizing the type of NMR data we described in the previous sections, we argue that log-transformed data should be used. Apart from the choice of using a transformation or not, we can also choose between normalizing standardized intensities ($C_i$ or $\log(C_i)$) or the raw peak areas ($A_i$ or $\log(A_i)$). As indicated above, a TSP-normalized dataset has better resolution of individual metabolites, but the additional scaling can cause artificial correlations. Normalizing the raw peak areas has a natural appeal, as we remain close to the original data. If a correlation map is the end goal, then we can use the peak areas directly, because the scaling with proton number will cancel out. Let $A_i/N_i$ be the peak area for metabolite $i$ scaled by proton number, then we have the following equation:

$$\begin{aligned}\mathrm{Corr}(A_i/N_i,\ A_j/N_j) &= \frac{\mathrm{Cov}(A_i/N_i,\ A_j/N_j)}{\mathrm{D}(A_i/N_i)\mathrm{D}(A_j/N_j)} \\ &= \frac{\mathrm{Cov}(A_i/A_j)/\ (N_i \cdot N_j)}{\mathrm{D}(A_i)/N_i \cdot \mathrm{D}(A_j)/N_j} = \mathrm{Corr}(A_i, A_j)\end{aligned} \tag{2}$$

and, as $\log(A_i/N_i) = \log(A_i) - \log(N_i)$, we get

$$\mathrm{Corr}(\log(A_i/N_i),\ \log(A_j/N_j)) = \mathrm{Corr}(\log(A_i),\ \log(A_j)) \tag{3}$$

where $\mathrm{D}(\cdot)$ denotes standard deviation.

### 3.3 ANOVA approach to normalization

Fixed effects models to normalize microarray data have been suggested in several papers (Kerr and Churchill, 2001a, b). We adapt the microarray approach as a comparison to correct for effects from the different cohorts, batches and samples. This ANOVA (Analysis of variance) model can be adapted for either raw peak areas or standardized areas. We have $i = 1, \ldots, I$ metabolites, $j = 1, \ldots, J$ cohorts, $k = 1, \ldots, K$ batches and $l = 1, \ldots, L$ samples. Let $Y_{ijkl}$ denote the log-transformed peak area/concentration for metabolite $i$ in cohort $j$, batch $k$ and sample $l$ as follows:

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \varepsilon_{ijkl} \tag{4}$$

where $\alpha_i$ is the metabolite effect, $\beta_j$ a cohort effect, $\gamma_k$ a batch effect, $\delta_l$ a sample effect and $\varepsilon_{ijkl} \sim N(0, \sigma^2)$ is a random error term. When no cohorts are defined, the model reduces to having only a metabolite, batch and sample effect.

### 3.4 Mixed model approach to normalization

Owing to the structure of the experimental data, the cohorts can be regarded as coming from a larger pool of HDF cohorts, a reasoning that also applies to the batches and samples. A mixed effects model is therefore a more appealing alternative to model the variation than a fixed effects model. Again, either log-transformed raw peak areas or standardized metabolite levels can be used as input. A suitable mixed model can be formulated as follows:

$$Y_{ijkl} = \mu + \alpha_i + b_j + b_{jk} + b_{jkl} + \varepsilon_{ijkl} \tag{5}$$

where $b_j \sim N(0, \sigma_1^2)$ is a cohort effect, $b_{jk} \sim N(0, \sigma_2^2)$ is a batch within cohort effect and $b_{jkl} \sim N(0, \sigma_3^2)$ represents a sample within batch within cohort effect.

As we are interested in estimating the correlation matrix for the metabolites, we can estimate this at the same time as the fixed and random effects by adopting an iterative procedure.

To achieve this, we reformulate the model in Equation 5 into a matrix form:

$$Y_l = X_l\beta + Z_l u + \varepsilon_l, \tag{6}$$

where $Y_l$ is the signal for sample $l$, which is a vector of length $I$ (the number of metabolites). The vector $\beta$ contains the fixed effects and $u$ is a vector of the random effects. The matrices $X_l$ and $Z_l$ are the sample-specific regressor matrices. The random error term $\varepsilon_l$ is a vector with a $N(0, \Sigma)$ distribution where $\Sigma$ is the $I \times I$ metabolite covariance matrix. The random effects $u$ have variance-covariance matrix

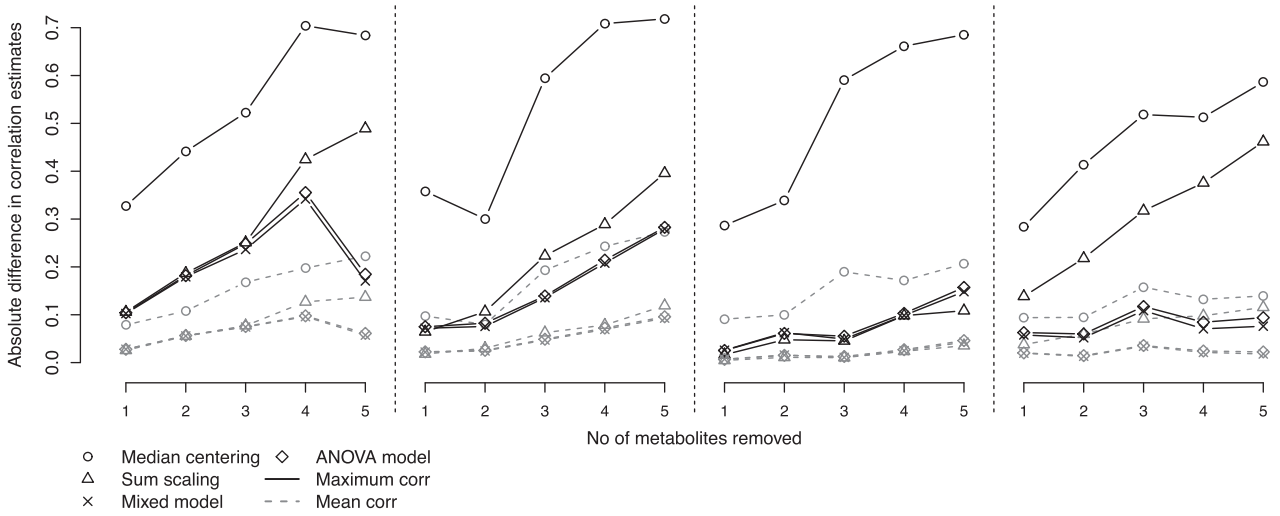


**Fig. 3.** Robustness of normalization methods with regard to correlations. Changes to correlation estimates when removing one, two, and up to five of the metabolites with the highest variance (left panel), the lowest variance (mid-left panel), median closest to the overall dataset median (mid-right panel) and finally the metabolites with the highest mean (right panel). The correlation estimates between the retained metabolites are compared when using the full dataset and the reduced sets. The maximum (solid line, black) and mean (dashed line, grey) of the absolute difference in correlation estimates are shown. The dataset used is the standardized log-transformed HDF set

$G = \text{diag}(\sigma_1^2, \ldots, \sigma_1^2, \sigma_2^2, \ldots, \sigma_2^2, \sigma_3^2, \ldots, \sigma_3^2)$. The iterative procedure is as follows [cf. Meng and Rubin (1993)]:

0. Initialize $\Sigma$, e.g. with a diagonal unit matrix.
1. Update the fitted values. (Fitted values are obtained by adding the fitted values from the fixed effects and the estimated contributions of the random effects on the highest level of grouping.) Calculate the matrix square root of the inverse of $\Sigma$: $B = \Sigma^{-1/2}$ (by using, e.g. diagonalization). Transform $Y_l, X_l, Z_l$ so that $Y'_l = BY_l$, $Z'_l = BZ_l$ and $X'_l = BX_l$. Fit the model in Equation 6 with the transformed matrices for each sample $l$. Calculate the fitted values $FV'_l$ and residuals $RV'_l$. Inverse transform the residuals and fitted values by $FV_l = B^{-1}FV'_l$ and $RV_l = B^{-1}RV'_l$.
2. Update $\Sigma$.
   Estimate $\Sigma$ by using $\frac{1}{L}\sum_{l=1}^{L} RV_l RV_l^T$.
3. Go to 1 and repeat the procedure until convergence.

When no cohorts are defined, the model will include a fixed metabolite effect, a random batch effect and a random sample within batch effect.

We fit this model with the open source software R (R Core Team, 2012) using the `hglm` package (Ronnegard *et al.*, 2010).

### 3.5 Simulation of data

Samples were simulated from a normal distribution with a metabolite-specific mean and under a fixed covariance matrix (of varying size) to mimic log-transformed real data. Levels for a control metabolite, with a small interaction to the other metabolites and with half the variance of the other metabolites, were also generated. Random effects for cohorts, batches and samples were added to the log-transformed data. To mimic the real data, scaling with the control metabolite was done on the non-log scale.

### 3.6 Comparisons with other methods of normalization

We mainly compare (in addition to the ANOVA approach) our proposed method with three global single-channel normalization methods that are not dependent on the presence of internal standards and are implemented in readily available R packages: first, a loess smoother, implemented as `normalize.loess` in the `affy` package (Gautier *et al.*, 2004); second, a median centering method where the median of each sample is subtracted. This normalization will re-center the values, but this has no effect when correlation estimates are concerned; and finally, we use a sum scaling method, which is based on scaling each sample with the total signal of that sample.

Quantile normalization (QN) (Bolstad *et al.*, 2003) and variance stabilizing normalization (VSN) (Huber *et al.*, 2002) are not applicable to the HDF dataset, but are used as a comparison for simulated datasets of larger dimensions. The implementations used are the `normalizeQuantiles` and `normalizeVSN` methods in the R package `limma` (Smyth, 2005). The `limma` package calls the `vsn` package to perform the normalization for VSN. All normalization methods are applied on log-transformed real data and on simulated data without transformation.

## 4 RESULTS

### 4.1 Robustness of different normalization methods

To compare the robustness of the different normalization methods when estimating correlations, we used the HDF dataset (both raw and standardized versions). Ideally, correlations between a subset of the metabolites should remain the same if we remove one or several other metabolites from the set and redo the normalization and estimation. However, if the number of features is relatively small, the normalization methods will be highly dependent on the levels of some metabolites.

In Figure 3, results from such an experiment are presented. We removed from one up to five of the metabolites with (i) highest variance (left panel), (ii) the lowest variance (mid-left panel), (iii) median closest to the overall dataset median (mid-right panel) and finally (iv) the metabolites with the highest mean (right panel). The dataset used is the standardized HDF set. Similar plots for the raw HDF set (Supplementary Fig. S1) and for both the raw and

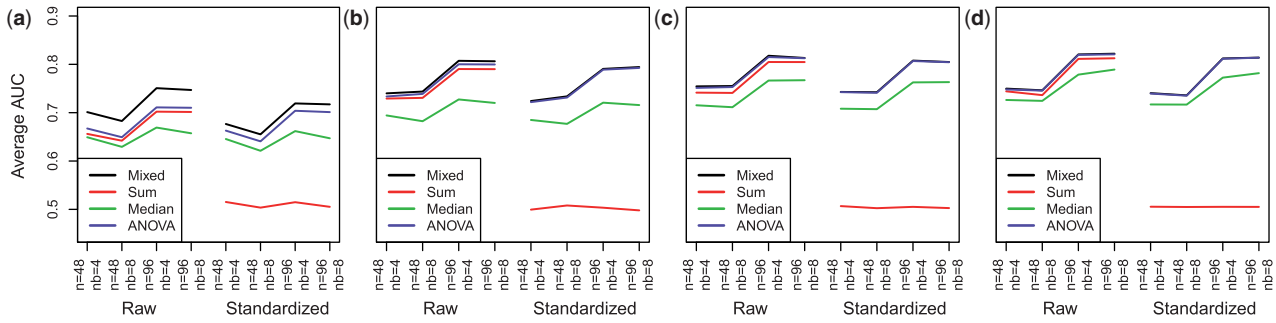


**Fig. 4.** AUC values for discovery of true correlations for different scenarios when varying the number of metabolites, batches and samples. The number of metabolites is 12 in panel (a), 24 in panel (b), 36 in panel (c) and 48 in panel (d). The number of samples was alternated between 48 and 96, whereas the number of batches was either 4 or 8. The AUC values given are the averages over 50 simulated datasets

standardized sets when random metabolites are removed (Supplementary Figs S2–S4) are given in Appendix C in the Supplementary Information. For each new dataset, we normalized using the ANOVA model, our proposed mixed model, sum scaling and median centering. A loess smoother was also included when applicable. We then compared how large the differences were in the estimated correlations for the retained metabolites in each dataset. We calculated the maximum and mean absolute differences in correlation estimates.

The median centering performs poorly in most scenarios (although the effects are not as dramatic when using the raw HDF set), whereas the sum scaling suffers most when removing high intensity metabolites. The ANOVA model and mixed model perform similarly, with a slight edge to the mixed model.

### 4.2 Reconstruction of correlation maps

A simulation study was performed to assess the different normalization methods concerning discovery of true correlations, as well as comparing raw and standardized data. A number of scenarios were simulated as described in Section 3. The number of metabolites was varied between 12 and 48, whereas the number of batches and samples was evaluated using two levels each, with the number of cohorts fixed to two throughout the simulations. Two different basic correlation matrices in a block-diagonal form were used to define the dependencies between metabolites in the different scenarios (details in Appendix D of the Supplementary Information).

One set of results is given in Figure 4. Using 50 replicates of each scenario, we evaluated the correlation estimates by deducing estimates of the true positive rate (TPR) and false positive rate (FPR) for a subset of the normalization methods in the robustness study. The TPR and FPR form the basis (using a range of cut offs for correlation) for calculating the area under the receiver-operating characteristic curve (AUC).

Based on both raw and standardized data, the mixed and ANOVA models outperform the median normalization method in all scenarios [difference in AUC, paired $t$-test with all $P$-values $< 0.001$]. The sum scaling performs well on the raw data, but deteriorates on the standardized dataset. The mixed model and ANOVA normalization have a comparable performance, although the mixed model outperforms ANOVA on the datasets with smaller number of metabolites ($p < 0.005$). The mixed and ANOVA models perform marginally better on the raw data than on the standardized counterpart. Results for the second set of correlation matrices are presented in Supplementary Figure S4 (Appendix D of the Supplementary Information).

Another way to evaluate the normalization methods is to cluster the metabolites using the normalized data. The blocks defined in the correlation matrix, as described briefly above, form natural clusters of metabolites. To compare the clustering from each method with the true groupings of metabolites, we use the adjusted Rand index [a measure in the range (0, 1), where 1 indicates perfect match]. As the division into groups is based on correlations, not differing metabolite levels, a distance measure for the clustering must be correlation based. By using $1 - |c_{ij}|$, where $c_{ij}$ is the correlation between two metabolites $i$ and $j$, we see that a clustering is simply a different approach to assess the correlation matrix estimates. Supplementary Tables S1 and S2 in Appendix D in the Supplementary Information show the average Rand index over 50 replicated simulations for the same data as presented in Figure 4 and Supplementary Figure S5. The mixed model edges the other approaches in most of the scenarios using this evaluation technique.

### 4.3 Raw versus standardized data

To further investigate the comparability of raw and standardized NMR data, we identified five metabolites in our HDF dataset with signals originating from exactly the same peaks in both the raw and standardized log-transformed versions (i.e. chemical shift calibration has no effect). Both versions of the data were normalized (using the complete sets of features; 26 in raw and 28 in standardized) with median centering, sum scaling and our mixed model. The 10 metabolite–metabolite correlations were then compared for the different versions of the data (raw or standardized) for the three normalization methods, including no normalization at all, as depicted in Figure 5.

For non-normalized data, the correlations are all close to zero or positive, which most likely is caused by artifacts like batch effects that induce correlations. It should also be noted that a subset of the metabolites exhibit strong positive correlations in the standardized data, with more modest positive correlations in the raw version, which indicates that the calibration standard affects the signal. However, for the normalized data the correlation estimates are more comparable (close to the diagonal line),

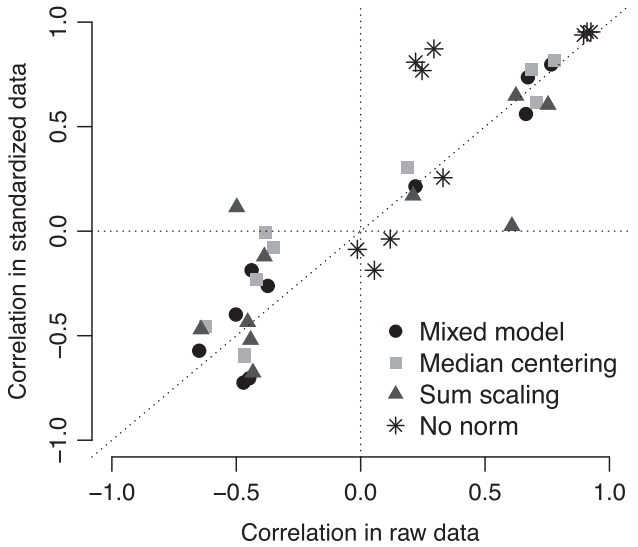


**Fig. 5.** Raw NMR data compared with the standardized counterpart. The inter-metabolite correlations for five metabolites (alanine, choline, glutamate, glycine and valine) are compared in raw and standardized data for three normalization methods and non-normalized data in the HDF dataset

although the sum scaling produces some outliers. The mixed model results in the most comparable correlations.

### 4.4 Low- to high-dimensional data

The simulation studies presented in the previous section indicate that model-based approaches outperform approaches based on median centering and sum scaling when normalizing datasets in the size range 10–50 features. When the number of features increases further, single-channel methods adapted from the microarray literature can be used. We compared how our mixed model performed in comparison with QN and VSN in identifying correlations for increasing numbers of metabolites (20, 40, 80, 160, 320 and 640 features). The number of samples was fixed and the correlation matrices estimated using a shrinkage approach (Schäfer and Strimmer, 2005). More details are given in Appendix E in the Supplementary Information.

Table 1 contains the average AUC values for each method over 50 replicates for each dataset size. The quantile normalization clearly has problems for small numbers of features, while the variance stabilizing normalization, by following the recommendations for usage, is not recommended for the simulations with the two smallest sizes. VSN eventually catches up to the mixed model in the normalization, whereas interestingly enough, QN does not, although the differences in AUC are not big. The fact that QN has a worse performance than VSN in the setting of correlation estimation between features can be because of the properties reported by Lim *et al.* (2007). Although QN works well in the differential expression settings, correlation artifacts in the data are also introduced. When the number of features exceeds 100–150, these results indicate that variance stabilizing normalization can be used with similar performance to a model-based approach, while quantile normalization would be less preferable to VSN.

**Table 1.** AUC values for estimating correlations with varying number of metabolites (20–640)

| Dataset size | Mixed | QN | VSN |
|---|---|---|---|
| $m = 20$ | 0.769 (0.04) | 0.662 (0.05) | NA |
| $m = 40$ | 0.795 (0.03) | 0.737 (0.03) | NA |
| $m = 80$ | 0.793 (0.02) | 0.761 (0.03) | 0.788 (0.02) |
| $m = 160$ | 0.795 (0.02) | 0.781 (0.01) | 0.792 (0.02) |
| $m = 320$ | 0.796 (0.01) | 0.790 (0.01) | 0.795 (0.01) |
| $m = 640$ | 0.797 (0.01) | 0.793 (0.01) | 0.797 (0.01) |

*Note*: The columns represent the different methods used; mixed model, quantile normalization (QN) and variance stabilizing normalization (VSN). The average AUC value of 50 repeats of each scenario is given, with the standard deviation in parenthesis. A graphical representation of the results is provided in Appendix E (Supplementary Fig. S6).

## 5 DISCUSSION

The purpose of this article is two-fold; first we propose a global normalization method intended for use with smaller metabolomics datasets, and second, we investigate how well correlations can be discovered when we have the option of using NMR data calibrated using a standard, or not.

When comparing raw and standardized metabolite levels (peak areas/concentration) for the real HDF dataset, we observed large positive correlations in the non-normalized data. Such large correlations are not biologically reasonable, as we expect both negative and positive correlations to be present, and we argue that they are caused by batch effects and the calibration standard (Fig. 1). The positive correlations were more pronounced for the standardized data, which is a cause for concern. However, after normalization, both the raw and standardized data exhibited better correspondence, and large differences in estimated correlations could not be found. Most likely, the influence of the calibration standard is (at least partly) removed by the inclusion of a random sample effect in our normalization model.

Our simulation study on the discovery of true metabolite-metabolite correlations included cohort, batch and sample effects in the simulated data, as well as a simulated calibration compound. Although true NMR data are more complex, and the calibration standard affects the data in a more intricate way than by simple scaling, we think the study gives some pointers concerning performance of different normalization methods, and how the power is affected when we use either raw or standardized data. In summary, our simulations show that using raw data compared with standardized data has a slight edge when it comes to performance (measured as average AUC). However, the loss in power using standardized data compared with raw data is small and the added advantages of better metabolite quantitation and chemical shift calibration render standardized data a feasible choice when estimating correlations.

In our correlation-robustness study (based on real data), we investigate how stable the correlations are when removing one or several metabolites from the set, and re-normalizing without the removed metabolites. The study shows that methods developed for high-dimensional datasets, like the loess smoother, perform poorly for smaller datasets, if at all applicable. The commonly

used median normalization is also problematic when removing certain sets of metabolites, and the same trend can be seen for the sum scaling method. The mixed model we propose is a robust choice.

In this article, we focus on low-dimensional metabolomics datasets for two main reasons. First, the normalization methods adopted from the microarray community (e.g. quantile normalization) are usually not applicable to these sets, and as current methods for metabolomics data have restrictions (either demanding presence of multiple standards or intended for differential analysis), customized methods are needed. Second, in MMCA, the focus is usually to unravel metabolite–metabolite correlations for a smaller set of metabolites, and *targeted* datasets are mainly used for this. When evaluating our mixed model for larger datasets (varying size between 20–640 features), contrasting it with quantile and variance stabilizing normalization (VSN), we conclude that a model-based approach is to prefer for datasets with less than ~100 features. For larger sets, the variance stabilizing approach performs equally well as a mixed model and better than quantile normalization. As mixed models are hard to adapt for large-scale sets, VSN is a good alternative.

The real data we use in this article have a complex structure with cohorts, batches and samples. The normalization method based on the mixed model can also be applied to simpler designs, e.g. when the nesting consists of samples within batches, making it applicable to many scenarios.

The estimation of the fixed and random effects in the mixed model is coupled to an iterative procedure to estimate the correlation matrix of the metabolites simultaneously. However, it is also possible to apply the method for just one iteration, and ignore the estimation of the correlation matrix in the first step, as we do in one of the simulation studies. This is a feasible option when the purpose is only to normalize the data, in e.g. a $p > n$ (more metabolites than samples) setting.

## 6 CONCLUSION

We present a mixed model approach to normalization of low-dimensional metabolomics datasets. We show that this method performs well compared with competing methods with respect to robustness and discovery of true correlations. We also use real and simulated data to infer how standardization with a calibration compound affects the estimation of correlations. Although the performance for non-standardized data is slightly better than for standardized data, the benefits of chemical shift calibration in identifying metabolites as well as in quantitation of metabolite concentrations motivate the use of a calibration standard in NMR experiments.

## ACKNOWLEDGEMENTS

The authors wish to thank Olle Nerman and Terry Speed for helpful comments on the material in this article. Two anonymous reviewers also provided valuable input, which greatly improved the article.

*Funding:* Cancer Research UK (in part) [Programme Grant C14303/A10825]; the core facilities of the Cancer Research UK Cambridge Institute; the Erik and Edith Fernström foundation.

*Conflict of Interest*: none declared.

## REFERENCES

Boes,T. and Neuhäuser,M. (2005) Normalization for Affymetrix GeneChips. *Methods Inf. Med.*, **44**, 414–417.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Chung,Y.L. *et al.* (2003) Metabolic profiling in tumors by in vivo and in vitro NMR spectroscopy. In: Harrigan,G.G. and Goodacre,R. (eds) *Metabolic Profiling - Its Role in Biomarker Discovery and Gene Function Analysis*. Chapter 5, Kluwer Academic Publishers, Dordrecht, the Netherlands, pp. 83–94.

De Livera,A.M. *et al.* (2012) Normalizing and integrating metabolomics data. *Anal. Chem.*, **84**, 10768–10776.

Fiehn,O. and Weckwerth,W. (2003) Deciphering metabolic networks. *Eur. J. Biochem.*, **270**, 579–588.

Gautier,L. *et al.* (2004) affy–analysis of Affymetrix Genechip data at the probe level. *Bioinformatics*, **20**, 307–315.

Huber,W. *et al.* (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18 (Suppl. 1)**, S96–S104.

Kerr,M.K. and Churchill,G.A. (2001a) Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183–201.

Kerr,M.K. and Churchill,G.A. (2001b) Statistical design and the analysis of gene expression microarray data. *Genet. Res.*, **77**, 123–128.

Kose,F. *et al.* (2001) Visualizing plant metabolomic correlation networks using clique-metabolite matrices. *Bioinformatics*, **17**, 1198–1208.

Kraskov,A. *et al.* (2004) Estimating mutual information. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **69** (Pt 2), 066138.

Lim,W.K. *et al.* (2007) Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*, **23**, i282–i288.

Meng,X.L. and Rubin,D.B. (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267–278.

Numata,J. *et al.* (2008) Measuring correlations in metabolomic networks with mutual information. *Genome Inform.*, **20**, 112–122.

R Core Team. (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org/.

Redestig,H. *et al.* (2009) Compensation for systematic cross-contribution improves normalization of mass spectrometry based metabolomics data. *Anal. Chem.*, **81**, 7974–7980.

Ronnegard,L. *et al.* (2010) hglm: a package for fitting hierarchical generalized linear models. *R J.*, **2**, 20–28.

Schäfer,J. and Strimmer,K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article32.

Smyth,G.K. (2005) Limma: linear models for microarray data. In: Gentleman,R. *et al.* (ed.) *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, New York, pp. 397–420.

Steuer,R. (2006) Review: on the analysis and interpretation of correlations in metabolomic data. *Brief. Bioinform.*, **7**, 151–158.

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