Validation of alternative methods of data normalization in gene co-expression studies

Antonio Reverter*, Wes Barris, Sean McWilliam, Keren A. Byrne, Yong H. Wang, Siok H. Tan, Nick Hudson and Brian P. Dalrymple

Bioinformatics Group, CSIRO Livestock Industries, Queensland Bioscience Precinct, St Lucia, QLD 4067, Australia

Received on October 17, 2004; revised on November 2, 2004; accepted on November 2, 2004

Advance Access publication November 25, 2004

1 INTRODUCTION

Microarray technology provides us with the opportunity to make inferences on protein associations and gene regulatory networks on a genome-wide scale. Genes of similar function yield similar expression patterns across a diverse range of conditions (Eisen et al., 1998; Hughes et al., 2000; Kim et al., 2003; Segal et al., 2003). Based on patterns of co-expression, often from merging several data sources, transcriptional regulatory networks have been proposed mostly within (Brown et al., 2000; Guelzim et al., 2002; Horak et al., 2002; Lee et al., 2002; Gardner et al., 2003; Yu et al., 2003; Nachman et al., 2004), but also across species (Stuart et al., 2003; Winter et al., 2004). Given an expression vector for each gene across various experimental conditions, the simplest kernel that can be used to measure the similarity between pairs of genes is the Pearson correlation coefficient. This correlation is equivalent to the dot product when the vectors have been normalized to zero mean and standard deviation (SD) of 1, so that each has a Euclidean length 1 (achieved by dividing each element in the vector by the square root of its norm). It gives a measure of the strength of the linear relationship between two vectors, it pays attention only to differences in pattern ignoring differences in level and amplitude, and (most importantly) assumes that each pair of values is i.i.d., an independent realization from the same (not necessarily normal) distribution. Violations of this assumption can generate spurious correlations that in turn will increase the false discovery rate (FDR) when making inference on gene regulatory networks.

Microarray data are notorious for being noisy due to systematic biases. Accounting for systematic effects during normalization, and methods to adjust for such biases is a subject of great importance (Smyth et al., 2003; and references therein), for which alternative approaches are continuously being proposed (Baird et al., 2004; Benito et al., 2004; Fan et al., 2004). Even when these effects have been accounted, it might be of interest to standardize each vector to unit variance (dividing by its SD) to avoid uninterpretable correlations resulting from either one or both vectors having a negligible variation. However, while dividing by the SD is necessary for many techniques, it amounts to giving equal prior importance to all genes (Frank and Friedman, 1993).

Most research involving co-expression of genes based on microarray data ignore alternate methods of data normalization because they are based on the quite reasonable, yet non-testable assumption, that gene pairs exhibiting co-expression in multiple species and/or tissues and across a large number of arrays are more likely to be functionally relevant (true positives). Yeung et al. (2004) concluded that the proportion of co-regulated genes is directly related to the number of hybridization experiments and not so much to the diversity of experiments. This could be attributed to the fact that noise can be better described with more technical replicates (i.e. power increases with more data), while diverse microarray experiments do not help in characterizing such noise. To standardize data, the authors used a z-score as their primary evaluation criterion. Such standardization is achieved by subtracting the average expression value of each gene across all experiments from the expression value of each gene and then divided by the SD of its expression level across all experiments. Assymptotically, and by means of the Central Limit Theorem (for details see Mood et al., 1974), such standardization is expected to yield a ‘normalized’ set of data that will follow a standard normal distribution (i.e. with zero mean and unit variance). For a bivariate normal distribution, the distribution of the correlation coefficients and density function under the null hypothesis is given by (Bevington

To whom correspondence should be addressed.

*To whom correspondence should be addressed.
and Robinson, 2003): 
\[ P(r) = \frac{1}{\sqrt{\pi}} \frac{\Gamma((v + 1)/2)}{\Gamma(v/2)} (1 - r^2)^{(v-2)/2}, \]

where \( \Gamma(\cdot) \) is the gamma function and \( v = N - 2 \) is the degrees of freedom.

We compare nine normalization methods to pre-adjust the data prior to computing all pair-wise clone × clone, clone × gene and gene × gene correlations. We tested the performance of each method by: (1) comparing the resulting empirical distributions of correlation coefficients with theoretical expectations in \( P(r) \); and (2) using validation on those clones for which a reliable functional annotation to a gene existed. Finally, once the normalization method of choice is identified, we show how the visualization of the resulting correlation matrices after sorting genes according to simple criteria can reveal structures that are consistent with functional biology.

## 2 Microarray data sets

We used five microarray studies in genetic improvement of beef cattle from our laboratory spanning a total of 78 microarray hybridizations, with mRNA from muscle and adipose tissues from 23 experimental conditions. These experiments were performed using the same bovine muscle and fat cDNA slide with 9600 elements (from now on refer to as clones) spotted twice (Lehnert et al., 2005). Spots were arranged in a layout of 48 printing blocks each of 20 × 20 spots. The 48 blocks were arranged in 12 rows of four blocks each. Importantly, duplicate clones on the array were spotted adjacent within a single block.

Several studies involving statistical and biological aspects for some of these experiments have been reported (Byrne et al., 2005; Reverer et al., 2003, 2004, 2005). For the present study, we did not perform any data pre-processing with any filtering criteria based on differential expression or absolute level of expression. We edited out readings with foreground signal ≤ background signal, and clones not observed in the 23 experimental conditions. This criteria resulted in a total of 2,095,999 fluorescent signals from 8074 clones of which 1530 contained accurate \( P < 0.01/8074 = 1.24 \times 10^{-6} \) functional BLAST annotation for 624 genes determined by searching the NCBI human reference sequence (RefSeq) collection of mRNA sequences. The method of Reverer et al. (2005) to infer transcriptome coverage estimates that, assuming a genome with 30,000 genes, a library with 9600 clones would contain 2533 unique genes while a library with 1530 (fully informative) clones would contain 716 genes. The redundancy of the cDNA libraries used to develop the microarray slide is reflected by the fact that a total of 195 genes were represented by more than one clone. These genes will provide the basis for the validation study to test the performance of each normalization method. Table 1 provides details of the five experiments including number of hybridizations, experimental conditions and summary statistics on signals.

## 3 Normalization methods: definition

We explore nine normalization methods based on raw means and on solutions to uni- and multi-variate mixed-model equations and after ignoring or accounting for between and within gene and experiment variation. These methods are briefly described in Table 2 and in more detail below.

### 3.1 Methods based on raw means

Normalization methods (1) RMNA, (2) RMCE and (3) RMEC (acronyms for Raw Means Non-Adjusted, Raw Means adjusted between Clones within Experiment and Raw Means adjusted between Experiments within Clones, respectively) were computed for each clone in \( e (e = 1, \ldots, 8074) \) as follows:

1. \( \text{RMNA}_c = \overline{x}_c \),
2. \( \text{RMCE}_c = \frac{k - \bar{p}_{c/e}}{\sigma_e} \); and
3. \( \text{RMEC}_c = \frac{k - \bar{p}_{c/E}}{\sigma_E} \),

where \( \overline{x}_c \) is the vector containing the average expression of the \( c \)-th clone in each of the 23 experimental treatments (or conditions) explored across the five microarray experiments (Table 1); \( \bar{p}_{c/E} \) is the vector containing the average expression of the \( c \)-th clone in the \( E \)-th experiment (thus, for each clone, there were 5 \( \bar{p}_{c/E} \) one for each experiment, and each value in \( \overline{x}_c \) was adjusted by the \( \bar{p}_{c/E} \) corresponding to its experiment of origin); \( \sigma_e \) is the SD of all the expressions in the \( E \)-th experiment; \( \sigma_E \) is the vector containing the average expression of the \( c \)-th clone in the \( E \)-th experiment; and \( \sigma_e \) is the SD of all the expressions in the \( c \)-th clone. It can be shown that, due to unit variance of RMEC, and assuming independence between \( \bar{p}_{c/E} \) and \( \overline{x}_c \), for two clones in \( i \) and \( j \) where \( i \neq j \), the correlation between RMNA and RMCE, and between RMCE and RMEC is equivalent to the correlation between RMEC and RMEC. Thus, and within rounding, results from RMNA are expected to be identical to those from RMEC.

### 3.2 Methods based on univariate mixed-model

Normalization methods (4) MM1NA, (5) MM1CE and (6) MM1EC (MM1 for univariate mixed-model and NA, CE and EC as defined earlier) were obtained after implementing the following univariate mixed-model:

\[ y = X\beta + Z_a a + Z_d d + Z_t t + \epsilon, \]

where \( y \) is the vector of all the intensity signals \( (N = 2,095,999) \), background corrected and base-2 log transformed; \( X \) is the incidence matrix relating signals in \( y \) with systematic fixed effects in \( \beta \) including array slide, printing block and fluorescent dye channel; \( Z_a \) is the incidence matrix relating \( y \) with random effects in \( a \) corresponding to the clones (\( N = 8074 \)) printed on the array; \( Z_d \) is the incidence matrix relating \( y \) with random effects in \( d \) corresponding to the interaction of clone by array and printing block; \( Z_t \) is the incidence matrix relating \( y \) with random effects in \( t \) corresponding to the interaction of clone by the two fluorescent dye channels; \( \epsilon \) is the random error associated with signals in \( y \).

We make standard stochastic assumptions about model (1). Accordingly, random effects are assumed to follow a normal distribution.
3.3 Methods based on multivariate mixed-models

Normalization methods (7) MM5NA, (8) MM5CE and (9) MM5EC (MM5 for penta-variate mixed-model and NA, CE, and EC as defined earlier) were obtained after implementing the following penta-variate (one for each experiment variation removed such skewness, stabilized the variance mostly to the inherent correlation existing in microarray data. Method RMCE, which attempted to correct for the between gene and within experiment variation removed such skewness, stabilized the variance but generated a uniform-like distribution that failed to approach the expected bell-shaped density. Hence, most correlations computed using the simplest models (i.e. raw-means methods) departed from the expectation. Methods RNA and RMEC based on raw means yielded similar distributions. These were vastly negatively skewed with small density on negative correlations while the bulk of the correlations were in the moderate to extreme positive. This skewness was attributed in part to the redundancy of the cDNA microarray platform, but mostly to the inherent correlation existing in microarray data. Method RMCE, which attempted for the between gene and within experiment variation removed such skewness, stabilized the variance but generated a uniform-like distribution that failed to approach the expected bell-shaped density. Hence, most correlations computed using the simplest models (i.e. raw-means methods) departed from the expectation. Methods RNA and RMEC based on raw means yielded similar distributions. These were vastly negatively skewed with small density on negative correlations while the bulk of the correlations were in the moderate to extreme positive. This skewness was attributed in part to the redundancy of the cDNA microarray platform, but mostly to the inherent correlation existing in microarray data. Method RMCE, which attempted for the between gene and within experiment variation removed such skewness, stabilized the variance but generated a uniform-like distribution that failed to approach the expected bell-shaped density. Hence, most correlations computed using the simplest models (i.e. raw-means methods) departed from the expectation. Methods RNA and RMEC based on raw means yielded similar distributions. These were vastly negatively skewed with small density on negative correlations while the bulk of the correlations were in the moderate to extreme positive. This skewness was attributed in part to the redundancy of the cDNA microarray platform, but mostly to the inherent correlation existing in microarray data. Method RMCE, which attempted for the between gene and within experiment variation removed such skewness, stabilized the variance but generated a uniform-like distribution that failed to approach the expected bell-shaped density. Hence, most correlations computed using the simplest models (i.e. raw-means methods) departed from the expectation.

In order to test the goodness-of-fit, we applied the Kolmogorov–Smirnov (K–S) test (for details see Law and Kelton, 1991, pp. 387–391). The K–S test statistic is simply the largest vertical
distance between the theoretical and the empirical distribution functions. The critical values for the K–S test statistic at \((1 - \alpha) = 0.95\) and 0.99 are 1.358 and 1.628, respectively. The computed values for the K–S test statistic for normalization methods 1–9 were 10.176, 4.185, 10.097, 1.434, 1.319, 1.421, 1.617, 1.189 and 1.636, respectively. Thus, at 5% significance level, MM1CE and MM5CE yielded correlations that are distributed according to expectations. At 1% significance level, this statement is only true for MM5CE.

4.2 Pair-wise comparisons

Table 3 compares the nine normalization methods in a pair-wise fashion by exploring the correlations and maximum discrepancies. The correlation of correlations resulting from methods based on raw means was small positive (\(\sim 0.20\)) between RMNA and RMCE, and between methods RMCE and RMEC, and nearly perfect (0.98) between RMNA and RMEC. Also, these raw mean-based methods produced results that were poorly correlated with those from methods based on mixed-models. Within the mixed-models normalization methods, results were highly correlated and above 0.90 in all cases. Between mixed-models methods, results were moderately correlated with the largest being observed at 0.69 between MM1CE and MM5CE. The maximum discrepancy between raw mean-based methods and mixed-model methods approached the upper bound of 2.0 in all comparisons. Smaller discrepancies, \(\sim 0.6\) within and \(\sim 1.2\) across, were observed for results from mixed-model methods.

Due to the strong parallelisms observed between methods MM1CE and MM5CE, the origin of the top 20 discrepancies was further scrutinized. For all the genes in the disputed

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMNA</td>
<td>0.53</td>
<td>0.20</td>
<td>0.98</td>
<td>0.11</td>
<td>0.12</td>
<td>0.11</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>RMCE</td>
<td>1.84</td>
<td>0.15</td>
<td>0.21</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
<td>0.09</td>
<td>0.10</td>
</tr>
<tr>
<td>RMEC</td>
<td>0.76</td>
<td>1.79</td>
<td>0.51</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>MM1NA</td>
<td>1.78</td>
<td>1.72</td>
<td>1.74</td>
<td>0.02</td>
<td>0.94</td>
<td>0.99</td>
<td>0.57</td>
<td>0.60</td>
</tr>
<tr>
<td>MM1CE</td>
<td>1.75</td>
<td>1.75</td>
<td>1.74</td>
<td>0.51</td>
<td>0.01</td>
<td>0.94</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>MM1EC</td>
<td>1.79</td>
<td>1.71</td>
<td>1.74</td>
<td>0.50</td>
<td>0.02</td>
<td>0.57</td>
<td>0.61</td>
<td>0.58</td>
</tr>
<tr>
<td>MM5NA</td>
<td>1.80</td>
<td>1.75</td>
<td>1.79</td>
<td>1.25</td>
<td>1.14</td>
<td>1.21</td>
<td>0.01</td>
<td>0.91</td>
</tr>
<tr>
<td>MM5CE</td>
<td>1.78</td>
<td>1.74</td>
<td>1.77</td>
<td>1.19</td>
<td>1.09</td>
<td>1.19</td>
<td>0.59</td>
<td>0.01</td>
</tr>
<tr>
<td>MM5EC</td>
<td>1.80</td>
<td>1.75</td>
<td>1.79</td>
<td>1.25</td>
<td>1.13</td>
<td>1.20</td>
<td>0.55</td>
<td>0.66</td>
</tr>
</tbody>
</table>
gene pairs, the SAGE Digital Northern libraries were downloaded from the Cancer Genome Anatomy Project (CGAP) site (http://cgap.nci.nih.gov/Genes/GeneFinder). These are across-tissue libraries that include the number of transcripts (in transcripts per 200,000) for the genes of interest. These values were correlated and compared with correlations obtained from methods MM1C and MMSC. Comparisons failed to unequivocally identify the most concordant method. However, ribosomal protein RP1 was involved in four of the 20 biggest discrepancies and in all cases the correlation obtained from method 8 was similar to the one from the SAGE libraries. The same conclusion was achieved with genes OPR51 and IFRD1, both involved in two discrepancies. Results from method MMSC were also more concordant to SAGE results when the latter originated from a large number of libraries (and thus, tissues). MMSC was favoured in comparisons involving a total of 666 libraries, while MM1CE was favoured in comparisons involving 527 libraries.

4.3 Validation studies

Genes represented in the microarray by more than one clone provided the basis to perform cross-validation studies. The assumption is made that, on average, clones of the same gene have a higher correlation among themselves than clones of different genes. Given in Table 4 are the average correlations among all pair-wise clones for genes with more than a specified number of clones and for each normalization method. Methods based on raw means show no particular trend in the average within gene correlations. These remained constant and equal to the average correlation across clones with their corresponding genes (Table 3, diagonals) for methods RMNA and RMEC, and constant and slightly higher than the average across all genes for RMCE. In contrast, normalization methods based on mixed-model equations, while having an average correlation across all genes close to zero (Table 3, diagonals), showed a monotonic increasing pattern for the average correlation among clones with genes highly represented in the microarray.

For the validation experiments, a random third of the 1530 clones with accurate functional annotation were treated as unknown and the performance of each normalization method in identifying the proper gene pairs, the SAGE Digital Northern libraries were downloaded from the Cancer Genome Anatomy Project (CGAP) site (http://cgap.nci.nih.gov/Genes/GeneFinder). These are across-tissue libraries that include the number of transcripts (in transcripts per 200,000) for the genes of interest. These values were correlated and compared with correlations obtained from methods MM1C and MMSC. Comparisons failed to unequivocally identify the most concordant method. However, ribosomal protein RP1 was involved in four of the 20 biggest discrepancies and in all cases the correlation obtained from method 8 was similar to the one from the SAGE libraries. The same conclusion was achieved with genes OPR51 and IFRD1, both involved in two discrepancies. Results from method MMSC were also more concordant to SAGE results when the latter originated from a large number of libraries (and thus, tissues). MMSC was favoured in comparisons involving a total of 666 libraries, while MM1CE was favoured in comparisons involving 527 libraries.

4.3 Validation studies

Genes represented in the microarray by more than one clone provided the basis to perform cross-validation studies. The assumption is made that, on average, clones of the same gene have a higher correlation among themselves than clones of different genes. Given in Table 4 are the average correlations among all pair-wise clones for genes with more than a specified number of clones and for each normalization method. Methods based on raw means show no particular trend in the average within gene correlations. These remained constant and equal to the average correlation across clones with their corresponding genes (Table 3, diagonals) for methods RMNA and RMEC, and constant and slightly higher than the average across all genes for RMCE. In contrast, normalization methods based on mixed-model equations, while having an average correlation across all genes close to zero (Table 3, diagonals), showed a monotonic increasing pattern for the average correlation among clones with genes highly represented in the microarray.

For the validation experiments, a random third of the 1530 clones with accurate functional annotation were treated as unknown and the performance of each normalization method in identifying the proper gene annotation tested. This procedure was repeated 100 times, each time using a different (random) third of the clones as test clones. In each run, gene expression vectors were obtained by averaging element by element (i.e. each of the 23 conditions) those vectors corresponding to clones that were annotated to the same gene. Finally, the correlation between each clone and the genes was computed. At various thresholds for the clone to gene correlation coefficient, each clone in the test sample can be categorized in one of four ways: true positives (TP; if the correlation is greater than the threshold and the clone corresponds to the gene), true negatives (TN; if the correlation is less than or equal to the threshold and the clone does not correspond to the gene), false positives (FP; if the correlation is greater than the threshold but the clone does not correspond to the gene), and false negatives (FN; if the correlation is less than or equal to the threshold but the clone corresponds to the gene). To judge overall performance, for each correlation threshold, three parameters were recorded: FDR from the proportion of clones wrongly assigned to a gene; sensitivity (SENS) from the proportion of clones correctly assigned to a gene; and specificity (SPEC) from the proportion of clones correctly not assigned to a gene.

### Table 4. Average correlation among all pair-wise clones for genes with more than x clones and by each normalization method

<table>
<thead>
<tr>
<th>x</th>
<th>Genes</th>
<th>Normalization method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>195</td>
<td>0.50</td>
</tr>
<tr>
<td>2</td>
<td>106</td>
<td>0.48</td>
</tr>
<tr>
<td>3</td>
<td>71</td>
<td>0.53</td>
</tr>
<tr>
<td>4</td>
<td>48</td>
<td>0.55</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>0.56</td>
</tr>
<tr>
<td>6</td>
<td>21</td>
<td>0.61</td>
</tr>
<tr>
<td>7</td>
<td>15</td>
<td>0.57</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>0.55</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>0.48</td>
</tr>
</tbody>
</table>

### Table 5. Percentage of FDR and SENS for correlation between clone and gene greater than x and by each normalization method

<table>
<thead>
<tr>
<th>x &gt;</th>
<th>Normalization method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.50</td>
</tr>
<tr>
<td>2</td>
<td>0.60</td>
</tr>
<tr>
<td>3</td>
<td>0.70</td>
</tr>
<tr>
<td>4</td>
<td>0.80</td>
</tr>
<tr>
<td>5</td>
<td>0.85</td>
</tr>
<tr>
<td>6</td>
<td>0.90</td>
</tr>
<tr>
<td>7</td>
<td>0.95</td>
</tr>
</tbody>
</table>

**FDR**<sup>a</sup> = TP/(TP + FN) is the proportion of clones wrongly assigned to a functional annotation.

**SENS**<sup>b</sup> = TP/(TP + TN) is the proportion of clones correctly assigned to a functional annotation.

Table 5 presents the results of the validation study. In all cases and as expected, FDR decreased with increasing threshold of correlations. However, at correlations as high as 0.85 and among the raw mean-based methods, only RMCE achieved an FDR below the nominal 5%. In contrast, and at this 0.85 threshold for the correlation, mixed-model-based methods achieved an FDR <0.5%. A higher threshold for the correlation coefficient before a decision is made amounts to a smaller statistical significance. This explains the reduction in SENS with increasing thresholds that was observed for all normalization methods. For a given correlation threshold, similar SENS was observed for raw mean-based and pentavariate mixed-model-based methods. These values were larger than those observed for the univariate mixed-model-based methods. However, at the same level of FDR (say 7–8% achieved at correlations >0.50 and >0.85 for mixed-model-based and raw mean-based methods, respectively), the

---

<sup>a</sup>FDR = TP/(TP + FN) is the proportion of clones wrongly assigned to a functional annotation.

<sup>b</sup>SENS = TP/(TP + TN) is the proportion of clones correctly assigned to a functional annotation.
Normalization in gene co-expression studies

1. RMNA  2. RMCE  3. RMEC

4. MM1NA  5. MM1CE  6. MM1EC

7. MM5NA  8. MM5CE  9. MM5EC

Fig. 2. Gene × gene correlation matrices generated by each of the normalization methods. Genes are sorted by RefSeq sequence identification.

SENS of the mixed-model-based methods was several magnitudes larger than the SENS of the raw mean-based methods (65–70% against 20–40%). The pattern for SPEC was the complement of that for FDR.

4.4 Visualization studies

In order to further compare normalization methods, the resulting correlation matrices were visualized.

Figure 2 illustrates the gene × gene correlation matrices (624 × 624) that resulted from each normalization method and where genes are sorted by their NCBI RefSeq accession number (i.e. NM_#####). As a result of their strong skewness, red dominates in RMNA, RMCE and RMEC, making the task of identifying specific patterns nearly impossible. The pictures from methods based on mixed-model equations are sparser and, mostly for the pentavariate methods, reveal a block of genes towards the upper left corner with a very particular pattern. These genes were identified as ribosomal proteins (RPs) that are known to be highly correlated to each other (Stuart et al., 2003) and possess (at least those from our data set) sequential RefSeq accession numbers.

The relative positioning of cDNA probes on microarray slides has been shown to introduce significant correlations between pairs of genes (Kluger et al., 2003). Correlation matrices corresponding to printing block × printing block (48 × 48) are given in Figure 3. Unadjusted means (methods RMNA and RMEC) fail to account for the spatial features of the slide layout and present extreme correlations among all blocks. Instead, the mixed-model-based methods (that included the random effect of clone × array slide × printing block interaction) generated pictures that clearly demonstrate improved accounting for spatial features. Mixed-model-based methods that normalized for the between clones and within experiment variance (MM1CE and MM5CE) produced pictures that were identical to those from unadjusted methods (MM1NA and MM5NA). This phenomenon was due to the fact that no two samples from different experiments were hybridized on the same microarray. In contrast, and because no clones were printed on two different blocks, \( \hat{\mu}_{E} \) and \( \hat{t}_{E} \) for two blocks \( i \) and \( j \) are not independent but affected by the spatial location of both blocks on the array. For this reason, MM1NA and MM1EC (and hence, MM5NA and MM5EC) yielded different pictures for the printing block × printing block correlations.
Finally, Figure 4 presents the gene × gene correlation matrices that resulted from (the possibly optimal) MM5CE after sorting the genes according to four different criteria as follows: average correlation with the rest of the genes, number of clones, gene symbolic name (using the standard human gene symbols), and chromosome location based on human gene order in the genome sequence. As expected, sorting genes by their symbolic name revealed a clear cluster of highly positively correlated RPs. In addition, smaller clusters of positively correlated genes corresponding to creatinine, collagens and cytochrome oxidases (negatively correlated to RP), follistatins (positively correlated to RPs and negatively correlated to collagens), and myosins (negatively correlated to RPs and positively correlated to collagens). Sorting genes by their chromosome did not reveal any particular pattern. Due to the sparse coverage of the genome in the present data set, none of the different genes are close enough to provide a reasonable expectation of correlated expression due to close physical association.

One rather unexpected finding was the observation that the expression of the collagen and myosin structural subunits was negatively correlated with the expression of the cluster of ribosomal proteins. Given the range of conditions examined, including animals on a starvation diet and very well fed animals, we expected to find a positive correlation between ribosomal proteins and muscle structural proteins. However, it is possible that this negative correlation is a spurious result due to the standardization of the amount of mRNA in the hybridizations, because the ribosomal protein transcripts are major components of muscle mRNA. An alternative explanation is that this observation is due to the total amount of RNA per cell not being constant. Potentially, starved animals may produce substantially less RNA per cell than well-fed animals. We intend to explore the reasons for this rather unexpected result in future analysis of the data.

The clusters of proteins that include the myosins, myosin-associated proteins and the collagens were investigated in more detail (Figure 5). A major cluster of positively correlated proteins

Fig. 3. Printing block × printing block correlation matrices generated by each of the normalization methods. The microarray platform contained 48 blocks (with 200 spots each) arranged in 12 rows by 4 columns.
### Fig. 4.
Gene × gene correlation matrices using normalization method MMSCE after sorting the genes according to four different criteria: (A) average correlation; (B) number of clones; (C) symbolic name and (D) chromosome location.

### Fig. 5.
Correlation matrix for the clusters of proteins that include the myosins and the collagens.

includes the myosin heavy chain proteins MYH1 (skeletal muscle) and MYH7 (cardiac, slow), the myosin essential light chain proteins MYL1 (skeletal, fast) and MYL3 (skeletal/ventricular, slow), and the myosin associated proteins MYBPC2 (myosin binding protein C, fast type) and MYOM2 (myomesin). The expression of most of these proteins is also highly correlated in the CGAP SAGE data. The expression of the regulatory light chain protein MYL2 (cardiac, slow) and a muscle sarcomeric protein, MYOZ2 (Frey et al., 2000), are highly positively correlated, but their expression is not correlated with the expression of the large cluster containing the other myosin subunits. A similar pattern of correlations is also present in the CGAP SAGE data. The expression of three of the four collagens is also correlated with the myosin cluster, albeit more weakly than within the myosin cluster. The co-expression of collagens I and III has been described (Miskulin et al., 1986), whilst collagen V is associated with different systems and in our analysis its expression is not significantly correlated with the expression of the other collagens. Again this matches the correlations observed with the CGAP SAGE data. Interestingly, the expression of the two muscle transcription factors, MYOG (myogenin) and MYF6 (myogenic factor 6) is not correlated, they are also not correlated with the expression of any of the other proteins in this section of the analysis.
5 CONCLUSIONS

Microarray experiments are notorious for producing data sets that are affected by a large number of systematic (non-biological) effects. Nevertheless, these data are an invaluable source for inferring protein associations and gene regulatory networks. Mixed-model equations are a natural framework to account for systematic sources of variation because they can accommodate correlation structures in a rather general manner. We have shown that raw mean-based methods of normalization are sub-optimal for inferring gene co-expressions as measured by the correlation between pairs of genes in various conditions. The choice of univariate or multivariate mixed-models will be greatly influenced by the accuracy of parameter estimates and by computing demands. More complex models would not provide better predictions unless their parameters can be estimated with enough accuracy. With respect to computing demands, in the present study, the univariate model with five variance components contained 694,098 equations and took 7 h and 32 min of CPU time on a Dell PC 2.2 GHz running Red Hat Linux 9. In contrast, the penta-variate model with 35 (co)variance components contained 925,167 equations and took 21 h and 3 min of CPU time on the same computer.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Cooperative Research Centre for Cattle and Beef Quality, a joint venture of The University of New England, NSW Department of Agriculture, Queensland Centre for Cattle and Beef Quality, a joint venture of The University and New South Wales Department of Primary Industries and CSIRO, for making available the microarray datasets.

REFERENCES