STATISTICAL CONSIDERATIONS IN THE ANALYSIS OF GENE EXPRESSION DATA FROM HETEROGENEOUS SOURCES

P.C. Thomson, M. Singh, and H.W. Raadsma

ReproGen, Faculty of Veterinary Science, University of Sydney, Camden NSW 2570

SUMMARY

Combining gene expression data from heterogeneous sources has the potential to increase our understanding of comparative genomics, amongst other things. However, the statistical analyses must address inherent differences across experiments before a combined analysis can be undertaken. A procedure to achieve this is described here, and the methods are illustrated by an analysis of gene expression data in the sheep and cow looking at three stages of lactation.

INTRODUCTION

Increasingly gene expression arrays are being used as a tool for understanding the genetic architecture of complex traits. Whilst initial work was mainly focused on human and model animal species, there is now considerable attention to the application of these techniques in livestock. One particular area that has not received much attention hoverer is the integration of gene expression data from heterogeneous sources, such as across different species, or across different platforms within the same species. However, there are many potential advantages if a combined analysis using gene expression data from multiple sources can be conducted. For example, a combined analysis across species will lead to a better understanding of comparative genomics, and may also elucidate how useful a model species is, as in the case of a sheep as a small-ruminant model for the cow. Further, situations arise where researchers have access to gene expression data from more than one platform, and it would be desirable to explore these data to get an overall picture of gene expression, as well as to understand possible differences between platforms, and incorporate these in formal meta-analyses.

Combining data from heterogeneous sources requires careful statistical consideration, and shares some of the same issues involved with meta-analysis of genomic data. The following describes a three-stage process that was developed for the analysis of a component of a lactation genetics study conducted by the CRC for Innovative Dairy Products. In particular, the methods developed will be illustrated by an application to a comparative study into patterns of gene expression over the lactation cycle of cows and sheep, with values recorded pre-peak, peak, and post-peak lactation from samples derived from mammary gland tissues using a common bovine gene expression array developed by Affymetrix. This array has been shown to have utility in both cattle and sheep and represents over 20,000 gene targets.

METHODS

The first stage involves normalisation of the raw expression data across the series of expression of arrays. In some situations, it may be possible to perform a joint normalisation using all data simultaneously by processing this through a standard procedure, such as RMA-normalisation (Irizarry *et al.* 2003). This is necessary to align the distributions (across genes) of all expression chips to be similar, and hence comparable. However, an additional manual step of quantile-normalisation will usually need to be applied to complete this process. Firstly, to introduce some notation, assume that there are two heterogeneous groups to consider, ovine and bovine in the current situation, but this may be generalised to multiple heterogeneous data sets. Assume we have expression arrays $i = 1, 2, \ldots, n_1, n_1 + 1, n_1 + 2, \ldots, n_1 + n_2$, where there are n_1 and n_2 arrays respectively for group 1 and 2, and that y_{ij} represents the (normalised) expression data for array *i* at gene *j*, *j* = 1, …, *g*. Then proceed as follows:

1. For each expression array, sort the (normalised) expression values from smallest to largest, say $y_{i(1)}, y_{i(2)}, \ldots, y_{i(g)}$, *i* = 1, … $n_1 + n_2$.
2. Obtain the means for each of these *g* order statistics by averaging over all the arrays, i.e. $\bar{y}_{(1)}, \bar{y}_{(2)}, \ldots, \bar{y}_{(g)}$. Similarly, obtain these *g* mean order statistics for each group by averaging over the n_1 and n_2 arrays for Groups 1 and 2, say, $\bar{y}_{1(1)}, \bar{y}_{1(2)}, \ldots, \bar{y}_{1(g)}$ and $\bar{y}_{2(1)}, \bar{y}_{2(2)}, \ldots, \bar{y}_{2(g)}$.
3. Adjustment is made separately for each group, by means of linear interpolation for each expression level, y_{ij}. The interpolation uses the following before-after (*x*,*y*) pairs: $\{ (\bar{y}_{1(1)}, \bar{y}_{(1)}), (\bar{y}_{1(2)}, \bar{y}_{(2)}), \ldots, (\bar{y}_{1(g)}, \bar{y}_{(g)}), \}$ for Group 1, and a similar set for Group 2. So if the value of y_{ij} in Group 1 lies between $\bar{y}_{1(k)}$ and $\bar{y}_{1(k+1)}$, then the adjusted expression value is calculated as $y_{ij}^{(\text{adj})} = \bar{y}_{(k)} + (y_{ij} - \bar{y}_{1(k)}) \times (\bar{y}_{(k+1)} - \bar{y}_{(k)}) / (\bar{y}_{1(k+1)} - \bar{y}_{1(k)})$, with a similar procedure for Group 2 arrays. A special contingency is required to handle the extreme expression values below $\bar{y}_{1(1)}$ or $\bar{y}_{2(1)}$, or above $\bar{y}_{1(g)}$ or $\bar{y}_{2(g)}$.

The result is that the overall distribution of expression levels for the two (or more) groups will be the same, thus allowing direct comparisons. Code to undertake this step has been written in R with the `approx()` function an efficient means to perform the interpolations.

The second stage involves estimating the effects of each gene across the different states, and this is achieved by fitting a single large-scale linear mixed model to all the expression data. Specifics of fixed and random effects to be included in the model will depend on the particular study, but the general rule is to include all relevant sources of variation in the model. This in an extension to the linear mixed model technique described in Sharp *et al.* (2008). A typical linear mixed model will be of the form

$$y = \mu + \text{Array} + \text{Gene} + \text{Gene.Group} + \text{Gene.State} + \text{Gene.Group.State} + \varepsilon$$

where $y = \log_e(\text{RMA})$, the normalised expression values; μ = overall mean expression value; and Array = fixed effect of array *i*, *i* = 1, …, $n_1 + n_2$. All the remaining terms are random, namely, Gene = effect of gene *j*, *j* = 1, …, *g*; Gene.Group = effect of gene in a particular group; Gene.State = effect of gene in a particular state; Gene.Group.State = effect of gene in a particular group-state combination; and ε = random error. Here, State refers to one of the comparisons of interest, e.g. expression levels at pre-peak, peak or post-peak of lactation. The sum of the Gene.State and Gene.Group.State BLUP solutions of the random effects will be used to assess the effect of a gene for subsequent analysis. ASReml (Gilmour *et al.* 2006) can be used to fit such models.

The third stage involves assessing which genes are differentially expressed (DE) in different states. Histograms of these effects for each Group×Stage combination show a heavy-tailed distribution, suggesting the existence of two groups of effects, one with a large variance (DE genes) and one with small variance (non-DE genes) (Figure 1). This is modelled is a two-component mixture distribution, with the DE component being $N(0, \sigma_1^2)$ and the non-DE component $N(0, \sigma_0^2)$ with $\sigma_1^2 > \sigma_0^2$. The (prior) probability of any gene being DE is π_1. The mixture model is fitted to each Group×Stage set of gene effects using the E-M algorithm (McLachlan and Basford 1988) and returns estimates of π_1, σ_1^2 and σ_0^2. The posterior probability (τ_j) of the gene being DE in that particular Group×Stage combination is calculated by an application of Bayes rule, $\tau_j = \pi_1 f_1(z_j) \big/ \left[\pi_1 f_1(z_j) + (1 - \pi_1) f_0(z_j) \right]$, where $f_1(\cdot)$ and $f_0(\cdot)$ are the normal probability density functions for the DE and non-DE components, respectively, and z_j is gene j effect. Values of $\tau_j > ½$ indicate the gene is more likely DE than non-DE. However, in order to reduce false positives, it is preferable to select a higher threshold, such as $\tau_j > 0.8$.

Having identified the subset of genes that are DE in a particular Group×Stage combination, comparisons across groups may then be of interest. For example, when groups may represent different species, it is of interest to see which genes are DE in both species, and to investigate the direction (up- or down-regulation of effects).

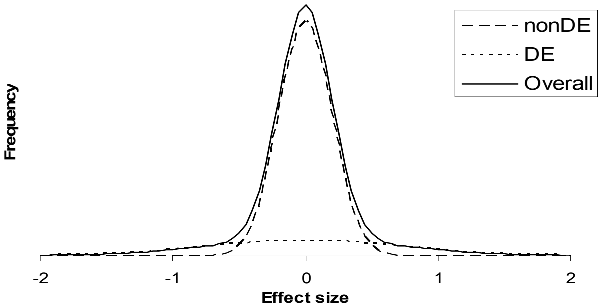

Figure 1. Mixture model for the distribution of DE and non-DE genes.

RESULTS AND DISCUSSION

As indicated above, the methodology outlined here will be illustrated by two gene expression studies conducted into lactation in sheep and cattle. Both studies examined three time points (states), pre-peak, peak, and post-peak lactation, and bovine Affymetrix arrays were used in both studies, For the sheep experiment, four arrays were used per time point, and for cattle five were used (although one array for the post-peak was not usable).

Despite joint RMA-normalisation of all 26 arrays, differences were still apparent between the ovine and bovine arrays, as shown in by the boxplots in Figure 2A. The additional normalisation step using the linear interpolation method was applied and the resultant distributions of adjusted expression values are shown in Figure 2B. Clearly the distributions are now very similar.

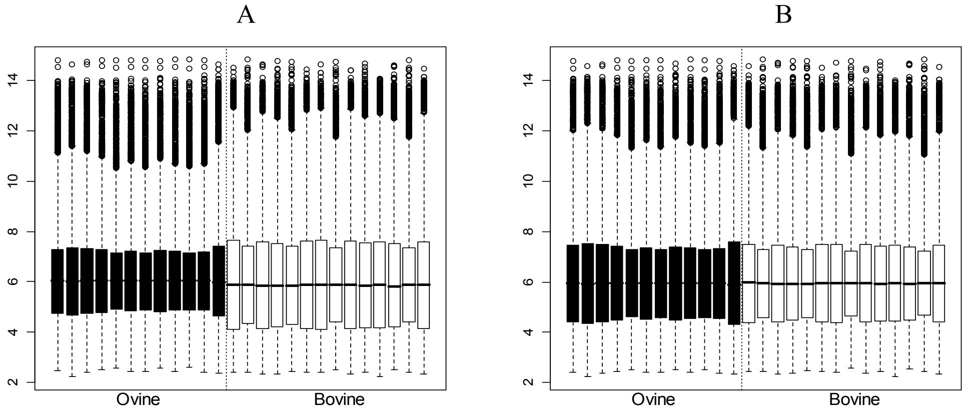

Figure 2. Boxplots of the distributions of expression values before (A) and after (B) the additional linear interpolation normalisation step.

Using the large-scale linear mixed model, BLUP solutions of the random effects were obtained, and the six sets (species × period) of genetic effects determined. In the process of fitting the mixture model, the probability of it being DE is calculated (τ_j), and these values for sheep in the pre-peak are shown in Figure 3. This shows a typical pattern with a small cluster of genes with a very high probability of being DE ($\tau_j \sim 1$), although any gene with $\tau_j > 0.8$ is considered DE.

These estimated effects (BLUP values) were obtained for both sheep and cow samples, and Figure 4 shows a smoothed scatter plot of sheep and cow genetic effects in the pre-peak phase. It is seen that many genes have similar expression patterns across sheep and cow, as indicated by the large strong positive association.

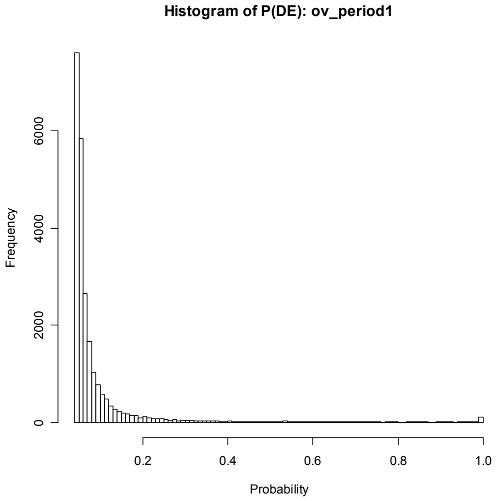

Figure 3. Distribution of the probability of a gene being DE for sheep in the pre-peak lactation stage

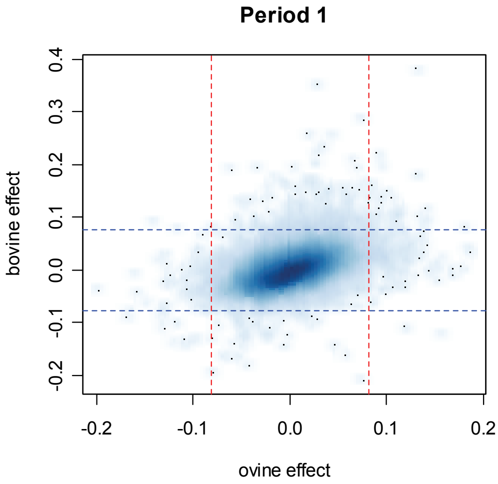

Figure 4. Joint distribution of genetic effects for sheep and cow in the pre-peak lactation stage. Vertical and horizontal dashed lines indicate thresholds for genes being DE.

CONCLUSION

Combining expression data from heterogeneous sources, be they from different species, tissues, or perhaps experimental platforms, has the potential to add considerable insight into our understanding of gene function, and will add additional value over what can be provided by studying gene expression patterns from a single isolated experiment. The methods outlined here provide a three-stage procedure that will allow these meta-analyses of expression data to be undertaken to move to an entire transcriptome analysis of a particular physiological state or organism.

REFERENCES

Gilmour, A.R., Gogel, B.J., Cullis, B.R., and Thompson. R. (2006) "ASReml User Guide. Release 2.0". VSN International Ltd, Hemel Hempstead.

Irizarry, R. A., Hobbs, B., Colin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). *Biostatistics* **2:**249.

McLachlan, G.J., and Basford, K.E. (1988) "Mixture Models: Inference and Applications to Clustering". M. Dekker, New York.

Sharp, J.A., Mailer, S.L., Thomson, P.C., Lefevre, C., and Nicholas, K.R. (2008) *Molecular Cancer* **7**:1.