LINEAR MIXED EFFECTS MODELS FOR MICROARRAY GENE EXPRESSION DATA

S.L. Rodriguez-Zas and B.R. Southey

Department of Animal Sciences, University of Illinois at Urbana-Champaign, Illinois 61801, USA

INTRODUCTION

Functional genomic studies investigate the simultaneous expression of multiple genes under different conditions or instances. In general, the microarray technology is based on the simultaneous hybridization of two labeled (e.g. dyes Cy3 and Cy5) cDNA samples to an array of target sequences. The microarray background intensity is subtracted from the sample intensity and follow-up normalization minimizes the impact of one channel (dye) being systematically brighter than the other. When samples from multiple instances and reference (in every array) are analyzed, the response is the difference between the logarithms of the sample to reference ratio between instances. The identification of expression patterns can be hindered by unaccounted noise including within array variation (e.g. variation in the rate of dye incorporation and spot size) and among array variation originating during the fabrication, hybridization and scanning phases. The large amount of information available poses computational, statistical and interpretational challenges partially addressed by clustering, network, projection and t-test methods. The previous response and approaches have limitations including restricted use of the information contained by the data, limited ability to account for all sources of variation and challenging interpretation of the results.

At the present, all livestock species microarrays contain less than 8,000 unique sequences. To evaluate models suitable for a wide variety of situations (e.g. genome-wise studies), a *Caenorhabditis elegans* microarray experiment described by Kim *et al.* (2001) was used. Most of the *C. elegans* genes (19,200) have been identified but many lack of known function. This situation is comparable to that of the livestock genome projects in the next 4 years. Further, the study of gene expression along the multiple developmental stages of the *C. elegans* offers a representative combination of data structure complexity, amount of information and multiple sources of variation. The resulting recommendations will be pertinent to livestock microarray studies in the short term. The objective of this study was to evaluate alternative models to describe gene expression profiles across multiple conditions while accounting for documented sources of variation. Two related issues were considered, a) normalization, and b) paired and independent sample approaches.

MATERIALS AND METHODS

Data and Microarray Protocol. Intensity records from 19,200 *C. elegans* genes along six developmental stages (egg, larvae stages L1 to L4 and adulthood) were studied using a 21,120 spot array. Messenger RNA at any one developmental stage was transcribed into Cy3-labeled cDNA and a multi-stage reference sample was transcribed into Cy5-labeled cDNA. Data from

three to four arrays (replicates) per developmental stage were available. A detailed description of the microarray construction was provided by Kim *et al.* (2001).

Models. A natural logarithmic transformation to the observations was used since the Shapiro-Wilk and Kolmogorov-Smirnov D statistics indicated substantial departure from Normality. A two-tier approach was used to analyze the intensities. In the first tier, the sample intensities were adjusted for linear, quadratic and cubic background trends and the impact of a dye effect, partially confounded with stage, was evaluated. The resulting estimates were compared to data normalized by equating the total Cy3 and Cy5 intensities.

In the second tier, two complementary descriptions of the adjusted intensities were used to investigate the gene expression profile across developmental stages. In a paired-sample (PS) model, the response variable was the difference between the adjusted intensities from each developmental stage and the reference sample. The model included a gene by stage fixed effect (115,200 levels) and a random array effect (20 levels) to account for the correlation among observations within an array. Matching sample and reference intensities can lead to a more powerful study due to the occurrence of natural dependencies within array spot. In an independent sample (IS) model, the stage and reference (instances) adjusted intensities were considered as independent and described with a linear combination of the effects of array (20 levels), instance (7 levels), gene (19,200 levels) and second order interactions (Kerr and Churchill 2001). Since the interest is to compare the changes in gene expression across instances, gene and instance were considered fixed effects and the interaction between array and gene was included as a random effect. The null hypothesis of no expression variation across instance was tested using an F statistic and Bonferroni adjustment to account for the multiple testing was conducted.

RESULTS AND DISCUSSION

Figure 1 presents the recorded background intensity of the reference sample in two arrays from the same developmental stage. The substantial variation across and within array substantiated the need for complex models that can account for multiple sources of variation and covariance structures. The correlation between the first-tier background adjusted and the conventionally normalized intensities ranged between 84% and 96% among arrays. This suggests that the conventional normalization fails to account for high order background effects. The null hypothesis of no dye effect was rejected based on likelihood ratio tests although dye and instance effects are partially confounded in this experiment.

The PS analysis detected 50% fewer significant (Bonferroni adjusted $P < 0.05$) expression changes than the IS analysis. The majority (95%) of the PS significant results were confirmed by the IS analysis.

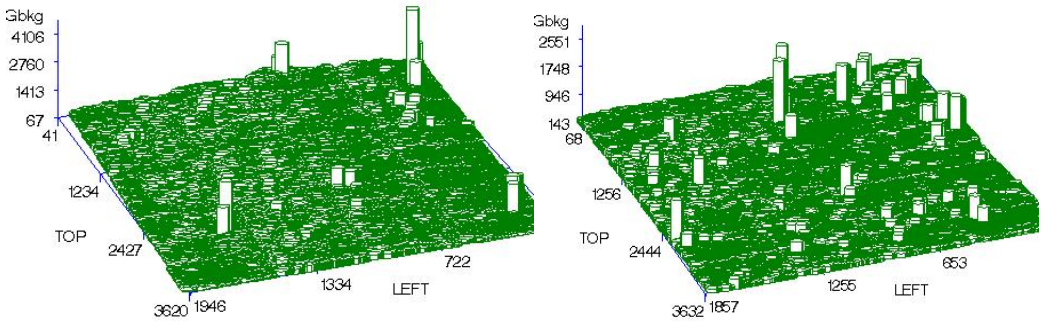

Figure 1. Reference background intensities (Gbkg) from two microarrays where TOP and LEFT denotes the array row and column number from the top and left corner, respectively

The PS model was effective in minimizing the competing intensity dissimilarity except for stage of development. The IS approach modeled the data from the stage and reference sample as independent observations. The results imply that sources of variation other than developmental stage and array were not fully accounted for by the PS model. The significant results unique to PS, reflect the higher power of this approach to detect smaller differences than the IS approach, when the assumptions are appropriate.

The ability of the proposed models to detect known co-regulated genes and to enhance the gene function assignment based on predictive bioinformatics is evident in figure 2. Results from the second-tier models indicated that known and putative members of the family of Actin genes (associated to muscle contraction and relaxation) and muscle and filament related protein genes showed a strong expression at the intermediate stages of development. Heat Shock genes associated with thermo-tolerance were grouped, based on their profile, into a group with moderate expression across developmental stages and another with lower expression at advanced stages. Two putative Heat Shock genes, based on bioinformatics tools, followed the patterns of the first group.

CONCLUSIONS

The proposed approach incorporated multiple sources of variation by adjusting the intensity for background and dye effects while accounting for the effects of gene, stage and array. Results from the PS and IS analyses suggest that a model that allows for variable dependency across array spots would be more adequate. Results from the second-tier models established expected co-regulation patterns and aided in the characterization of lesser-known sequences. The repeated measurement model allowed to study trends across stage and can be extended to more complex variance-covariance structures. The proposed approach provided statistically robust and biologically meaningful answers to the analysis and interpretation of functional genomic studies.

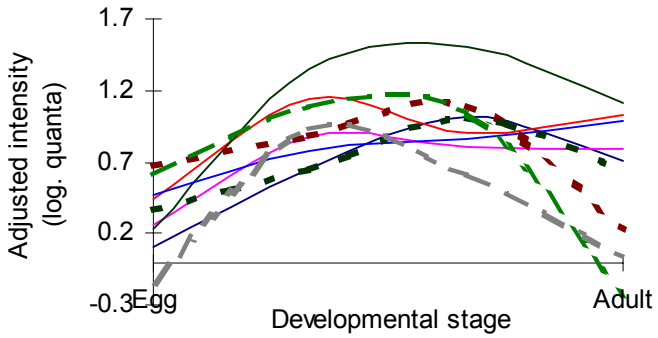

Figure 2. Estimated expression of known (solid lines), putative (small-dotted lines) Actin genes and muscle related genes (large-dotted lines) across developmental stages

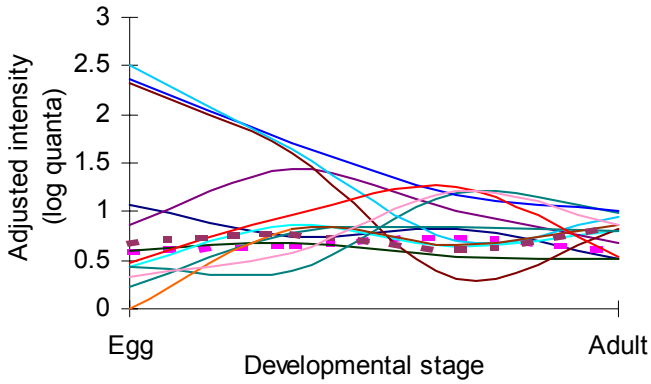

Figure 3. Estimated expression of known (solid lines) and putative (dotted lines) Heat Shock genes pertaining across developmental stages

ACKNOWLEDGEMENTS
We wish to thank Dr. S. K. Kim for providing the microarray data.

REFERENCES

Kerr, M.K. and Churchill, G.A. (2001) *Genet. Res.* **77** : 123-128.
Kim, S.K., Lund, J., Duke, K., Jiang, M., Stuart, J.M., Eizinger, A., Wylie, B.N. and Davidson, G.S. (2001) *Science* **293** : 2087-2092.