SYSTEMS GENETICS APPROACHES FOR GENETIC IMPROVEMENT OF LIVESTOCK

H. N. Kadarmideen

CSIRO Livestock industries, JM Rendel Laboratory, PO Box 5545 Rockhampton Mail Centre,
Rockhampton, QLD 4702, Australia

SUMMARY

Identification of causative genes and quantitative trait loci (QTL) is ineffective unless we identify, on a systems level, also regulatory genes elsewhere on the genome that significantly control their expressions. Until the invention of high-throughput genomic and transcriptomic techniques, it has not been possible to identify such major regulatory genes for QTL and candidate genes on a whole genome scale. *Genetical genomics* is a novel method in integrative biology that helps us to investigate the inheritance of such regulatory loci, so called *expression quantitative trait loci* or eQTL. This article formulates common biological questions relevant to livestock microarray gene expression profiling (MGEP) experiments and introduces genetical genomics methods to identify eQTLs. Main objective of this article is to present statistical methods and models (with an emphasis on BLUP mixed models) and tests of hypothesis for interval *eQTL* mapping. Illustrations from our own recent investigations are provided. Further, some concepts in the design of genetical genomics experiments with an aim to obtain maximum power at reduced costs for MGEP and Single Nucleotide Polymorphisms (SNP) genotyping of animals are developed via *selective expression profiling* and *selective genotyping* for eQTL mapping. Finally possible genetic improvement of livestock using combined genetic, genomic and transcriptomic data in an expression selection index (*eSI*) framework is considered.

INTRODUCTION

Systems genetics is a new emerging branch of systems biology which consists of integrated genetic, genomic, transcriptomic and proteomic investigations. Systems genetics paradigm was first investigated for diabetes and obesity related genes in recombinant inbred lines of mouse (Kadarmideen *et al*. 2006); this was followed by applications in human and pigs (Kadarmideen and Janss 2007) and in plants (VonRohr *et al*. 2007). *Functional genomics* or precisely *transcriptomics* via microarrays involves study of gene expressions as measured by the quantity of mRNA for all the genes in the microarray chip. Each microarray experiment can accomplish the equivalent of thousands of traditional 'one gene-one experiment' genetic tests in parallel, which triggered dramatic use of this technology in all branches of life sciences in the last decade. Many livestock research organizations and universities have began microarray gene expression profiling (MGEP) experiments with specific biological hypothesis mostly to underpin the functional biology and genomics but no coherent ideas have been developed on how to apply and integrate such MGEP experiments with traditional (molecular) breeding in livestock production. It is important therefore to formulate hypotheses and design experiments which will improve genetic merit of livestock and subsequently the livestock production. Based on this background, one of the objectives of this article is to formulate common biological questions and problems that are relevant to livestock MGEP experiments.

Statistical and molecular genetic methods for interval quantitative trait loci (QTL) mapping are well established. The *genetical genomics (GG)* or e*xpression genetics* is an integrative method that combines Mendelian genetics with high throughput transcriptomics. The concept was proposed by Jansen and Nap (2001) and applied by many others (e.g. Schadt *et al*. 2003). These methods allow us to investigate if transcript variation of genes profiled on microarrays is inheritable and under genetic control (of few regulatory genes). With the brief overview of genetical genomics methods, the other objective of this article is to present and discuss statistical methods and models under genomic BLUP framework and tests of hypothesis for interval *eQTL* mapping methods. Illustrations from our own recent genetical genomic investigations in mouse for diabetes, obesity and stress related genes are provided. The single most limiting factor in integrated genetical genomics study is the experimental costs and labour involved in collection and processing of samples and conduct both MGEP and high density SNP genotyping. In this paper, some concepts in design of genetical genomics experiments with consideration for reducing costs for MGEP and high density SNP genotyping for eQTL mapping are provided.

MATERIALS AND METHODS

Detection of differentially expressed (DE) genes. In terms of genetic improvement of livestock for better production/performance efficiency, MGEP experiments can be categorized into two major types: 1. To identify major (regulatory) genes for possible use in gene-assisted breeding programs and 2. To identify gene and regulatory networks underlying complex traits to develop systems genetics approaches for genetic improvement. The second type of investigations is not currently the main priority in livestock experiments.

Biological Cases / Questions in Livestock MGEP for Detection of Major (regulatory) Genes are,

1) Which genes show differential expression between different or extreme phenotypes? For example, disease vs healthy (between disease types), aggression vs. submission, fertile vs. infertile, susceptible vs. resistant/tolerant, lean vs. fat, slow vs. rapid growth etc.,
2) Which genes show differential expression between different or extreme environmental conditions? For example, high vs. low nutrition, high vs. low challenge with parasites/pathogens, stressed vs. normal condition, hot vs. cold or temperate climate etc.
3) Which genes show differential expression between different developmental stages of animals? For example, birth to slaughter age in pigs or across different lactation stages or parities in dairy cows, across birth to 200-d, 400-d weights in beef cattle etc.
4) Which genes show differential expression between different genetic backgrounds? For example, purebred, crossbred or composite animals.

Results on number of differentially expressed genes from case 1 shows possible causal genetic factors / candidate genes that are causing different phenotypes, case 2 shows effect of environmental conditions on gene expressions patterns, case 3 shows sets of genes that are up-, down or null-regulated across stages / ages, and case 4 shows genetic basis of differential gene expression patterns. Information obtained from each one of the above cases has implications in how we can breed and manage animals. For instance, possible uses in marker/gene assisted selection (case 1), genotype x environmental interactions (case 2), time or age specific breeding strategies (case 3) and directed breed-specific changes in a breeding program (case 4).

Genetical Genomics or Expression Genetics. In this method, each individual in a pedigreed population is subject to MGEP and high density marker genotyping. QTL mapping techniques are then used, treating expression level of each gene on a microarray as a quantitative trait, to identify genomic control points (eQTL) that influence transcript levels (Jansen and Nap 2001, Schadt *et al.* 2003, Kadarmideen *et al.* 2006). *Cis-acting eQTLs* are a result of genomic sequence variation that resides within or in the close proximity of the gene being regulated. These *cis-eQTLs* are interesting causative or candidate genes for regular QTL mapped into the same location because it pinpoints to a candidate mutation / sequence difference within QTL region which may be spanning several Kb or even Mb. This is one of the reasons why eQTL mapping significantly improves the power of detecting causal mutations and candidate genes in regular gene/QTL mapping. *Trans-acting eQTLs* are regulatory loci, often occurring in clusters or hubs, remotely controlling expression of several other genes elsewhere on the genome. *Trans-eQTLs* are good candidate regions harbouring master regulator genes and common transcription factors as we have recently shown (Von Rohr *et al.* 2007). This eQTL mapping is a complex and highly demanding statistical and computational procedure. In principle, expression levels of thousands of genes measured on a microarray can be mapped onto chromosomes, transcript-by-transcript, but in practice eQTL mapping is restricted to a subset of DE genes or to a few known (candidate) genes or to clusters of gene. Instead of linking each transcript across tens of thousands of markers, one could test one marker at a time for significant linkage across all transcripts as shown by Kendziorski *et al.* (2006). This method amounts to grouping DE genes under marker genotypes and does not separate eQTL regions specific for each transcript. In a way, this is similar to our earlier investigations on mapping clustered transcripts onto the genome (Figure 2, Kadarmideen *et al.* 2006).

Linear mixed model for detection and mapping of eQTLs. Bueno *et al.* (2006) provided some consideration on design of MGEP experiments for genetical genomic studies but not in the context of interval eQTL mapping using grid search and in outbred populations. Other genetical genomics studies have so far conducted eQTL interval mapping by ‘fixed effect’ models. Here a concept based on theory of BLUP mixed models (Henderson 1984) is provided, extending ‘fixed effect’ eQTL mapping to random (mixed) model framework, based on similar principles as found in Bueno *et al.*, (2006). The methodology for out bred populations is provided based on the methodologies of Kadarmideen and Dekkers (1999) and Kadarmideen *et al.* (2000 and 2006) for regular QTL mapping. The model is shown for two-channel arrays but is equally valid for single-channel arrays.

Interval eQTL Mapping. The linear statistical model is

$$y_{ijklmn} = \mu_i + D_{ik} + A_{ik} + E_{ik} + \beta_m p(eQTL)_l + u_{lk} + e_{ijklmn} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Where $\boldsymbol{y}_{\mathrm{ijklmn}}$ is a vector of expression level of gene transcript *i* measured in array *j* belonging to treatment *k* in individual *l*, μ_{i} is an overall mean for the expression of gene transcript *i*, D_{ik} is the effect of dye (green or red), A_{ik} is the effect of array (or slide), E_{ik} is the effect of the experiment/treatment *k*, u_{ik} is the genetic effect (other than eQTL; for instance sire or dam or animals' own) and e_{ijklmn} is a residual term. The term $p(eQTL)_l$ represents the conditional probability of transmission of one of the two eQTL alleles from either parent to the offspring conditional on marker and linkage information on individual and its relatives. The conditional probabilities for eQTL allele transmission from parent to offspring could be assigned based on well established methods (e.g. as in Kadarmideen *et al.* 2000 and 2006 for within-family half-sib designs). The regression coefficient β_{m} is the eQTL effect on transcript abundance, which should be interpreted based on experimental design used (e.g. sib-pair, full-sib, half-sib, BC, F2 regression). In a single channel array (such as Affymetrix), there is one array per experimental unit or an individual and hence the vector $\boldsymbol{y}_{\mathrm{ijklmn}}$ would contain intensity values measured by one dye whereas in two-channel array (cDNA and long oligos), a spot will contain intensity values for the same gene measured in a pair of individuals. In this case, the corresponding values can be read into two elements of vector $\boldsymbol{y}_{\mathrm{ijklmn}}$. The aim is to have each element in the data vector represent one gene in one individual.

In matrix notation, the model (1) is written as

$$\mathbf{Y=Xb+Zu+e}$$

where **y** is a vector of expression phenotypic records, **X** is a design matrix relating fixed effects in **b** to **y**. The vector of random effects is included in **u** where the design matrix, **Z** relate records to these genetic effects. The vector of random residuals is in **e.** The expectation (**E**) and Variances (**V**) of model terms are as follows.

E=Xb and **V(y)=V(Zu+e)** = **ZGZ'+R** where **G** = $\mathbf{A}\sigma_u^2$ where **A** is the additive relationship matrix and **R** is residual variance-covariance matrix. Each expression trait is assumed to follow (log) normal distributions with the above mean and variance-covariances.

Tests of Hypothesis. To test for significance of presence of an eQTL in the marker bracket, the Liklihood Ratio (LR) test statistic can be computed (based on Kadarmideen *et al.* 2000 and 2006) as:

$$\mathrm{LR}=N\,ln\,(RSS_{red}\,/\,RSS_{full})$$

Where N is a total number of individuals, $\mathrm{RSS_{red}}$ is a residual sums of squares obtained from fitting a model under the null hypothesis $\beta_m = 0$, and $\mathrm{RSS_{full}}$ is obtained from fitting the full model under the alternate hypothesis, $\beta_m \neq 0$. To derive significance threshold or cut-off values, data permutation techniques and bootstrapping techniques need to be used to adjust for multiple testing at marker positions. Note that this is over and above the procedures to account for multiple testing to identify DE genes in microarrays.

Marker Regression eQTL Mapping. The above mixed model (1) is extremely time consuming and may even be impractical when thousands of gene transcripts need to mapped onto highly dense genome maps by grid search interval mapping approaches. A simplified interval eQTL mapping model could be based on marker regression QTL mapping (MRM) method developed by Kadarmideen and Dekkers (1999 and 2001) which assumes no genetic model but simply regress phenotype on flanking marker genotypes in a pair-wise multiple regression and filters out the significant marker brackets as a possible QTL region. With the high density SNP genotyping arrays, it would be possible to sort the most significant marker brackets as potential region for eQTL. Therefore the model (1) can be re-arranged as

$$y_{ijklmn} = \mu_i + D_{ik} + A_{ik} + E_{ik} + \beta_{m1}.p(SNP)_{lL} + \beta_{m2}.p(SNP)_{lR} + u_{lk} + e_{ijklmn} \;\ldots\ldots\ldots\ldots\ldots..(2)$$

where, $\mathrm{(SNP)_{lL}}$ and $\mathrm{(SNP)_{lR}}$ are the left and right SNP marker bracketing a potential eQTL. The probability of marker transmission to individual *l* could either take on a score of 0 or 1 or 0.5 if an allele transmission is known with certainty (as in BC or F2 populations) or take on 0-1 probability values in other outbred populations (as in half-sib populations e.g. Kadarmideen and Dekkers 1999).

RESULTS AND DISCUSSION

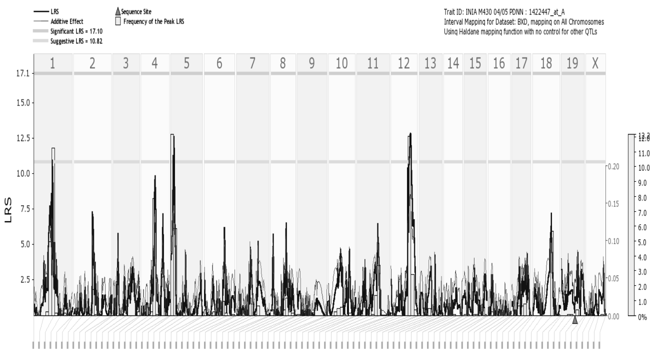

Figure 1 (taken from Kadarmideen et al., 2006) shows a genome-wide detection and mapping of eQTL for *ins1* gene involved in diabetes and obesity in mouse. Three trans-eQTL regions were found one each on chromosomes 1, 5 and 12. Peaks in the heavy line (LRS) show locations of a putative QTL, and the histogram beneath it shows frequent peak location for bootstrap samples. Another example of expression genetics is mapping of *crhr1* gene expression levels (involved in cortisol, an indicator of stress) in mouse to chromosomes to 2 and 13 (Kadarmideen and Janss 2007). These trans-eQTL regions would be potential candidate genomic regions in which regulatory loci (e.g. transcription binding factors) are located.

Figure 1. Whole genome eQTL mapping in mouse for expression of *Ins1* gene located on Chromosome 19 (at 51.83 Mb) showing 3 trans-eQTL peaks at other genomic locations

Design and Power of Genetical Genomics Experiments. Unlike QTL mapping, phenotypes are very expensive for eQTL mapping. It is a known limiting factor across all species as cost of microarrays is still prohibitive to conduct genetic investigations. In addition, the nature of record collection in eQTL mapping is different from QTL mapping in that tissue biopsies need to be conducted which requires trained paramedics and proper experimental facilities as per the animal ethics law. The later problem means that these experiments typically contain only a small number of animals (often in the range of 50 to 200). Small sample sizes in eQTL mapping leads to high False Discovery rates (FDR) of (mostly trans) eQTLs. In general the power to detect cis- and trans-eQTL, at a given type I error rate, depends on, among other factors,

1. Size of eQTL effects
2. Whether gene transcript is cis- or trans-regulated
3. Heritability of gene transcript
4. Population size and number of replicates for each gene on an array
5. Type of experimental design
6. Type and density of genetic markers and their recombination rates with eQTL

eQTL effects depend on source of the phenotypes vis-à-vis tissue type that was used to collect mRNA. Power is usually higher to detect cis- than trans eQTLs. Some tissues may have high heritabilities than others for the same transcript. Experimental designs from crossing inbred lines have higher power than that from out bred populations. Similar to regular QTL mapping, the power is at maximum depending on what combinations of the 6 criterion (or more) we have. There is neither a deterministic way to calculate power nor standard software/algorithm is available yet to calculate power. This is indeed an area for quantitative geneticists to explore based on similar principles for regular QTL mapping. The use of small sample sizes could still result in coarse identification of (mostly non-significant) interesting regions which can be used for further fine mapping and validation by systems biology methods.

Selective genotyping for eQTL mapping. Selective genotyping for QTL mapping (Weller 2001) is well established. In selective genotyping, the aim is to achieve similar power of QTL detection with only a subset of individuals genotyped, similar to that obtained by genotyping all individuals. For eQTL mapping, it involves looking at the distribution of normalized expression phenotype for gene *l* and selecting top and bottom x% (e.g. 5 %) of arrays (in this case units / animals) from the whole population of arrays. At the lower tail of the distribution, it is expected that there would be a high frequency of trans-genes that down regulate gene expression compared to frequency of trans-genes that up regulates gene expression. In the upper tail, the opposite case is expected. That is, in case of bi-allelic trans-genes or eQTL (with Q and q alleles), there would be more *qq* genotypes at the lower tail than at the top where *QQ* genotypes would be more frequent; with the middle of the distribution representing *Qq* heterozygotes. These three groups of animals are therefore more likely to show sequence variation with respect to gene expression differences. The major problem with this approach is that this technique has to be applied to each gene on a microarray and by repeating this for each gene among thousands of genes profiled, one may have genotyped all animals for all markers. Hence the best strategy is to conduct hierarchical clustering and perform selective genotyping for clustered genes and conduct eQTL *heat* mapping as seen in Figure 2 (from Kadarmideen *et al.* 2006). These strategies are particularly important when a GG experiment is conducted with a handful of microsatellites rather than with SNP genotyping arrays.

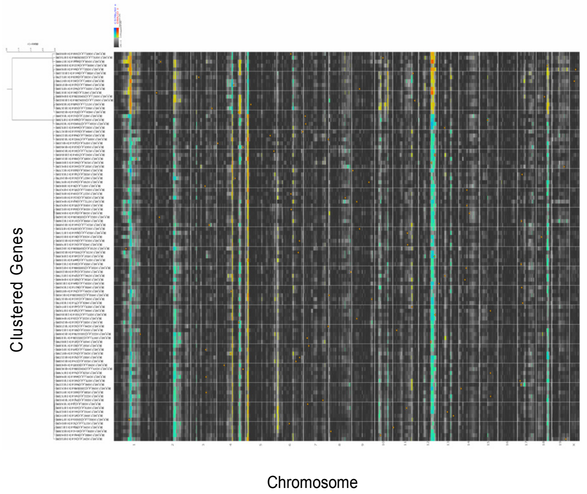

Figure 2: Heat maps, showing eQTL hotspots (red-to-orange) on chromosomes (x-axis) affecting expression of all genes in hierarchical clusters (y axis).

Selective transcriptional profiling for eQTL mapping. For genetical genomics, genotyping at large number of markers is not a major constraint (with availability of high throughput SNP arrays) but the phenotyping (i.e., expression profiling) is. Jin et al. (2004) proposed selective phototyping for improving the efficiency of genetical genomics studies and QTL studies in general. Their method chooses a sub-sample of individuals that are as dissimilar as possible with respect to marker genotypes across the genome or over genomic regions of interest. Selecting animals only based on marker differences across the genome would be randomly linked to expression differences of genes across the genome. However, the use of phenotypic differences will further strengthen the selection process. Wang and Nettleton (2006) and Nettleton and Wang (2006) proposed selective transcriptional profiling approach where they use both traditional trait and marker data to select individuals for transcriptional profiling and then use available data from both selected and unselected individuals to detect eQTL. They introduce 'missing data concept' for expression profiles of unselected individuals but use their trait and marker information in the analysis. The fundamental concept in their approach is evaluating association of transcript abundance of randomly selected genes profiled on microarray with an identified trait QTLs. If the link is statistically significant then the trait QTLs will turn out to be eQTLs for the microarrayed genes. In all the above methods, the eQTLs not associated with the trait and/or not at the same position as a trait QTL will be missed. But such eQTLs may exist for the subsets of genes on a microarray. Therefore clear selective transcript profiling strategies to capture all eQTLs for all the genes (or clusters) on a microarray would be a future research in this area.

Livestock genetic improvement using genetical genomics. Following Kadarmideen *et al.* (2006), and based on Model 1 and 2, individual animals in a pedigree could have an estimated breeding value (EBV) for each gene transcript, an EBV at the candidate gene or regular marker-QTL (which can appear as fixed or random effect in model 1) and an EBV for SNP or eQTL, and an EBV for polygenes. It could be speculated that a future expression selection index (*e*SI) in animal breeding may be a combination of a weighted average score of each EBV as

$$e\text{SI} = b_1\text{*POLY\_EBV} + b_2\text{*Exp\_EBV} + b_3\text{*QTL\_EBV} + b_4\text{*eQTL\_EBV}$$

where weights $b_{1,}$ $b_{2,}$ b_3 and b_4 relate to index weights (based on economic, family information and perhaps tissue- and time-specific criteria) for polygene, gene expression, QTL and eQTL information, respectively. Strictly speaking, for *e*SI, almost all eQTL information to be included here will be a trans-eQTL because cis-eQTL information would already be accounted for in normal QTL. The optimisation of the weights $b_{1,}$ …, b_4 is classically done by evaluating the correlations of the predictors with an end-phenotype and computing these weights as regression coefficients. However, this approach depends on the assumption of linearity (constancy) of correlations and regressions under changing means, and it is doubtful whether this assumption still holds when including expressions and regulator genes (eQTL) affecting pathways is non-linear. Further research on how to derive index weights and optimise selection procedures in these non-linear systems may therefore be anticipated.

CONCLUSIONS

In this article, a general framework for developing hypothesis and designing MGEP experiments in livestock were classified into 4 categories. Genetical genomics links traditional Mendelian genetics with high throughput –omics science in understanding complex polygenic traits. In this paper, a framework for random or BLUP based mixed statistical models to identify and map eQTL were discussed with the opportunities and challenges in implementation of such highly complex and demanding analyses. Under the random model approach, two statistical methods were introduced, one based on interval eQTL mapping approach using high density SNP arrays and the other based on non-interval eQTL mapping using multiple marker (SNP) regression approach. While these methods are appealing, there are statistical and computational challenges in how to collate, analyse and interpret results. Hence this will be an active research for the next generation of animal geneticists. Selective phenotyping and genotyping strategies for both MGEP and eQTL experiments are discussed, which will reduce the costs for such experiments. Finally, possible inclusion of eQTL EBVs in a genomic selection index (so called expression Selection Index) for genetic improvement of livestock is discussed.

REFERENCES

Bueno J.S.S., Gilmour S.G., and Rosa G.J.M. (2006). *Genetics* **174**: 945.

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*, University of Guelph, Guelph, ON, Canada

Jansen R.C., and Nap J.P. (2001) *Trends Genet* **17:** 388.

Jin, C., Lan, H., Attie, A. D., Churchill, G. A., Bulutuglo, D., and Yandell, B. S. (2004). *Genetics* **168:** 2285.

Kadarmideen, H.N., and Dekkers, J.C.M. (1999). *Genet. Sel. Evol*. **31**: 437.

Kadarmideen, H.N., Janss, L.L.G., and Dekkers, J.C.M., (2000). *Genet. Res.* **76**: 305.

Kadarmideen, H.N., and Dekkers, J.C.M. (2001). *J. Anim. Breed. Genet.* **118**: 297.

Kadarmideen, H.N., Li, Y., and Janss, L.L.G (2006). *Genet. Res.* **88**: 119.

Kadarmideen, H.N., Von Rohr, P. and Janss, L.L.G., (2006). *Mamm.Genome* **17**: 548.

Kadarmideen, H.N., and Janss, L.L.G. (2007). *Physiol. Genomics* **29**: 57.

Kendziorski C.M., Chen M., Yuan M., Lan H., and Attie A.D. (2006) *Biometrics* **62**: 19.

Nettleton, D and Wang, D (2006). *Animal Genetics* **37** (suppl.1): 13.

Schadt E.E., Monks S.A., Drake T.A., Lusis A.J., Che N., *et al.* (2003) *Nature* **422**: 297.

Von Rohr, P. Friberg, M., Kadarmideen, H.N., (2006) *J. Comput. Biol. and Bioinformatics* (In Press).

Wang D and Nettleton, D. (2006). *Biometrics* **62**: 504.

Weller, J.I. (2001). *Quantitative Trait Loci Analysis in Animals*, CABI publishing, New York, USA