COMBINING INFORMATION ACROSS TRAITS USING A FACTOR ANALYTIC MODEL INCREASES THE POWER OF QTL DETECTION

W.S. Pitchford¹, A. Esmailizadeh Koshkoii³ and A.R. Gilmour²

¹ School of Agriculture, Food and Wine, University of Adelaide, Roseworthy SA 5371 Australia
² Orange Agricultural Institute, NSW Department of Primary Industries, Orange NSW 2800 Australia
³ Present address: Department of Animal Science, Faculty of Agriculture, Shahid Bahonar University of Kerman, Kerman, Iran

SUMMARY
A multiple marker analysis approach was extended to multi-trait and multiple family situations. It was shown through a simulation study that modeling multiple phenotypes and multiple families in a single linkage analysis simultaneously could markedly increase power, compared with modeling each phenotype or family separately. Applying the approach to mapping data from beef cattle clearly showed its ability to gather information across traits and families.

INTRODUCTION
A QTL may affect more than one trait and this is termed a pleiotropic effect. It is common in gene mapping experiments to measure many traits as a way of value-adding to the investment in genotyping costs. Univariate analysis of relationships between genetic variants (e.g. microsatellite allele, single nucleotide polymorphism, haplotype) and phenotypes are commonly conducted using regression of the phenotype on the marker (e.g. Seaton et al. 2002).

The simplest way to deal with multivariate data is by mapping traits separately and assessing whether the confidence intervals for QTL overlap for some combinations of traits. Hence, in almost all cases, multivariate traits are condensed to allow univariate analysis. One approach is selecting one of the traits as the primary trait and considering the remaining traits as covariates, modifying the mean behaviour of the primary trait. For example, in mapping genes for carcass fatness, treating carcass weight as a covariate runs the risk of masking linkage evidence for genes that impact both traits.

Alternatively, the multivariate data is replaced by one or more linear combinations of the traits. For example, Weller et al. (1996) considered the use of canonical variables which have the advantage of modeling all traits, but the magnitude of the estimated effects are difficult to interpret in terms of the underlying traits. Furthermore, a transformation that produces traits that are either phenotypically or genetically uncorrelated does not ensure that the QTL only influences a single canonical trait. This is because different QTL affecting a trait may have different patterns of pleiotropy, for example some QTL affect only one trait whereas others affect two or more traits (Knott and Haley, 2000). It is not possible to find a canonical transformation that ensures all QTL influence only one canonical trait. Consequently, it cannot be assumed that QTL found to be affecting two different canonical variables in the same location are actually different QTL (Weller et al. 1996). Meta-analysis of results from different studies (Allison and Heo, 1998) or joint analysis of the original data (Walling et al., 2000) have also been proposed to improve QTL mapping resolution.
The methods developed herein were motivated by the desire to maximise the power of QTL detection in the Adelaide-AgResearch Gene Mapping Project (Pitchford et al., 2003). The Project comprises many correlated traits recorded on progeny of 3 F1 sires born in two countries (total 6 sires). While care was taken to standardise protocols across countries, there were differences in markers genotyped, growth rate, slaughter age and sometimes subtle differences in measurement techniques. Thus, a method was required that would combine information across both correlated traits and half-sib families. Multiplicative models are increasingly being used in plant breeding multi-environment trials (Smith et al. 2001) and have been proposed for multi-trait animal analyses (Kirkpatrick and Meyer 2004). A multiplicative model approach is proposed herein.

**METHODS**

**Analysis methods.** Rather than testing an individual marker independent of all other markers, it is preferable to model all potential QTL at once. Gilmour’s (2007) method allows this and has the added advantage of the estimation of variance components for either each chromosome, or the whole genome. A single genome variance component is simpler than chromosomal variances, but could be very small because of inclusion of the large number of markers of zero or negligible effect.

The Adelaide-AgResearch Gene Mapping Project comprises 6 Jersey x Limousin F1 sires crossed back to both parent breeds to produce at least 125 progeny per sire (Pitchford et al., 2003). The progeny were genotyped for microsatellite markers that the sires were heterozygous for. The progeny genotypes are coded as -1 or 1 depending on which of the two alleles were inherited from the heterozygous sires. Markers are fitted as independent random effects and in the absence of any QTL represent error contrasts with expected variance component of zero. A significant variance component indicates that there is additional variation, attributed to QTL, and the best linear unbiased predictions (BLUPs) for the marker effects are then interpreted assuming the correlation between markers is $e^{-2\delta}$, where $\delta$ is the distance between markers in Morgans. Based on the magnitude and prediction error of the BLUP given the data, they can be converted to a probability of being zero (Verbyla et al. 2003). These probabilities were converted to log scale (-2ln(P)) which is equivalent to a LOD score for each marker and is distributed as a $\chi^2_2$ (Fisher 1954). Since the method thus far is a marker selection method, the most significant marker is added to the fixed effects to estimate size and significance. The model is then re-fit to identify whether significant genome variance remains and if so, the process of identifying most probable markers and re-fitting as fixed effects continues. Markers are selected conservatively compared to individual marker regression.

Multiplicative models have recently been popularised (Smith et al. 2001) in the analysis of plant variety trials to model genotype by environment interactions in the analysis of the data from multi-environment trials. The key aims of a multi-environment trial analysis are to provide accurate and precise estimates of overall variety performance and to aid with the interpretation and understanding of variety by environment interaction (Smith et al., 2001). As for plant analyses, in the multiplicative modelling of the marker by trait interaction herein, both common ($\xi$) and specific ($\delta$) factors are of interest. From a breeding point of view, common factors help select for desirable genetic correlations and specific factors help select against undesirable genetic correlations. The general factor model is linear in the common factors, $\xi$:
where \( y \) represents a multivariate observation, \( \mu \) a vector of means, \( \Lambda \) a matrix of factor loadings (\( \lambda_{ij} \) is the loading of the \( i \)th variable on the \( j \)th factor), \( \xi \) a vector of common factors, and \( \delta \) a vector of specific or residual factors. The vectors \( \xi \) and \( \delta \) are not observed and are generally assumed independent. With so many unobservable quantities, direct verification of the factor model from observations on \( y_1, y_2, \ldots, y_p \) is not possible. However, with some additional assumptions about the random vectors \( \xi \) and \( \delta \), the model implies covariance relationships, which can be checked.

Thompson et al. (2003) presented a sparse implementation of the average information algorithm for REML estimation of the factor analytic variance parameters which is computationally efficient as it exploits the regression underpinning the factor analytic model. Additionally, the common case of factor analytic variance structures with less than full rank (reduced rank variance models) has been accommodated in the algorithm, which is useful in the multivariate analysis. The algorithm has been implemented in ASReml (Gilmour et al., 2006) as the "XFA" variance model.

Multi-variate model fitting was conducted in the same way as univariate model fitting for selection for significant markers. Since the diagonal (DIAG) model with independent variances (equivalent to analysing all trait-sire combinations separately) and the factor analytic model with one factor (FA1) are nested models, the REMLRT statistic, \( -2\Delta \lambda \) (twice the log likelihood difference), can be approximated by the \( \chi^2 \) distribution with the degrees of freedom equal to the difference in the number of free parameters in the two nested models (Stuart et al., 1999). Significant markers were identified for the factor first. Once the factor variance was no longer significant, specific variances were tested and the most likely markers identified. This continued until the specific genome variances were not significant. In theory, more than a single factor could be fitted. However, this was not possible herein because, as outlined above, the whole-genome variances were small.

**Simulation study.** To investigate the power of the FA1 model, four normally distributed quantitative traits were simulated, each with a residual standard deviation of unity. The simulation was based on sample sizes of 125, 250, 500 and 750 reflecting the design of the Adelaide-AgResearch Mapping Project. There were 8 chromosomes with 6 markers for each chromosome, and an average marker distance of 20 cM and 1000 replications. Inheritance of all loci was determined assuming random assortment and that recombination events occurred independently, allowing use of Haldane’s (1919) mapping function. A total of 10 QTL were set (Table 1) for the whole genome, two having pleiotropic effects. Three sets of simulations for each population were generated, QTL with small effects, QTL with medium effects and QTL with large effects, in which each QTL (on average) explained 7, 10 and 13% respectively, of the phenotypic variance in the backcross. The variation explained by all the QTL for the four traits was the same. However, the traits had differentheritabilities because four QTL were simulated for traits 1 and 2 and three QTL for traits 3 and 4. Two QTL were set in repulsion phase on chromosome 6 affecting trait 2. Two QTL in coupling phase on chromosome 5 affected trait 1 and two QTL in coupling phase on chromosome 8 affected trait 4. One QTL at the centromeric end (first marker) of chromosome 2 and another at the telomeric end of chromosome 7 (last marker) both affected trait 2. The scenarios were designed to reflect those encountered in the Adelaide-AgResearch project.
Table 1 Simulated pleiotropic and trait specific QTL

<table>
<thead>
<tr>
<th>Trait</th>
<th>QTL1</th>
<th>QTL2</th>
<th>QTL3</th>
<th>QTL4</th>
<th>QTL5</th>
<th>QTL6</th>
<th>QTL7</th>
<th>QTL8</th>
<th>QTL9</th>
<th>QTL10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chr1</td>
<td>Chr2</td>
<td>Chr4</td>
<td>Chr5</td>
<td>Chr5</td>
<td>Chr6</td>
<td>Chr6</td>
<td>Chr7</td>
<td>Chr8</td>
<td>Chr8</td>
</tr>
<tr>
<td>Position (cM)</td>
<td>20</td>
<td>0</td>
<td>60</td>
<td>20</td>
<td>80</td>
<td>20</td>
<td>60</td>
<td>100</td>
<td>40</td>
<td>80</td>
</tr>
</tbody>
</table>

PLTC: Pleiotropic QTL. #: Trait specific QTL. COUP: Two linked QTL in coupling phase. REPL: Two linked QTL in repulsion phase

**Application to experimental data.** Early results from the Adelaide-AgResearch Project identified a major QTL on BTA2 which was segregating in a number of families and affected a number of carcass traits. A functional SNP in myostatin (*MSTN* or growth differentiation factor 8, *GDF8*) reported in Limousin cattle (Grobet et al. 1998) was genotyped and shown to be associated with many carcass traits (Sellick et al., 2007). However, it had still not been established that *MSTN* was indeed implicated. Thus, *MSTN* was used as a model to investigate the behaviour of the multivariate multiple QTL approach in the real dataset.

Three carcass traits (silverside weight (kg), eye muscle area (cm²) and channel fat (kg)) from six half-sib families were analysed. The model fitted to the data included fixed effects of trait means, country, country by breed interaction, sire within country, farm of origin within New Zealand, birth type within breed of dam, cohort within Australia (6 combinations of year of birth and sex), and slaughter group within New Zealand, (28 levels that includes adjustments for sex and year). The DIAG model included the random effect of marker by sire by trait effect with 18 variances (3 traits by 6 families). Also, the residual (co)variances were allowed to differ between the Limousin and the Jersey backcross animals as well as between countries. The subsequent FA1 model fitted to the trait by sire dimension of the trait by sire by marker interaction term required 18 factor loadings and 18 specific variances.

**RESULTS**

**Simulation study.** The DIAG model and the FA1 model were fitted across the four traits of the trait by marker effects in the simulation study. These involved four and eight parameters (4 factor loadings and 4 specific variances) respectively. As the QTL size of effect increased, the marker variances estimated from DIAG model for all four traits increased. The marker variances for traits 1 and 2 were always higher than those of traits 3 and 4, as expected because the heritability for these later traits was lower than that of the former. The impact of the two pleiotropic QTL was evident in the factor loadings with higher loadings on traits 1 and 3 than on traits 2 and 4 because QTL1 did not affect trait 2 and QTL3 did not affect trait 4.

Three specific QTL (QTL2, 6 and 7) were simulated for trait 2 and only one specific QTL was simulated for trait 3 (QTL8). This was reflected in the specific variances so that, in general, traits 2 and 3 had highest and lowest specific variances, respectively. In some replicates for small populations and small QTL size, the estimated marker variance using the DIAG model was on the
boundary (zero) for one trait. Also in some replicates, using the FA1 model, the specific variances were zero for one or two traits when the sample size and QTL effect were small.

A true positive was declared whenever the LOD score for a marker exceeded a pre-defined criterion (LOD=2) and a QTL was present at that marker. Conversely, a false positive was declared whenever the LOD score for a marker exceeded the criterion but no QTL was present. To avoid identifying adjacent markers as separate QTL, it was required that the LOD score drop by at least 1.0 between ‘peaks’ before declaring they represented separate QTL.

The power of the experiment was calculated as the average number of QTL detected divided by the number of QTL present. The results showed that the ability to detect QTL using both univariate (DIAG) and multivariate (FA1) analyses was strongly influenced by the QTL size of effect, and sample size so that the power to detect QTL improved significantly with increasing sample size and QTL effect (Figure 1). The overall power of detecting a QTL using the FA1 model was generally higher than that obtained in the univariate analysis. As expected, the increasing power using the FA1 model was more evident when two pleiotropic QTL were affecting the trait (Traits 1 and 3, Figure 1). For the large sample size and large QTL effect, the two methods had similar power to identify QTL.

The main feature to be noticed is the greater ability of the FA1 model compared to univariate analysis to detect QTL1 and QTL3 (not presented), which were simulated to have common effects on three of the traits. As expected, multivariate and univariate analyses were equally efficient in detecting trait specific QTL with large effects. However, trait specific QTL with small effect were detected with very low efficiency in both multivariate and univariate analyses (Figure 1). In the case of the probability for false QTL detection, both methods gave small likelihoods of finding false QTL. Again, as expected, the highest likelihood of detecting false QTL was for small sample size.

In the situation where the two linked QTL were in coupling phase, both methods sometimes choose the marker between two correct markers, particularly when the sample size and QTL effect was small. However, in the case of the two linked QTL in repulsion phase, the intervening marker was rarely selected.

Both univariate and multivariate techniques identified a low proportion of unlinked loci as QTL. This effect was not evident with large sample sizes. Both approaches seem quite conservative, delivering 0-4.7% (univariate) and 0-7.1% (FA model) of false positive unlinked loci.
Figure 1 Comparison of the power of univariate (Uni) and multivariate (FA1) for QTL detection. The power was defined as average number of QTL detected for each trait divided by the number of QTL present. Legend: Small QTL +, Medium QTL ▲, Large QTL ●; Univariate analysis solid lines and multivariate dashed lines.

Application to experimental data. Fitting the DIAG model, assuming no marker correlation between traits and families, gave a log-likelihood value of -3125.10, while fitting the FA1 model gave a log-likelihood value of -3096.66. The resulting REMLRT of 56.88 on 18 degrees of freedom indicates a highly significant (P<0.0001) improvement in the fit. The results from the DIAG model suggested a significant marker on BTA2 but it was not clear which marker was indicated. The FA1 model showed higher LOD scores for markers close to the myostatin gene (Figure 2). The highest LOD score was for the myostatin marker (labelled as MSTN).
DISCUSSION

A multi-trait multiple QTL approach in the framework of the mixed-effects model was developed for joint analysis of multiple traits and multiple families in gene mapping studies. The simulation scheme to evaluate the power of the test consisted of four QTL located on three chromosomes with no QTL on 6 other chromosomes (66 unlinked markers and 29 linked markers). Thus the majority of genetic markers across the genome were not linked to QTL. It is possible that allowing a common variance for all the markers leads to a low power for the REMLRT. However, the results from the simulation study showed that the REMLRT was remarkably robust.

In terms of the number of QTL correctly identified, the FA1 model performed much better than the univariate analysis for QTL of small effect and in small populations. However, it was only slightly better than univariate analysis for large QTL sizes and large populations. In addition to the issue of power, it is important to understand the nature of a genetic correlation between traits, which can provide relevant information for selection decisions. The superior performance of multivariate analysis was due largely to its ability to detect the QTL with common effect on different traits. In this regard, the key advantage of the FA1 over univariate analysis is that it provides a formal test for pleiotropic effects. This could also be used to formally distinguish between linked and pleiotropic QTL.

The method was tested in data from a beef cattle QTL mapping experiment where the MSTN gene was previously found to have pleiotropic effects on meat yield and fatness. The results clearly showed that fitting a FA1 model across traits and families gave a much better indication of the QTL position than single trait analysis or separate family analysis.
In contrast to previous joint analyses (e.g. Walling et al. 2000), this method is a one-stage process, which models residuals and genetic effects simultaneously. In addition, it includes all the markers in one analysis. Moreover, the approach utilises widely available statistical procedures, namely the linear mixed model and restricted maximum likelihood. It can easily accommodate covariates, extra sources of variation, fixed or random including polygenic effects and it can easily be generalised to experimental and crossing designs commonly used. The method should be able to be used with high density marker association studies as well as linkage studies.

REFERENCES
Haldane, J. B. S. 1919. J. Genet. 8: 299.