

MULTI-TRAIT QTL ANALYSIS BASED ON MIXED MODELS WITH PARSIMONIOUS COVARIANCE MATRICES

M. Malosetti^{1,2}, M.P. Boer^{1,3}, M.C.A.M. Bink³ and F.A. van Eeuwijk^{1,2}

¹ Laboratory of Plant Breeding, Wageningen University, P.O. Box 386,
6700 AJ, Wageningen, The Netherlands

² C.T. de Wit Graduate School for Production Ecology & Resource Conservation (PE&RC)

³ Biometris, 6700 AC Wageningen, The Netherlands

INTRODUCTION

In comparison to single trait QTL mapping approaches, multi-trait approaches have typically increased power to detect QTLs and allow investigation of the mechanism underlying genetic correlations between traits as pleiotropy versus genetic linkage (Jiang and Zeng, 1995; Knott and Haley, 2000). Recent examples in both the animal and plant breeding literature reflect the current interest in multi-trait approaches (Caliński *et al.*, 2000 ; Hackett *et al.*, 2001 ; Lund *et al.*, 2003; Turri *et al.*, 2004 ; Olsen *et al.*, 2005). Several methods have been proposed, including multivariate regressions (Knott and Haley, 2000), mixture models (Jiang and Zeng, 1995), and transformations of the multiple traits to canonical variables, like principal components (Mangin *et al.*, 1998). In this paper, we will look at a multi-trait approach within a mixed model framework, in which QTLs for multiple traits can be modelled as fixed or random in combination with a flexible structured variance-covariance matrix (VCOV) to model residual genetic variation and correlation due to polygenic effects.

Multivariate QTL mixed models have successfully been applied in animals (Lund *et al.*, 2003; Szyda *et al.*, 2003; Mercadé *et al.*, 2005), where the residual genetic VCOV between traits was modelled by an unstructured model, i.e., separate variances for individual traits and separate correlations for pairs of traits. Even with few traits, an unstructured VCOV might cause convergence problems and can be time consuming. To remedy convergence problems and excessive time demands in fitting, some structuring of the VCOV is necessary. Such structuring, however, should not reduce too much the flexibility of the VCOV to represent the residual genetic variation. An interesting option are factor analytic (FA) structures that were first proposed in a plant breeding context by Gogel *et al.* (1995), with further elaborations by Smith *et al.* (2001, 2005). Application of FA models in multi-environment QTL mapping for plants were given by Verbyla *et al.* (2003) and Malosetti *et al.* (2004). Modelling of multi-environment data is equivalent to modelling of multi-trait data when environments are interpreted as traits. We think that FA structures would also be interesting for animal breeding applications, but so far we did not find examples. We want to illustrate the use of FA models in multi-trait modelling by an example from plant breeding. We intend to show not only how multiple trait data can be analysed efficiently by using FA structures, but also how the additional complication of having multiple traits in multiple environments can be analysed within the same framework. For clarity we will talk about ‘characters’ when referring to the set of observed traits in each environment, while the term ‘trait’ will be used for combinations of characters and environments.

MODEL DESCRIPTION AND GENOME SCAN STRATEGY

A formulation for a multi-character multi-environment QTL model that can also serve as a multiple trait model is

$$Y_{i,(jm)} = \mu_m + E_{(jm)} + \sum_{q=1}^Q x_{iq} \alpha_{(jm),q} + e_{i,(jm)}$$

with $Y_{i,(jm)}$ the observation of genotype i ($i = 1 \dots I$), in environment j ($j = 1 \dots J$), for character m ($m = 1 \dots M$). The fixed terms in the model include a general intercept term for each of the characters (μ_m), an environmental main effect for each character-by-environment combination (i.e. each trait) ($E_{(jm)}$), and a term for the pleiotropic effects of q QTLs ($q = 1 \dots Q$) affecting the traits. The model term for the QTLs themselves consists of firstly the additive genetic predictors for QTL locus q in individual i , x_{iq} , representing conditional probabilities for QTL genotypes as inferred from marker information (Jiang and Zeng, 1997), and secondly the additive effect of QTL q in environment j for character m , $\alpha_{(jm),q}$. The random term, $\underline{e}_{i,(jm)}$, describes polygenic variation and plot error and is assumed to be multi-normally distributed with mean zero and VCOV Σ .

Stacking the observations by first genotype (slowest moving index), then environment, and finally character (fastest moving index), the VCOV for the data will be a block diagonal matrix $\Sigma = I_I \otimes \Sigma^*$ with I_I the identity matrix of dimension I (the number of genotypes) and Σ^* a symmetric matrix with $J \cdot M = N$ rows (=traits). Σ^* has on the diagonal the variances of individual traits and off-diagonally it has covariances between traits. With N traits, an unstructured VCOV model will require the estimation of $N \cdot (N+1)/2$ parameters. Alternatively, a good approximation to the unstructured model using considerably less parameters is given by the FA model of order K , where K is usually 1 or 2. The VCOV matrix becomes $\Sigma^{FA} = AA' + \Psi$, with A an $(N \times K)$ matrix of trait-specific multiplicative terms $\lambda_{(jm),k}$ ($k = 1 \dots K$) and Ψ a diagonal N matrix of trait-specific residuals ($\psi_{(jm)}$). With one multiplicative term ($K=1$), the diagonal elements of Σ^{FA} would be $\lambda_{(jm)}^2 + \psi_{(jm)}$, and the off-diagonal elements would be $\lambda_{(jm)} \lambda_{(jm)^*}$, for character-environment combinations (jm) and $(jm)^*$, with a total of $2 \cdot N$ parameters.

A QTL identification strategy can be as follows. First scan the genome with a model that includes one fixed pleiotropic QTL with separate effects for each character by environment combination ($Q=1$) and a FA model for residual genetic variation containing one multiplicative term ($K=1$). Fit the model by restricted maximum likelihood at intervals of 5 cM along the genome. Test the null hypothesis of no QTL by a Wald test for the fixed effects in a mixed model (Verbeke and Molenberghs, 2000). For decisions on significance use a false discovery rate procedure as described in Benjamini and Yekutieli (2005). To increase the power of the analysis, perform a second round of genome scanning, now including cofactors identified in the first round (Jansen and Stam, 1994). Finally, estimate QTL effects by fitting a model including all positions that showed significant indications for QTLs in the earlier genome scan.

Within the genome scan, model selection for the VCOV structure might have been included in various ways. Before fitting a QTL at a particular chromosome position, a best VCOV model could be chosen for a purely phenotypic model, i.e. a model without genetic predictors, by optimizing some criterion like the Bayesian Information Criterion (BIC) (Schwarz, 1978). The selected VCOV structure is then used to test for possible QTLs. After a genome scan, a multiple QTL model can be identified, starting from the collection of individually significant QTLs during the scan. Upon the determination of a final multiple QTL model, the VCOV model may be inspected for possible simplification. It is expected that after identification of QTLs, a less complex VCOV structure will be adequate for representing residual correlation between environments than before fitting QTLs. When indeed the VCOV model can be simplified, the tests for the fixed QTLs need to be adjusted. Our experience from analyses on real and simulated data is that FA models form a good a priori choice to perform genome scans, even when sometimes alternative VCOV models fit slightly better for the phenotypic model. Furthermore, after fitting QTLs, FA models still perform well in accounting for the

residual correlations, although often less complex VCOV models do slightly better in terms of BIC.

ILLUSTRATION

For a population of 150 doubled haploid barley lines we analysed the characters grain yield and flowering time across 10 environments in the US and Canada (Hayes *et al.*, 1993). Phenotypic and molecular data were obtained from <http://wheat.pw.usda.gov/ggpages/SxM/>.

A total of 12 QTLs were detected along the seven chromosomes. For the most important QTL, at chromosome 2, the effects are presented in Table 1. This QTL had a significant effect on yield and flowering time in most of the environments. The sign of the effect could differ between environments (gene-environment interaction), which was consistent with observed changes in correlation between yield and flowering time (Table 1). The full QTL model predicted very well the observed phenotypic correlations between yield and flowering time (Table 1).

Table 1. Estimated effects (\pm s.e.) of the QTL on chromosome 2(2H) for grain yield (YLD) and flowering time (FWT) across environments (significant effects in bold). The two last columns contain the correlations between YLD and FWT per environment as observed (Ph_Corr) and as predicted from the final QTL model (QTL_Corr).

	YLD	FWT	Ph_Corr	QTL_Corr
ENV_1	-0.166 \pm 0.086	2.7\pm0.17	-0.22	-0.22
ENV_2	-0.407\pm0.050	3.8\pm0.26	-0.41	-0.51
ENV_3	0.283\pm0.044	4.1\pm0.22	0.47	0.69
ENV_4	-0.327\pm0.038	2.7\pm0.20	-0.41	-0.46
ENV_5	-0.032 \pm 0.051	3.1\pm0.18	0.09	0.05
ENV_6	-0.240\pm0.051	2.9\pm0.20	-0.28	-0.41
ENV_7	-0.082 \pm 0.056	2.2\pm0.13	-0.05	-0.05
ENV_8	0.245\pm0.033	3.7\pm0.16	0.52	0.84
ENV_9	0.182\pm0.066	2.5\pm0.18	0.16	0.26
ENV_10	0.012 \pm 0.045	2.8\pm0.16	0.21	0.23

DISCUSSION

QTL detection will benefit from exploiting correlated information between traits. The FA model used here, offers a low-cost VCOV alternative in terms of parameters to unstructured models, while being flexible enough to model complex correlations between traits. For our example, we ran without major difficulties a genome-wide QTL scan for 20 traits, including several QTLs in the model. Attempts to do a genome scan using an unstructured VCOV model failed for our data. The use of simpler models for the VCOV (i.e. diagonal and constant correlation) produced considerably worse fits (BIC=6457 and BIC=5741, respectively) than a FA model (BIC=3535). Furthermore, by using simpler VCOV models we appeared to negatively affect the power of the analysis as we detected fewer QTLs. A flexible modelling framework by means of a FA VCOV model facilitates a further investigation of the causes of genetic correlations as being due to pleiotropy or linkage. Detailed results of the multi-character multi-environment analyses described in this paper will be published elsewhere, including analyses to distinguish pleiotropy from linkage as cause of genetic correlation.

REFERENCES

- Benjamini Y. and Yekutieli D. (2005) *Genetics* **171**: 783-790
 Caliński T., Kaczmarek Z., Krajewski P., Frova C. and Sari-Gorla M. (2000) *Heredity* **84**: 303-

- Hackett C.A., Meyer R.C. and Thomas W.T.B (2001) *Genet. Res.* **77**: 95-106
- Hayes P.M., Liu B.H., Knapp S.J., Chen F., Jones B., Blake T., Franckowiak J.D., Rasmuson D.C., Sorrells M., Ullrich S.E., Wesenberg D.M. and Kleinhofs A. (1993) *Theor. Appl. Genet.* **87**: 392-401
- Gogel B.J., Cullis B.R. and Verbyla A.P. (1995) *Biometrics* **51**: 744-749
- Jansen R.C. and Stam P. (1994) *Genetics* **136**: 1447-1455
- Jiang C. and Zeng Z-B (1995) *Genetics* **140**: 1111-1127
- Knott S.A. and Haley C.S. (2000) *Genetics* **156**: 899-911
- Lund M.S., Sorensen P., Guldbrandtsen B. and Sorensen D.A. (2003) *Genetics* **163**: 405-410
- Malosetti M., Voltas J., Romagosa I., Ullrich S.E. and van Eeuwijk F.A. (2004) *Euphytica* **137**: 139-145
- Mangin B., Thoquet P. and Grimsley N.H. (1998) *Biometrics* **54**: 88-99
- Mercadé A., Estellé J., Noguera J.L., Folch J.M., Varona L., Silió L., Sánchez A. and Pérez-Encizo M. (2005) *Mamm. Genome* **16**: 374-382
- Olsen H.G., Lien S., Gautier M., Nilsen H., Roseth A., Berg P.R., Sundsaasen K.K., Svendsen M. and Meuwissen T.H.E. (2005) *Genetics* **169**: 275-283
- Schwarz, G. (1978) *The Annals of Statistics* **6** : 461-464
- Smith A.B., Cullis B.R. and Thompson R. (2001) *Biometrics* **57**: 1138-1147
- Smith A.B., Cullis B.R. and Thompson R. (2005) *J. Agric. Sci. Cambr.*: **143**: 1-14
- Szyda J., Grindflek E., Liu Z. and Lien S. (2003) *Genet. Res.* **81**: 65-73
- Turri M.G., DeFries J., Henderson N. and Flint J. (2004) *Genome* **15**: 69-76
- Verbeke G. and Molenberghs G. (2000) *Linear mixed models for longitudinal data*. Springer.
- Verbyla A., Eckermann P.J., Thompson R. and Cullis B. (2003) *Austr. J. Agric. Res.* **54**: 1395-1408