Session 53, Abstract number 3355

Variance component method for QTL mapping in F_2 populations

Daisy Zimmer, Manfred Mayer, Norbert Reinsch

Research Institute for the Biology of Farm Animals (FBN), Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany;
zimmer@fbn-dummerstorf.de

Introduction

In animal science it is important to identify regions on a chromosome, which have a significant influence on a complex trait. These regions are called quantitative trait loci (QTL). A complex trait is typically influenced by multiple QTL with small additive and non-additive genetic effects relative to the phenotypic variance, but the exact number of affecting QTL is unknown. Non-additive effects are interactions within locus (dominance) or between two or more loci (epistasis). It is desirable to fit multiple QTL simultaneously using flanking marker information. QTL and marker have two alleles, where two alleles with the same origin are identical-by-descent (IBD). So far the variance component method (VCM) was often used in the field of QTL analysis. We modified the VCM from Xie (1998) for detecting multiple QTL in an F_2 design from a cross between two divergent parental inbred lines, i. e. they are alternatively homozygous on every locus. Generally, the VCM considers the QTL effects as random effects in a linear mixed model (LMM). In an F_2 population the observations $\mathbf{Y}$ can be modelled as follows when additive genetic, dominance and pairwise epistatic effects of the QTL are considered. A pair of QTL is indicated by ℓ and k. The LMM is

$$\mathbf{Y} \quad = \quad \mathbf{X}\boldsymbol{\beta} + \sum_{\ell=1}^{m} \mathbf{Z}_\ell(\mathbf{u}_{a_\ell} + \mathbf{u}_{d_\ell}) + \sum_{\ell=1}^{m-1} \sum_{k=\ell+1}^{m} \mathbf{Z}_{\ell k}(\mathbf{u}_{aa_{\ell k}} + \mathbf{u}_{ad_{\ell k}} + \mathbf{u}_{da_{\ell k}} + \mathbf{u}_{dd_{\ell k}}) + \mathbf{e}\,,$$

where m is the number of involved QTL. The model contains a vector of fixed effects $\boldsymbol{\beta}$ and $\mathbf{X}$ is the related design matrix. The vector $\mathbf{u}_s$ with $s \in \{a_\ell, d_\ell, aa_{\ell k}, ad_{\ell k}, da_{\ell k}, dd_{\ell k}\}$ denotes the additive genetic, dominance and the four pairwise epistatic effects. The design matrices $\mathbf{Z}_\ell$ with $\dim(\mathbf{Z}_\ell) = p \times n_\ell$ and $\mathbf{Z}_{\ell k}$ with $\dim(\mathbf{Z}_{\ell k}) = p \times n_{\ell k}$ assign the observations to the marker genotypes at the putative QTL. n_ℓ and $n_{\ell k}$ denote the number of effects to be estimated for the ℓ-th QTL and for a pair of QTL, respectively. The number of individuals with an observation is p. The residuals are independently and identically distributed with $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where $\mathbf{I}$ is the identity matrix and σ_e^2 is the residual variance component. The expectation of a QTL effect is $E(\mathbf{u}_s) = \mathbf{0}$ and the covariance matrix is $Var(\mathbf{u}_s) = \mathbf{V}_s\sigma_s^2$, where σ_s^2 is the related QTL variance component and $\mathbf{V}_s$ is the corresponding relationship matrix. The relationship matrices are determined from flanking marker information and they are also called marker genotype IBD matrices. We found an efficient way to set up the reduced relationship matrices with QTL genotype probabilities and fundamental QTL relationship matrices.

Theory of our VCM approach

The QTL genotypes can not be observed, but marker genotypes are known. The QTL genotype probabilities are conditional probabilities, which depend on the recombination rate between marker location and position of QTL. Taking the flanking markers of a QTL nine different marker genotypes can be distinguished and the QTL genotype probabilities can be derived similarly to Haley & Knott (1992). Furthermore, it is assumed that QTL and marker position are not identical and we assume that there is maximal one QTL in a marker interval. The QTL alleles are denoted by Q, q for the first QTL and H, h for the second QTL. The QTL allele Q is linked to marker allele 1 and q is linked to marker allele 2. All probabilities for the genotypes QQ, Qq and qq at the ℓ-th QTL conditional on flanking marker information M (e. g. $Pr(QQ|M)$) are written in the columns of a matrix $\mathbf{L}_\ell$ with $\dim(\mathbf{L}_\ell) = n_\ell \times 3$. Note that the genotypes Qq and qQ are considered as the same state. The individual QTL genotypes only depend on flanking

markers and they are conditionally independent. $\mathbf{L}_{\ell k}$ with $\dim(\mathbf{L}_{\ell k}) = n_{\ell k} \times 9$ contains the joint probabilities for both QTL genotypes conditional on marker genotypes of each individual. Due to a informative marker between both QTL, the joint probability is the product of both single probabilities, e. g. $\Pr(QQHH|M) = \Pr(QQ|M)\Pr(HH|M)$. Double recombination events are considered. Any of the numerous programs available to calculate these probabilities for F_2 line-cross experiments can be applied for this purpose.

As a second ingredient we need fundamental QTL relationship matrices, assuming that the recombination rate between marker and QTL is zero. In this case marker and QTL are identical and the QTL genotype can be clearly predicted from the marker genotype. The fundamental QTL relationship matrices are

$$\mathbf{A} = \begin{array}{c} \\ QQ \\ Qq \\ qq \end{array} \begin{array}{c} \begin{array}{ccc} QQ & Qq & qq \end{array} \\ \begin{pmatrix} 2 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix} \end{array} \quad \text{and} \quad \mathbf{D} = \begin{array}{c} \\ QQ \\ Qq \\ qq \end{array} \begin{array}{c} \begin{array}{ccc} QQ & Qq & qq \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array},$$

containing the additive and dominance relationship of the QTL for three possible genotypes QQ, Qq and qq. The matrix $\mathbf{A}$ was derived similar to Xie (1998). The diagonal elements of $\mathbf{A}$ are $1 + F_i$, where F_i is the inbreeding coefficient of the i-th animal. The matrix $\mathbf{D}$ follows the rules of Smith (1984) and Xie (1998). Both matrices, $\mathbf{A}$ and $\mathbf{D}$ consider the inbreeding coefficient correctly. We use the Kronecker product (symbol $\otimes$) of the fundamental matrices to compute the four different fundamental QTL interaction matrices, $\mathbf{A}\otimes\mathbf{A}, \mathbf{A}\otimes\mathbf{D}, \mathbf{D}\otimes\mathbf{A}, \mathbf{D}\otimes\mathbf{D}$, which include pairwise epistatic effects between two loci. The dimensions of the fundamental QTL interaction matrices are 9×9 and all possible QTL genotype combinations $(QQHH, QQHh, QQhh, QqHH, QqHh, Qqhh, qqHH, qqHh, qqhh)$ between both QTL are considered.

Now it is possible to set up the additive genetic, dominance and pairwise epistatic IBD matrices of F_2 individuals for a desired QTL position with QTL genotype probabilities and fundamental QTL relationship matrices. The IBD matrices will be positive definite at non-marker positions. If the QTL genotypes of two individuals i and j are known, then the relationship coefficient for all possible nine QTL genotype combinations may be directly taken from $\mathbf{A}$. If the recombination rates are different from zero, then we have to specify the QTL genotype probabilities for each individual. The relationship coefficient is a weighted average of the QTL genotype combinations (Xie, 1998). The additive genetic and dominance relationship matrices $\mathbf{V}_{a_\ell} = \left\{a^\ell_{ij}\right\}_{i,j}$ and $\mathbf{V}_{d_\ell} = \left\{d^\ell_{ij}\right\}_{i,j}$, which include one row and column for each individual $(i, j = 1, \ldots, p)$, are set up by $a^\ell_{ii} = (\mathbf{L}_\ell \mathrm{diag}(\mathbf{A}))_i$ and $a^\ell_{ij} = (\mathbf{L}_\ell \mathbf{A}\mathbf{L}'_\ell)_{ij}$ as well as $d^\ell_{ii} = (\mathbf{L}_\ell \mathrm{diag}(\mathbf{D}))_i$ and $d^\ell_{ij} = (\mathbf{L}_\ell \mathbf{D}\mathbf{L}'_\ell)_{ij}$, respectively, at the ℓ-th QTL $(\ell = 1, \ldots, m)$.

The basic idea considers genetic effects for each individual, i. e. n_ℓ and $n_{\ell k}$ are equal to the number of phenotyped individuals $(n_\ell = n_{\ell k} = p)$. The estimation of the conditional genotypic effects for all animals with these IBD matrices becomes computationally slow or even infeasible, because the number of mixed model equations (MME) increases with the number of F_2 individuals. It makes sense to reduce the number of effects to be estimated. This is also useful, when the marker location and QTL are identical or close together, because then there are (nearly) linear dependencies in the IBD matrices and convergence problems may appear. In our reduced approach the F_2 individuals are grouped depending on their marker genotypes and an average conditional genotypic effect is estimated. The number of MME is reduced and so we need considerably less computing time. We call this procedure the short VCM (versus the long VCM described above). This new reduced method is supposed to be flexible and easy to compute. The diagonal and off-diagonal elements of the reduced relationship matrix $\mathbf{V}_{a_\ell}$ at the ℓ-th QTL are calculated also in different steps by

$$a^\ell_{ii} = \begin{cases} (\mathbf{L}_\ell \mathrm{diag}(\mathbf{A}))_i & \text{when} \quad n^\ell_i = 0, \\ \frac{1}{n^\ell_i}\left((\mathbf{L}_\ell \mathrm{diag}(\mathbf{A}))_i + (n^\ell_i - 1)(\mathbf{L}_\ell \mathbf{A}\mathbf{L}'_\ell)_{ij}\right) & \text{else}, \end{cases}$$

$$a^\ell_{ij} = (\mathbf{L}_\ell \mathbf{A}\mathbf{L}'_\ell)_{ij},$$

with $i, j = 1, \ldots, n_\ell$ and $i \neq j$. We consider the exact number n_i^ℓ of observed levels for each marker genotype. The vector of diagonal elements of $\mathbf{A}$ is denoted by $\mathrm{diag}(\mathbf{A})$. $\mathbf{L}_\ell$ is adjusted, meaning the number of marker genotype effects is n_l and it is $p = \sum_{i=1}^{n_\ell} n_i^\ell$. The calculation of $\mathbf{V}_{d_\ell}$ at the ℓ-th QTL is done similarly to the notes above, but $\mathbf{A}$ has to be substituted by $\mathbf{D}$. The relationship coefficient a_{ij}^ℓ (d_{ij}^ℓ) represent the conditional probability of (pairwise) IBD alleles at the ℓ-th QTL between individuals i and j with respect to marker genotypes. For $\mathbf{V}_{a_\ell}$ and $\mathbf{V}_{d_\ell}$ $n_\ell = 9$ given that the markers are fully informative and we are going to estimate an conditional genotypic effect for each marker genotype i. The epistatic relationship matrices $\mathbf{V}_{aa_{\ell k}}, \mathbf{V}_{ad_{\ell k}}, \mathbf{V}_{da_{\ell k}}, \mathbf{V}_{dd_{\ell k}}$ at the ℓ-th and k-th QTL are computed analogously to $\mathbf{V}_{a_\ell}$, but we use the corresponding Kronecker product instead of $\mathbf{A}$. In total we estimate $n_{\ell k}$ levels of the epistatic effects. Using the reduced epistatic relationship matrices and fully informative markers, then $n_{\ell k} = 27$, if the QTL are in two adjacent marker intervals and $n_{\ell k} = 81$ otherwise.

Results and Discussion

The properties of VCM to detect multiple QTL are compared by simulations. We studied the QTL mapping for an F_2 population from two divergent breeds with large phenotypic differences. The progeny size (F_2) was 200 for each of the 1000 replicates. Six markers were evenly spaced on the chromosome at 0/10/20/30/40/50 cM. Two QTL linked to the markers were simulated at 25 and 35 cM and they were in repulsion. Under these assumptions we performed a QTL scan on the chromosome. Only the F_2 individuals were phenotyped. The phenotypic value included only a population mean, additive genetic effects for each QTL ($a_1 = -a_2 = 1$) and a residual effect. The genetic effects were simulated using Cockerham's model (Kao & Zeng, 2002) and the residual variance component was $\sigma_e^2 = 0.181$. To estimate the positions of the QTL we solved the MME and received the EBLUP and EBLUE estimates with use of the software ASReml (Gilmour et al., 2006). The detection of QTL was achieved by a residual likelihood ratio test. The test statistic was $LR = 2\,(\log L_{H_A} - \log L_{H_0})$, where $\log L_{H_A}$ and $\log L_{H_0}$ were the logarithmic residual likelihood functions under H_A and H_0, respectively. The model under H_A coincided with the proposed LMM. Under H_0 it was assumed that $\sigma_s^2 = 0$. A QTL was supposed on that position, where LR was maximal and greater than a previous determined threshold value. The significance threshold for the short and long VCM were here 4.34 and 5.75, respectively. The criteria of comparison were the accuracy of the estimated mean, empirical standard error (sd), mean squared error (mse) as well as the 5 % and 95 % quantiles of the estimated QTL positions. The contour plot of the average likelihood profile between the reduced and animal model is shown in the Figure. It is obvious that both methods identified nearly the same QTL positions, but the short VCM prossessed smaller maximal likelihood ratio ($\max LR = 44.524$) than the long VCM ($\max LR = 60.621$). However, the short VCM needed considerably less computing time than the long VCM. The short VCM was also good for QTL detection (see Table), but the mse of the estimated position of the QTL was higher than for the animal model.

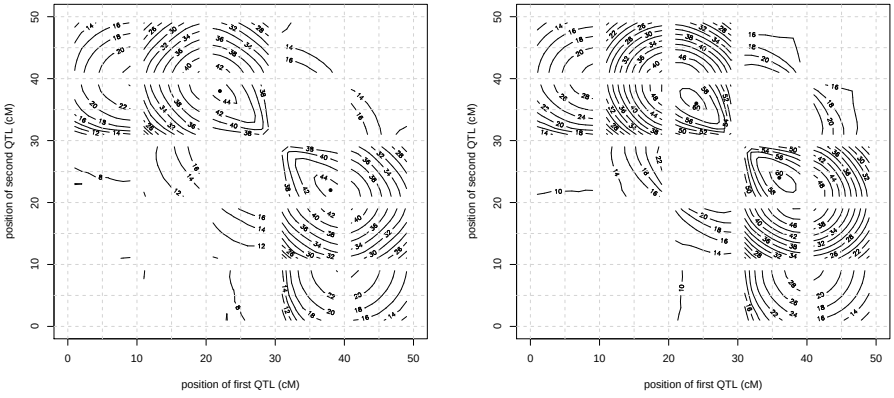

Figure: Contour plot of the average likelihood profile for the reduced model (left) with $\max$ LR = 44.524 and the animal model (right) with $\max$ LR = 60.621. QTL identified in reduced model at 22/38 cM and in animal model at 24/36 cM.

Table: Average estimated positions of QTL (P1, P2) between the reduced and animal model for 1000 replications, each with 200 individuals. Criteria for comparison were the mean, standard error (sd), mean squared error (mse), 5 % and 95 % quantile of the estimated QTL position. Further interesting parameter was the estimates of the residual variance component. The QTL were simulated at 25/35 cM and markers were set at 0/10/20/30/40/50 cM. The residual variance was $\sigma_e^2 = 0.181$.

	reduced model		animal model	
	P1	P2	P1	P2
mean	21.67	38.38	23.70	36.37
sd	2.56	2.44	1.56	1.49
mse	17.64	17.40	4.12	4.10
5% quantile	17.00	35.00	22.00	34.00
95% quantile	25.00	42.00	26.00	38.00
σ_e^2	0.27		0.18	

The estimated residual variance of VCM tends to be slightly overestimated, because the variance of the residual deviation includes the original residual variance plus the variance between the QTL genotype combinations. The reduced relationship matrices depend on the exact number of observations per marker genotype, which are used as weights. Also the size of relationship matrices for the short VCM is independent of the number of F_2 individuals which is a great advantage of this approximation. If there are less observed levels of marker genotypes, then we compute the corresponding row in the IBD matrix with expected probabilities for each QTL genotype. Therefore, it is a useful method to locate the regions in the genome with the strongest support for a QTL.

The short VCM is a flexible approach to map multiple linked QTL regarding multiple marker intervals. A major advantage of the presented method is that it is easy to implement and it is a useful genome scan method for detecting QTL which is not so computationally demanding like the original approach from Xie (1998). The proposed method can be currently used for F_2 populations from inbred parental populations or lines which are homozygous at the QTL. The short VCM is computationally fast and numerically stable. The advantage of the short VCM is high with increased marker density and progeny size. Therefore, this method is expected to have a practical importance for future QTL analyses. Due to diminished accuracy for QTL mapping with the reduced model and the computationally demanding of the animal model it may be desirable to use a combination of animal and reduced model. Therefore, a rough QTL mapping can done by short VCM and after this procedure we can use the long VCM to accurate mapping of QTL and estimating parameters.

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, UK*.

Haley, C. S. and Knott, S. A. (1992): A simple regression method for mapping quantitative trait loci in line crosses using anking markers. *Heredity 69, 315–324*.

Kao, C.-H. and Zeng, Z.-B. (2002): Modeling epistasis of quantitative trait loci using Cockerham's model. *Genetics 160, 1243–1261*.

Smith, S. P. (1984): Dominance relationship matrix and inverse for an inbred population. *Columbus: Mimeo, Dep. Dairy Sci., The Ohio State Univ.*

Xie, C., Gessler D. D., Xu S. (1998): Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. *Genetics 149, 1139–1146*