A new method to fine mapping quantitative trait locus using linkage disequilibrium.

H. Gilbert1,2, M.Z. Firat2,3, L.R. Totir2, J.C.M. Dekkers2, R.L. Fernando2

1Station de Génétique Quantitative et Appliquée, Institut National de la Recherche Agronomique, 78352 Jouy-en-Josas cedex, France

2Department of Animal Science, Iowa State University, Ames, Iowa 50011, USA

3Akdeniz University, Faculty of Agriculture, Department of Animal Science, 07059 Antalya, Turkey

Abstract A new approach was developed to fine map a biallelic QTL using linkage disequilibrium , relating means and covariances of the QTL gametic values to the QTL allele effect and their frequencies among the founders' marker haplotypes. This reduces the number of parameters compared to usual models. MCMC was used to derive the conditional probabilities of inheriting maternal and paternal QTL alleles. A residual maximum likelihood method was implemented to map the QTL, using a Newton-Raphson algorithm to jointly estimate QTL position and effect of the QTL, QTL allele frequencies of the founders' marker haplotypes carrying the mutant QTL allele, and polygenic and residual variance components for each interval. Simulated populations were analyzed to compare its ability to fine map a QTL to two others methods: one based on identity by descent QTL covariances, and the other modeling independently the QTL effect means and covariances of the QTL gametic values.

Introduction Combination of molecular tools with phenotypic data to better estimate genotypic values for selection represents a recent challenge in quantitative genetics. For so-called marker assisted selection, methodology has been developed (Abdel-Azim and Freeman, 2001; Fernando and Grossman, 1989; Fernando and Totir, 2002; Goddard, 1992; van Arendonk *et al.*, 1994; ăWang *et al.*, 1994; ăWang *et al.*, 1998 to exploit the information on a locus having an effect on a quantitative trait ă(QTL). Most of those strategies used only mendelian cosegregation between alleles at a marker locus linked to a QTL (marked QTL: MQTL) ăalleles, at the family level (often called linkage equilibrium, LE). This implies modeling the ăcovariances between the genetic values due to the QTL. But to combine this information with preferential ăassociations between alleles at the population level (linkage disequilibrium, LD), one should also model ăthe expectation of those values. There, authors usually suggested an independent model of each of those ăelements (see Meuwissen and Goddard, 2001; Fernando and Totir 2003). For this contribution we developed a new strategy based on Fernando and Grossman (1989) genetic model ăto tie them together, to improve the forthcoming parameters estimated for selection.

1 Integrating genotypic information in BLUP evaluation

Selection from a total population of n, based on information $\mathbf{D}$ is usually performed estimating their individual genotypic values g_i (i=1,n). The conditional mean of the k selected is maximized, selecting the candidates with the largest estimates $\hat{g}_i$=E(g_i/$\mathbf{D}$) (Bulmer, 1980; Fernando and Gianola, 1986; Henderson, 1984).

BLUP EVALUATION: Traditionally, $\mathbf{D}$ consists of pedigree relationships and trait phenotypes $\mathbf{P}$. Multivariate normal distributions are assumed for a vector $\mathbf{y}$ of trait phenotypes and a vector $\mathbf{g}$ of unobservable genotypic values. Thus, the conditional mean of $\mathbf{g}$ is a linear function of $\mathbf{y}$:

$$E(\mathbf{a}/\mathbf{D}) = \mu_{\mathbf{g}} + \mathbf{C}\mathbf{V}^{-1}(\mathbf{y} - \mu_{\mathbf{y}}), \quad (1)$$

ăwhere $\mu_{\mathbf{g}}$ and $\mu_{\mathbf{y}}$ are the expected values of $\mathbf{g}$ and $\mathbf{y}$ conditionally on pedigree ăinformation, $\mathbf{C}$ is the covariance matrix between $\mathbf{g}$ and $\mathbf{y}$ and $\mathbf{V}$ is the covariance matrix of ă$\mathbf{y}$. Regardless the joint distribution of $\mathbf{g}$ and $\mathbf{y}$, this linear function of $\mathbf{y}$ gives the best linear predictor of $\mathbf{g}$ (Henderson, 1984).

ăIf $\mathbf{y}$ can be modeled as $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{g}^* + \mathbf{e}$, where $\mathbf{X}$ and ă$\mathbf{Z}$ are known incidence matrices, β is an unknown vector of fixed effects, $\mathbf{g}^* = \mathbf{g} - E(\mathbf{g})$ and $\mathbf{e}$ is a vector of residuals with null means and covariance matrix $\mathbf{R}$, then ă$\mu_{\mathbf{y}} = \mathbf{X}\beta$, $\mathbf{C} = \mathbf{A}\mathbf{Z}'$ and $\mathbf{V} = \mathbf{Z}'\mathbf{A}\mathbf{Z} + \mathbf{R}$, where ăVar($\mathbf{g}/\mathbf{P}$)=$\mathbf{A}$. An efficient computation of $\mathbf{A}^{-1}$ is known from Henderson (1976). When the μ_{g} and $\mu_{\mathbf{y}}$ are unknown and replaced by their generalized least squares estimates, it ăgives the best linear unbiased predictor (BLUP) for $\mathbf{g}$ (Henderson, 1984). ă

BLUP USING TRAIT AND MARKER DATA: ă$\mathbf{D}$ can also contain marker data $\mathbf{M}$ related to a marker closely linked to a quantitative trait locus (MQTL) which ăcontributes to the genetic variability of the trait. Under assumption of multivariate variability, equation ă1 is still a valid expression for the conditional mean of $\mathbf{g}$. μ_{g}, $\mu_{\mathbf{y}}$, ă$\mathbf{C}$ and $\mathbf{V}$ are modified to account for this new information.

ăTwo cases are to be considered when dealing with marker data. The marker locus is in linkage equilibrium (LE) with the QTL if ăthe alleles for these two loci are independently distributed. Even if the two loci are physically ălinked, knowing the marker genotype for a randomly sampled individual does not provide information concerning its QTL ăallele. However, if two relatives receive the same marker allele from a common ancestor the likelihood they received ăthe same MQTL allele can be calculated. Those genotypes for a linked marker are thus used to model the genetic ăcovariances between relatives, but μ_{g} and $\mu_{\mathbf{y}}$ are not modified by this information.

ă ăThe loci are in linkage disequilibrium (LD) if information on the marker genotypes provides information on the QTL genotypes. ăTypically, one of the marker allele will be preferentially associated with one particular MQTL allele. In this case, ăE($\mathbf{g}/\mathbf{P}$) $\neq$ E($\mathbf{g}/\mathbf{P}$,$\mathbf{M}$). The marker genotypes are then ăused to model covariances between relatives, but also the expected genetic values.

1.1 Model under equilibrium (co-segregation)

In order to use the Henderson mixed model equations (HMME) for marker assisted BLUP, Fernando and Grossman (1989) modeled $g_i = v_i^m + v_i^p + u_i$, where v_i^m (v_i^p) are the MQTL additive effects of the maternal (paternal) alleles, and u_i is the additive effect of the remaining trait loci. $\mathbf{y}$ is then modeled as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{W}\mathbf{v} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{2}$$

ăwhere $\mathbf{v}$ is the vector of gametic effects at the QTL and $\mathbf{W}$ a known incidence matrix relating the ăgametic effects to trait values. An efficient tabular method for the computation of the inverse ăof the conditional ăcovariance matrix $\mathbf{\Sigma}_{\mathbf{v}}^{-1}$ of $\mathbf{v}$ has been described by Fernando and Grossman (1989) and Wang *et al.* ă(1995). The efficient computation of the inverse of $\mathbf{\Sigma}_{\mathbf{u}}$, conditional covariance matrix of $\mathbf{u}$ and ăproportional to $\mathbf{A}$, is known from the inverse of $\mathbf{A}$. The HMME can then be solved to obtain estimates for ăβ, $\mathbf{v}$ and $\mathbf{u}$. ăTo estimate the variance components, we compute the likelihood of the $\mathbf{y}$, using an automatic ădifferentiation strategy to obtain the first and second derivatives. A Newton-Raphson ăalgorithm maximizes the likelihood.

1.2 Combining with linkage disequilibrium

When modeling LD, we divide the MQTL effect between a genetic fixed effect α and the random effect already described. Equation 2 is then changed in

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{X}_{\mathrm{g}}\alpha + \mathbf{W}\mathbf{v} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{3}$$

ăwith E($\mathbf{g}/\mathbf{M}$) = $\mathbf{X}_{\mathrm{g}}\alpha$. $\mathbf{X}_{\mathrm{g}}$ is a matrix with one column of probabilities. ă ăThis communication describes a new model to efficiently compute the expected values and covariances ăof gametic effects. In most of the previous works, each expected value is estimated ăindependently from the corresponding variance components. Our model ties the components together and with the expected ăvalues. Parameters are then meaningful and can possibly be used directly to select individuals based on the marker ăalleles they carry. ă

1.3 A new model for LD

We assume a biallelic MQTL with alleles Q1 and Q2 (mutant allele). a, the MQTL effect, corresponds to half the difference between expected trait performances of homozygous individuals: $2a = \mu_{Q2Q2} - \mu_{Q1Q1}$. For individual i, lets define $p_i^x = \Pr(Q_i^x = \mathrm{Q2}/\mathbf{P}, \mathbf{M})$, $x = m, p$, the probability it inherited the mutant allele from its dam ($x = m$) or its sire ($x = p$) conditionally on the information. It is then easy to write as $\mathrm{E}(g_i/\mathbf{P}, \mathbf{M}) = a * (p_i^m + p_i^p)$ the **conditional expectations of the gametic values**, so that $\mathbf{X}_g$ is composed of elements $(p_i^m + p_i^p)$ for each individual, and $\alpha = a$. **The conditional variances of the gametic effects** are written as $\mathrm{Var}(v_i^x/\mathbf{P}, \mathbf{M}) = a^2 * p_i^x * (1 - p_i^x)$, $x = m, p$. Concerning the **covariances** $\mathrm{Cov}(v_i^m, v_j^x/\mathbf{P}, \mathbf{M})$, if j is not a descendant of i, we propose a tabular method similar to the method used in Fernando and Grossman (1989) to compute $\mathbf{\Sigma_v}$:

$$Cov(v_i^m, v_j^x/\mathbf{P}, \mathbf{M}) = Pr(Q_i^m < -Q_d^m/\mathbf{P}, \mathbf{M}) * Cov(v_d^m, v_j^x/\mathbf{P}, \mathbf{M}) + Pr(Q_i^m < -Q_d^p/\mathbf{P}, \mathbf{M}) * Cov(v_d^p, v_j^x/\mathbf{P}, \mathbf{M}), \tag{4}$$

where v_d^x are the gametic effect for the dam d of i. Similarly, v_s^x, the gametic effect for the sire s of i, is used to compute $\mathrm{Cov}(v_i^s, v_j^x/\mathbf{P}, \mathbf{M})$. $Q_i^m < -Q_d^m$ represents the MQTL allele inherited from the maternal haplotype of the dam. $\Pr(\mathrm{Q}_i^{x1} <\text{-}\ \mathrm{Q}_d^{x2}/\mathbf{P}, \mathbf{M})$, $x1 = m, p$, $x2 = m, p$, were called PDQ in Wang *et al.* (1995).

To **compute probabilities** p_i^x, if j is not a descendant of i, we propose a similar tabular method, using the PDQ:

- if i is not a founder

$$p_i^m = Pr(Q_i^m < -Q_d^m/\mathbf{P}, \mathbf{M}) * p_d^m + Pr(Q_i^m < -Q_d^p/\mathbf{P}, \mathbf{M}) * p_d^p, \tag{5}$$

- if i is a founder, we arbitrarily label 1 and 2 each of the QTL alleles (the maternal or paternal origins are now meaningless): for $x = 1, 2$,

$$p_i^x = \sum_h Pr(H_i^x = H_h) * \pi_h, \tag{6}$$

where H_i^x is the x haplotype of i, H_h is the h^{th} haplotype available in the founder population, and π_h is the frequency of the mutant allele in H_h, $\Pr(Q_i^x = Q2/H_i^x = H_h)$. π_h are parameters to estimate, which will explicitly represent the linkage disequilibrium between the alleles in the population.

1.4 Practical implementation

The software was developed in C++, with an intensive access to the opportunities available through the matvec library developed for scientific programming (Kachman and Fernando, 2002). It was used to solve the HMME for the estimation of fixed and random effects. Automatic differentiation techniques were applied to get the first and second derivatives of the likelihood, so that the Newton-Raphson algorithm can be used to get variance components estimates.

Our practical implementation requires intensive use of Monte Carlo Markov Chain (MCMC) strategies we can not extensively described in this paper, for the estimation of the PDQs and $\Pr(\mathrm{H}_i^x = \mathrm{H}_h)$ (Pita *et al.*, 2004). An efficient strategy was implemented using algorithms such as the descent graph (Sobel and Lange, 1996), to get the segregation indicators probabilities at each marker position before beginning the mapping process. It makes it possible to compute the required probabilities for each position considered on the linkage group during the mapping process with no additional MCMC computation.

1.5 Discussing the model

With our strategy, we reduce the number of parameters to estimate to a and the π_h, and tie together the estimations of expected values and variance components for the gametic effects, which is supposed to be easier to interpret. Compared to the strategies based on Meuwissen and Goddard (2001) computation of the IBD matrix, where the covariance matrix combines LE and LD, our model disentangle them, easing the interpretation and uses. Finally, the π_h can be used as criteria for selection of individuals, being indicators of the QTL state of the individuals.

The major assumption concerns the biallelic state of the QTL. This might not be such a constraint if we consider this hypothesis as "the mutant allele of interest *versus* the pool of the other alleles". How this is pertinent when the alleles in the "pool of the other alleles" have different effects on the trait should be tested.

2 Simulations and first results

First tests of this method were carried out on simulated data. Its mapping accuracy was compared to methods based on Meuwissen and Goddard (2001) calculation of IBD matrix. In this preliminary work, the segregation indicators are considered as known from the simulations.

2.1 Simulated designs

The population came from 100 generations of random mating, with effective size of 100. In the 100th generation, 5 sires and 25 dams were picked up at random to create a working population. Five new generations of 100 were simulated, each from mating of 5 sires and 25 dams. Marker genotypes and trait values were known for the all 1030 individuals. On a 9 cM linkage group, we simulated 10 biallelic genetic markers evenly spaced (intervals between genetic markers are numbered from 0 to 8) The simulated QTL was in the middle of the linkage group (4th interval) or in the middle of the 2nd interval between two markers (position 2.5 cM).

The phenotypic trait variance was 100, with two designs: 1) big QTL effect, equal variances from the QTL (σ^2_Q) and the polygenic part (σ^2_u): $\sigma^2_Q = \sigma^2_u = 30$, $\sigma^2_\epsilon = 40$; 2) small QTL effect: $\sigma^2_Q = 2.5$, $\sigma^2_u = 22.5$, $\sigma^2_\epsilon = 75$.

2.2 Preliminary results and conclusions

Table I presents three criteria for the estimation of the mapping accuracy, based on the interval where the maximum likelihood was found: the averaged, the standard deviation and the mean absolute difference with the true parameter. They were estimated over 200 to 500 simulations depending on the cases.

Table I: Location of the maximum likelihood: averaged interval, standard deviation and mean absolute difference with the true value. σ^2_Q = simulated QTL variance, Interval = interval truly simulated

ă Interval	σ^2_Q	averaged interval	standard deviation	mean absolute difference
4	30.0	3.97	1.11	0.64
4	2.5	3.65	2.11	1.69
2	30.0	2.06	1.15	0.75

These first results show good accuracy, with a 1 cM error in average, which locates the QTL in the true interval in more than 80% for the more accurate case. Using the option based on Meuwissen and Goddard computation of IBD matrix, we obtained a mean absolute difference of 0.4 for the first case described in

this table, which is better. A further look at the results showed troubles in the likelihood maximization: in at least 40% of the simulations, the true maximum was not found, giving bad estimates for the parameters. The mostly affected parameters were the effect and the πs, the variances being consistent and accurately estimated.

Two different steps are considered to overcome these troubles: choose a new, more efficient, algorithm to get good likelihood maxima, and add a step in the model for the means to avoid the direct estimation of a and the πs, and eventually make them easier to estimate. The first step begins to show interesting resulting, with strategies based on Fisher information matrix. Once fixed, we plan comparisons of the models to define the optimal conditions (haplotype sizes and population structures) when LD strategies should be implemented. In an additional computational step, techniques to deal with multiple traits and missing data will be made available.

Acknowledgement: We acknowledge Monsanto and Sygen, which grant part of this work, and INRA and ISU.

Bibliography

Abdel-Azim G. and Freeman A.E. (2001). A rapid method for computing the inverse of the gametic covariance matrix between relatives for a marked quantitative trait locus, *Genetics Selection Evolution*, 33: 153-173.

Bulmer M.G. (1980). *The Mathematical Theory of Quantitative Genetics*, Clarendon Press, Oxford.

Fernando R.L. and Gianola D. (1986). Optimal properties of the conditional mean as a selection criterion, *Theoretical and Applied Genetics*, 72: 822-825.

Fernando R.L. and Grossman M. (1989). Marker assisted selection using best linear unbiased prediction, *Genetics Selection Evolution*, 21: 467-477.

Fernando R.L. and Totir L.R. (2003). Incorporating molecular information in breeding programs: methodology, In Muir W.M. and Aggrey S.E., editors, *Poultry Breeding and Biotechnology*, CABI Publishing, Cambridge.

Fernando R.L. and Totir L.R. (2002). Advances in genetics and statistical models to predict breeding values, *7th World Congress on Genetics Applied to Livestock Production*, august 19-23, 2002, Montpellier, France, CD-ROM Communication N 20-01.

Goddard M. (1992). A mixed model for analysis of data on multiple genetic markers, *Theoretical and Applied Genetics*, 83: 878-886.

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

Kachman S.D., Fernando R.L. (2002). Analysis of generalized linear mixed models with matvec, *7th World Congress on Genetics Applied to Livestock Production*, august 19-23, 2002, Montpellier, France, CD-ROM Communication N 28-04.

Meuwissen T.H.E. and Goddard M.E. (2001). The use of marker haplotypes in animal breeding schemes, *Genetics Selection Evolution*, 28: 161-176.

Pita F.V.C., Fernando R.L., Totir L.R., Lopes P. (2004). An improved approximation of the gametic covariance matrix for marker assisted genetic evaluation by BLUP, *Genetics Selection Evolution*, submitted.

Sobel E. and Lange K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics, *American Journal of Human Genetics*, 58: 1323-1337.

van Arendonk J.A.M., Tier B., Kinghorn B.P. (1994). Use of multiple genetic markers in prediction of breeding values, *Genetics*, 137: 319-329.

Wang T., Fernando R.L., Grossman M. (1998). Genetic evaluation by BLUP using marker and trait information in a multibreed population. *Genetics*, 148: 507-515.

Wang T., Fernando R.L., van der Beek S., Grossman M., van Arendonk J.A.M. (1995). Covariance between relatives for a marked quantitative trait locus, *Genetics Selection Evolution*, 27: 251-274.