Incorporating molecular markers into genetic evaluation

Session G6.1

R. L. Fernando
Iowa State University,
Department of Animal Science,
225 Kildee Hall,
Ames 50011-3150
rohan@iastate.edu

ABSTRACT

Molecular markers can be broadly classified into two types: I) those that have a direct effect on a trait, and II) those that do not have a direct effect on any trait but are linked to a trait locus. A marker of type I can be incorporated into genetic evaluation by including it as a fixed effect in the model used for genetic evaluation. Even in this ideal situation, genetic evaluation may not be straightforward when marker genotypes are missing on a significant proportion of the pedigree. Markers of type II can be further classified into two types: IIa) markers that are in linkage disequilibrium with the trait locus, and IIb) markers that are in linkage equilibrium with the trait locus. When disequilibrium is strong, a marker of type IIa can be treated as a type I marker. If disequilibrium between the marker and the trait locus is weak, the marker will have little effect on the trait mean. When the marker locus and trait locus are in linkage equilibrium, the allele states at the two loci are independent, and thus, the maker has no effect on the trait means. But, even in this situation, marker information can be used to model trait covariances by treating marker within animal as a random effect. Strategies for including these types of markers in genetic evaluation will be discussed

1 Introduction

Due to advances in molecular genetics, increasing amounts of molecular information are becoming available for genetic evaluation. The molecular information considered here consists of molecular genotypes at polymorphic loci. These loci can be broadly classified into two types: I) those that have a direct effect on a trait, and II) those that do not have a direct effect on any trait but are linked to a trait locus. Loci of type II can be further classified into two types: IIa) loci that are in linkage disequilibrium with the trait locus, and IIb) loci that are in linkage equilibrium with the trait locus. The focus of this paper is on the principles underlying mixed linear model methods

that incorporate these types of loci into genetic evaluation. It can be shown that two types of linkage information can contribute to genetic evaluation or linkage analysis. Methodology for genetic evaluation that combines these two types of information is presented here. The application of this methodology to incorporate loci of types I, IIa and IIb into genetic evaluation is discussed. For simplicity, theory is presented for incorporating genotypes at a single locus into genetic evaluation.

2 Two Types of Linkage Information

2.1 Disequilibrium Information

Consider a marker locus A with alleles A_1 and A_2 that is linked to a trait locus Q with alleles Q_1 and Q_2. Let S_A be a variable that specifies if the allele at the A locus on some haplotype is A_1 or A_2. Similarly, let S_Q be a variable that specifies if the allele at the Q locus on this haplotype is Q_1 or Q_2. We will refer to S_A and S_Q as allele state indicators. Now, for a randomly sampled haplotype, if the allele state indicators S_A and S_Q are independent, then A and Q loci are said to be in gametic phase equilibrium (or linkage equilibrium); on the other hand, if S_A and S_Q are dependent, A and Q are said to be in gametic phase disequilibrium (or linkage disequilibrium) [1].

Even when allele states at Q cannot be observed, inferences on the joint distribution between S_A and S_Q can be made by computing trait means for each of the genotypes at the A locus. Significant differences between these means implies that the A locus is in disequilibrium with a locus Q that has an effect on the trait. This disequilibrium is usually taken as evidence for linkage between loci A and Q [7], and this information for linkage that comes from the dependence between allele state indicators will be referred to as disequilibrium information. However, even when A and Q are linked, it is it is possible that the loci are in equilibrium [1]. Thus, disequilibrium information may not provide evidence for linkage even when loci are linked. Fortunately, a second type of information can provide evidence of linkage between loci even when they are in gametic phase equilibrium.

2.2 Cosegregation Information

Any allele on a haplotype either originates in the maternal or the paternal allele of the parent that transmitted the haplotype. Let O_A denote the grand-maternal or grand-paternal origin of the allele at locus A on some haplotype. Similarly, let O_Q

denote the grand-maternal or grand-paternal origin of the allele at locus Q on the same haplotype. We will refer to O_A and O_Q as allele origin indicators. If loci A and Q are linked, regardless of the joint distribution of allele state indicators, O_A and O_Q will be dependent, and this dependence is known as cosegregation. On the other hand, if loci A and Q are not linked, O_A and O_Q will be independent, and this independence is known as independent segregation.

Even when alleles at the Q locus cannot be observed, as explained below, inferences on the joint distribution between O_A and O_Q can be made by computing trait covariances between relatives. Consider a sire with genotype A_1A_2 at the A locus. Suppose the offspring of this sire can be classified into two groups: group I with offspring that received their sire's A_1 allele, and group II with offspring that received their sires A_2 allele. If A and Q loci are tightly linked, then, with high probability, offspring in marker group I will receive their sire's Q allele that is "linked" to the sires A_1 allele, and those in marker group II will receive their sire's other Q allele that is "linked" to the sires A_2 allele. Thus, covariances between trait values of offspring within either group I or II would be greater than covariances across the marker groups. On the other hand, if A and Q are not linked, the within group covariances will be identical to the across group covariances. Thus, the difference between within group and between group covariances can be used to infer linkage between A and Q. This information for linkage that comes from the dependence between allele origin indicators will be referred to as cosegregation information.

3 Genetic Evaluation

Now, we will see how disequilibrium information and cosegregation information from molecular genotypes are combined with pedigree and phenotypic information for genetic evaluation using mixed linear model methods. We will assume additive gene action for the quantitative trait locus linked to the marker (MQTL) and also for the other loci affecting the trait (RQTL). As usual, the RQTL will be assumed to be unlinked to the markers and the MQTL. Further, we will assume in this presentation that only two alleles, Q_1 and Q_2, are segregating at the MQTL. This last assumption is not necessary for the mixed linear model approach, but it does provide some simplifications that are worth observing.

3.1 Genotypes at Linked Locus in Disequilibrium

Genetic evaluation with genotypes at a marker locus that is linked and has an arbitrary level of gametic phase disequilibrium with the MQTL is considered here. This

type of analysis was first proposed by Goddard [4] and was further developed by Wang et al. [11], when disequilibrium was entirely due to crossbreeding and the marker locus was assumed to be in equilibrium with the MQTL in the purebreeds. Methodology to accommodate purebreeds with disequilibrium was considered by Fernando and Totir [3].

A marker locus with an arbitrary level of disequilibrium includes genotypes of types I, IIa and IIb. When disequilibrium is complete, i.e., when the MQTL genotype is known with certainty conditional on the observed genotype at the linked marker locus, type IIa genotypes are indistinguishable from type I genotypes. This special case, however, will be considered separately as there are some noteworthy simplifications.

Suppose genotypes at the MQTL were observed. Then, trait phenotypic values can be modeled as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{Q}\boldsymbol{\mu} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{y}$ is the vector of trait phenotypic values, $\boldsymbol{\beta}$ is a vector of non-genetic fixed effects, $\boldsymbol{\mu}$ has a single element, which is the additive effect of a Q_2 allele, $\boldsymbol{u}$ is the vector of additive effects of the RQTL, $\boldsymbol{e}$ is a vector of residuals, and $\boldsymbol{X}$, $\boldsymbol{Q}$ and $\boldsymbol{Z}$ are known incidence matrices. Given data from p animals, the incidence matrix $\boldsymbol{Q}$ will have p rows and a single column, with row i of $\boldsymbol{Q}$ containing the number of Q_2 alleles for animal i.

Now, for the situation considered here, the genotypes at the MQTL are not observed, and genotypes are available only at a linked locus. Thus, $\boldsymbol{Q}$ is an unobservable random matrix. The usual mixed model methodology cannot accommodate models with unobservable incidence matrices. However, we can write

$$\boldsymbol{a} = \boldsymbol{Q}\boldsymbol{\mu} - \mathrm{E}(\boldsymbol{Q}|\boldsymbol{M})\boldsymbol{\mu}, \tag{2}$$

where $\boldsymbol{M}$ denotes the observed genotypic information, and $\mathrm{E}(\boldsymbol{Q}|\boldsymbol{M})$ the conditional expectation of Q given $\boldsymbol{M}$. In the following, we will denote this conditional expectation by $\hat{\boldsymbol{Q}}$. In (2), $\boldsymbol{a}$ is a random vector with null mean, and now the model for the trait phenotypic values can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\hat{\boldsymbol{Q}}\boldsymbol{\mu} + \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}. \tag{3}$$

Provided we can compute $\hat{\boldsymbol{Q}}$, all the incidence matrices in this model are known, and the mixed model equations for this model can be setup provided we can compute the inverse of the covariance matrix for each of the random vectors $\boldsymbol{a}$ and $\boldsymbol{u}$. The covariance matrix for $\boldsymbol{u}$ is proportional to the additive relationship matrix. The inverse of the additive relationship matrix is sparse, and thus it can be computed efficiently [5]. On the other hand, the inverse of the covariance matrix for $\boldsymbol{a}$ is not

sparse, and thus its computation is not efficient. However, $\boldsymbol{Za}$ can be written as $\boldsymbol{Wv}$, where

$$a_i = v_i^m + v_i^p,$$

v_i^m and v_i^p are the additive effects of the maternal and paternal MQTL alleles of individual i, and $\boldsymbol{W}$ is a known incidence matrix relating $\boldsymbol{v}$ to $\boldsymbol{y}$. It can be shown that the covariance matrix, $\boldsymbol{\Sigma}_v$, for $\boldsymbol{v}$ can be calculated using a simple recursive formula that also leads to an efficient algorithm to invert $\boldsymbol{\Sigma}_v$ [3]. The model for trait phenotypic values now becomes

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\hat{Q}\mu} + \boldsymbol{Wv} + \boldsymbol{Zu} + \boldsymbol{e}. \tag{4}$$

When the marker locus is in equilibrium with the MQTL, as we will see in detail later, each row of $\hat{\boldsymbol{Q}}$ will be equal to a constant. Thus, $\boldsymbol{Z\hat{Q}\mu}$ can be dropped from the model. In this situation, only cosegregation information will contribute to the analysis through the modeling of covariances among MQTL effects. When disequilibrium is complete and all marker genotypes are observed, $\mathrm{E}(\boldsymbol{Q}|\boldsymbol{M}) = \boldsymbol{Q}$. Thus, in this situation, $\boldsymbol{v}$ is null, and after utilizing the disequilibrium information, cosegregation information does not contribute to the analysis. When disequilibrium is partial, $\mathrm{E}(\boldsymbol{Q}|\boldsymbol{M}) \neq \boldsymbol{Q}$, and $\boldsymbol{v}$ is not null. In this situation, disequilibrium information will contribute to the analysis through the model for the mean of MQTL effects, and cosegregation information will contribute to the analysis through the model for covariances between MQTL effects. These points should become clearer as we describe, in the following sections, how to compute $\hat{\boldsymbol{Q}}$ and the covariance matrix for $\boldsymbol{v}$.

3.1.1 Mean of MQTL additive genetic values

Recall that the mean of MQTL effects is $\hat{\boldsymbol{Q}}\mu$, where row i of $\boldsymbol{Q}$ has the number of Q_2 alleles for animal i. Thus, the i^{th} element of $\boldsymbol{Q}$ is the sum of two Bernoulli variables and has expected value:

$$\hat{Q}_i = p_i^m + p_i^p, \tag{5}$$

where

$$p_i^m = \Pr(O_Q(m,i) = Q_2|\boldsymbol{M}), \quad p_i^p = \Pr(O_Q(p,i) = Q_2|\boldsymbol{M}),$$

and $O_Q(m,i)$ is the maternal MQTL allele state and $O_Q(p,i)$ the paternal MQTL allele state of individual i. Let $F_Q(m,i) = A_j$ denote the event that the maternal MQTL allele of individual i originated in a founder haplotype with marker allele A_j. Then, p_i^m can be written as

$$p_i^m = \sum_j \Pr(F_Q(m,i) = A_j|\boldsymbol{M})\pi_j, \tag{6}$$

where π_j is the conditional probability that a founder haplotype with marker allele A_j has MQTL allele Q_2. Similarly, p_i^p can be written as

$$p_i^p = \sum_j \Pr(F_Q(p,i) = A_j|\boldsymbol{M})\pi_j. \tag{7}$$

Markov chain Monte Carlo (MCMC) methods can be used to compute the founder haplotype origin probabilities: $\Pr(F_Q(m,i) = A_j|\boldsymbol{M})$ and $\Pr(F_Q(p,i) = A_j|\boldsymbol{M})$ [8].

The π_j in (6) and (7) are the disequilibrium parameters. Thus, under equilibrium, where allele states S_A and S_Q are independent, the conditional probability of a Q_2 allele on a founder haplotype does not dependent on the marker allele on that haplotype, and so $\pi_1 = \pi_2 = \ldots = \Pr(Q_2)$. It follows that $p_i^m = p_i^p = \Pr(Q_2)$ for all i because

$$\sum_j \Pr(F_Q(m,i) = A_j|\boldsymbol{M}) = \sum_j \Pr(F_Q(p,i) = A_j|\boldsymbol{M}) = 1,$$

for all i. However, under disequilibrium, where allele states S_A and S_Q are not independent, the π_j are not all equal and it follows that $\hat{\boldsymbol{Q}}$ is not a vector of constants. Thus, disequilibrium information contributes to modeling the mean of MQTL effects.

3.1.2 Covariance of MQTL additive genetic values

The additive genetic value v_i^m is a Bernoulli variable times μ. Thus, the variance of v_i^m is

$$\mathrm{Var}(v_i^m) = \mu^2 p_i^m (1 - p_i^m), \tag{8}$$

and similarly, the variance of v_i^p is

$$\mathrm{Var}(v_i^p) = \mu^2 p_i^p (1 - p_i^p). \tag{9}$$

Under equilibrium, $p_i^m = p_i^p = \Pr(Q_2)$ and thus the variance of MQTL additive genetic values does not depend on the marker genotypes. When disequilibrium is present, however, p_i^m and p_i^p and thus the variance of MQTL additive genetic values depend on the marker genotypes.

To compute the off-diagonal elements of $\boldsymbol{\Sigma}_v$, we use the same formula that is used under equilibrium:

$$\begin{aligned}\mathrm{Cov}(v_i^m, v_k^p|\boldsymbol{M}) = {} & \Pr(O_Q(m,i) = m|\boldsymbol{M})\mathrm{Cov}(v_d^m, v_k^p|\boldsymbol{M}) \\ & + \Pr(O_Q(m,i) = p|\boldsymbol{M})\mathrm{Cov}(v_d^p, v_k^p|\boldsymbol{M}),\end{aligned} \tag{10}$$

where $O_Q(m,i) = m$, for example, is the event that the maternal MQTL allele of individual i originates in its dam's maternal allele [3]. Due to cosegregation of marker

and MQTL alleles, the conditional allele origin probabilities, $\Pr(O_Q(m,i) = m|\boldsymbol{M})$ and $\Pr(O_Q(m,i) = p|\boldsymbol{M})$ depend on marker genotypes, regardless of the level of disequilibrium between the marker locus and the MQTL. Thus, (10) shows that cosegregation information contributes to modeling covariances between MQTL additive genetic values.

The advantage of using (10) to compute $\boldsymbol{\Sigma}_v$ is that this leads to an efficient algorithm to invert this covariance matrix [3], and without such an algorithm, genetic evaluation with large pedigrees would not be possible. The disadvantage of using (10) is that it leads to an approximation of the covariance matrix and its inverse when marker data are not complete. Complete marker data in this situation are the ordered genotypes at the marker locus. Wang et al. [12] gave a recursive formula that gives exact results with unordered genotypes at a single locus. Recently, Thallman et al. [9] have developed a recursive formula that gives exact results with missing genotypes for a pedigree with loops.

3.2 Genotypes at Trait Locus

Here we consider genetic evaluation when genotypes at a trait locus are available. As mentioned earlier these genotypes are indistinguishable from type IIa genotypes when disequilibrium is complete. In this case, when genotypes are available for every animal in the pedigree, the mixed linear model reduces to

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{ZQ\mu} + \boldsymbol{Zu} + \boldsymbol{e}, \tag{11}$$

where the incidence matrix $\boldsymbol{Q}$ is observed. However, if some genotypes are missing, those elements of $\boldsymbol{Q}$ corresponding to the missing genotypes are not observed. In this case, the unobserved elements of $\boldsymbol{Q}$ can be replaced by their conditional expectations given the observed genotypes. If the MQTL has a large effect on the trait, for each individual i with a missing trait genotype, the deviation a_i of the MQTL genotypic value from its conditional expectation could be included in the model. So, when some genotypes are missing, the model becomes

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\hat{Q}\mu} + \boldsymbol{Z}_a\tilde{\boldsymbol{a}} + \boldsymbol{Zu} + \boldsymbol{e}, \tag{12}$$

where $\tilde{\boldsymbol{a}}$ is the vector with MQTL "deviations" for animals with missing trait genotypes, and $\boldsymbol{Z}_a$ is a known incidence matrix relating the elements of $\tilde{\boldsymbol{a}}$ to $\boldsymbol{y}$. As $\tilde{\boldsymbol{a}}$ is a vector of deviations, it will have a null mean. The variance of a_i can be calculated as

$$\mathrm{Var}(a_i) = \mu^2[p_i^m(1-p_i^m) + p_i^p(1-p_i^p)], \tag{13}$$

where, even for large pedigrees, the probabilities p_i^m and p_i^p can be calculated by iterative peeling [10, 2], conditional on the observed trait genotypes. The covariance

between any pair of elements a_i and a_j can be computed recursively as follows. Given any pair of individuals, one of them is not a descendant of the other. Suppose that j is not a descendant of i. Then, the covariance between a_i and a_j can be written as

$$\text{Cov}(a_i, a_j) = \frac{1}{2}[\text{Cov}(a_{d_i}, a_j) + \text{Cov}(a_{s_i}, a_j)], \tag{14}$$

where d_i is the dam and s_i the sire of i. If i is a founder, the two covariances on the right hand side of (14) are null. Further, if the dam of i is genotyped, a_{d_i} would be null, a_{d_i} would not be included in $\tilde{\boldsymbol{a}}$, and $\text{Cov}(a_{d_i}, a_j) = 0$; similarly, if the sire of i is genotyped, a_{s_i} would be null, a_{s_i} would not be included in $\tilde{\boldsymbol{a}}$, and $\text{Cov}(a_{s_i}, a_j) = 0$. The recursive formula (14) used here is identical to that used to compute the additive covariance matrix. Thus, the inverse of the covariance matrix for $\tilde{\boldsymbol{a}}$ can also be computed efficiently.

4 Discussion

In this paper we have discussed how molecular information can be incorporated into genetic evaluation. In this presentation we considered genotype information at a single locus. The principles presented here can also be used to incorporate genotype information at several linked loci. In this situation, the founder haplotype origin probabilities in equations (6) and (7), and the grand-parental origin probabilities in equation (10) are computed conditional on the observed genotypes at all the linked loci. In this case, the π_j disequilibrium parameters can be defined conditional on the allele states of all the marker loci. If several marker loci are used, the number of disequilibrium parameters may be too many to estimate accurately. One alternative is to define the disequilibrium parameters conditional on the allele states of the two marker loci flanking the MQTL.

Meuwissen and Goddard [6] have used a different approach to combine disequilibrium information and cosegregation information. Their approach has the advantage of a single disequilibrium parameter, regardless of the number of markers. On the other hand, while the method presented here does not make any assumptions on how disequilibrium was generated, their approach assumes that disequilibrium is due to a mutation a specified number of generations ago. Further, for their method it is necessary to specify the effective size of the population before pedigree data became available.

References

[1] J. F. Crow and M. Kimura. *An Introduction to Population Genetics Theory.* Harker and Row, Publishers, 1970.

[2] R. L. Fernando, C. Stricker, and R. C. Elston. An efficient algorithm to compute the posterior genotypic distribution for every member of a pedigree without loops. *Theor. Appl. Genet.*, 87:89–93, 1993.

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

[4] M. E. Goddard. A mixed model for analysis of data on multiple genetic markers. *Theor. Appl. Genet.*, 83:878–886, 1992.

[5] C. R. Henderson. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*, 32:69–83, 1976.

[6] T. H. E. Meuwissen and M. E. Goddard. Prediction of identity by descent probabilities from marker-haplotyes. *Genet. Sel. Evol.*, 33:605–634, 2001.

[7] M. Soller, T. Brody, and A. Genizi. On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theor. Appl. Genet.*, 47:35–39, 1976.

[8] C. Stricker, M. Schelling, F. Du, I. Hoeschele, S. A. Fernández, and R. L. Fernando. A compariosn of efficient genotype samplers for complex pedigrees and multiple linked loci. In *7th World Congress Genet. Appl. Livest. Prod.*, 2002. CD-ROM Communication N^o 21-12.

[9] R. M. Thallman, K. J. Hanford, S. D. Kachman, and L. D. Van Vleck. Sparse inverse of QTL covariance matrix with incomplete marker data. *Statistical Applications in Genetics and Molecular Biology*, 2004. in press.

[10] J. A. M. Van Arendonk, C. Smith, and B. W. Kennedy. Method to estimate genotype probabilities at individual loci in farm livestock. *Theor. Appl. Genet.*, 78:735–740, 1989.

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

[12] T. Wang, R. L. Fernando, S. van der Beek, M. Grossman, and J. A. M. van Arendonk. Covariance between relatives for a marked quantitative trait locus. *Genet. Sel. Evol.*, 27:251–274, 1995.