Fabiano Veraldo da Costa Pita

Construction of the Gametic Covariance Matrix for Quantitative Trait Loci Analyses in Outbred Populations

A dissertation submitted to the Genetics and Breeding Graduate Program of the Federal University of Viçosa in partial fulfillment of the requirements of the degree *Doctor Scientiae* .

Viçosa

Minas Gerais - Brazil

2003

Fabiano Veraldo da Costa Pita

Construction of the Gametic Covariance Matrix for Quantitative Trait Loci Analyses in Outbred Populations

A dissertation submitted to the Genetics and Breeding Graduate Program of the Federal University of Viçosa in partial fulfillment of the requirements of the degree *Doctor Scientiae* .

APPROVED: September 5th, 2003

Prof. Ricardo Frederico Euclydes (Committee member)	Profa. Simone E. Facioni Guimarães (Committee member)
Dra. Carmen Silva Pereira	Dr. Mário Luiz Martinez

Prof. Paulo Sávio Lopes
(Adviser)

Acknowledgments

I would like to thank my major advisers, Prof. Dr. Paulo Sávio Lopes at Federal University of Viçosa and Dr. Rohan Luigi Fernando at Iowa State University, for their support, patience, and friendship.

I am thankful to the National Research Council for Development-CNPq, to the Foundation for the Coordination of Higher Education and Graduate Training-CAPES Foundation, and to the Brazilian people for providing financial resources to my studies.

I would like to extend my gratitude to my committee members, Prof. Dr. Ricardo Frederico Euclydes, Prof. Dr. Robledo de Almeida Torres, and Profa. Dra. Simone Elisa Facioni Guimarães.

Thanks to all friends who I met during my graduation course, Brazilians from Viçosa, in special Rogerio and Fernanda, Ricardo and Luciana, Adriana, Urbano, and Brazilians from Ames, in special Cristiano and Carla, Aguimar and Grace, Raquel and Marcos, Artur. I am also very grateful to the friendship from Radu, Kim, Gretchen, Hauke, Petek and Mehmet.

All my love and gratefulness to all my family, in special my father Jaime, my mother Sandra, and my brother Juliano, for their orientation, support, and encouragement.

All my love to my wife Renata, who is always beside me.

Contents

RESUMO v

ABSTRACT viii

1 Introduction **1**

2 Review of Literature **3**

CALCULATION OF IBD PROBABILITIES 3

MARKER ASSISTED GENETIC EVALUATION 6

MARKOV CHAIN MONTE CARLO METHODS 13

LITERATURE CITED 17

3 A comparison of public domain programs for computing identical by descent coefficients. **22**

ABSTRACT 23

RESUMO 24

INTRODUCTION 25

METHODS 26

SIMULATIONS 28

RESULTS AND DISCUSSION 32
CONCLUSIONS 36
LITERATURE CITED 39

4 An improved approximation of the gametic covariance matrix for marker assisted genetic evaluation by BLUP. 42

ABSTRACT 43
RESUMO 44
INTRODUCTION 45
USE OF MCMC TO ESTIMATE PDQ'S 47
The SL sampler 48
The Haplotype Sampler 49
The Cascading-Origin Sampler 50
Simulations 51
RESULTS AND DISCUSSION 54
CONCLUSIONS 58
LITERATURE CITED 59

RESUMO

PITA, Fabiano Veraldo da Costa, D.S., Universidade Federal de Viçosa, Setembro 2003. **Construção da Matriz de Covariância Gamética para Análises de QTL em Populações Exogâmicas**. Orientador: Paulo Sávio Lopes. Comissão Orientadora: Ricardo Frederico Euclydes e Simone Elisa Facioni Guimarães.

A aplicação de análises de "Quantitative Trait Loci" (QTL) em populações exogâmicas é desafiadora porque pressuposições simplificadoras não podem ser aplicadas (por exemplo, os alelos QTL nao podem ser assumidos fixados em diferentes famílias, o número de alelos QTL segregantes não é conhecido *a priori*, não há desequilíbrio de ligação entre um dado alelo marcador e um dado alelo QTL). Quando o efeito genotípico do QTL é assumido aleatório no modelo de análise, a matriz de covariância gamética deve ser calculada para a realização das análises em populações exogâmicas. A acurácia dessa matriz é importante para a obtenção de estimativas confiáveis da posição ou efeito do QTL em análises de mapeamento, ou de valores genotípicos em avaliação genética assistida por marcadores. O objetivo do primeiro estudo foi avaliar diferente estratégias já implementadas em programas computacionais (SOLAR, LOKI, ESIP e MATVEC) para calcular a matriz de coeficientes Idênticos por Descendência (IBD), que é necessária para o mapeamento de QTL em populações exogâmicas. SOLAR utiliza um método baseado em regressão linear, LOKI e ESIP são ambos baseados em "reverse peeling" e o amostrador implementado em MATVEC amostra indicadores de segregação. Um pedigree com estrutura F2 típica foi simulado com uma família F2 pequena (2 indivíduos) ou grande (20 indivíduos) e marcadores

flanqueadores localizados a 2 cM, 5 cM ou 10 cM de distância um do outro, com o QTL localizado no meio do intervalo. A habilidade dessas estratégias em lidar com informações de marcadores perdidas foi avaliada assumindo um dos pais da geração F2 com ou sem informação de marcador. SOLAR nao estimou os coeficientes IBD corretamente para a maior parte das situações simuladas, enquanto que LOKI apresentou problemas quando o tamanho da família F2 era grande. ESIP e o amostrador em MATVEC apresentaram bom desempenho em todas as situacões simuladas, com estimativas de coeficientes IBD próximas aos coeficientes verdadeiros. Portanto, ESIP e MATVEC são os softwares mais indicados quando análises genéticas são realizadas em pedigrees com estruturas complexas. O objetivo do segundo estudo foi avaliar o efeito da utilização de uma melhor aproximação da inversa da matriz de covariância gamética para a avaliação genética de grandes populações de animais domésticos. Algoritmos eficientes, baseados no rastreamento dos alelos QTL de um indivíduo em relação aos de seus avós (Probabilidade de Descendência de um QTL - PDQ), podem ser usados para construir a inversa da matriz de covariância gamética diretamente. Mas essa inversa é uma aproximação quando há informação incompleta de marcador. Também, o cálculo exato de PDQ's torna-se difícil quando a informação de marcador é incompleta. Nesse estudo, a inversa da matriz de covariância gamética para uma população exogâmica simulada foi calculada usando o algoritmo eficiente, mas as PDQ's foram calculadas usando um algoritmo Monte Carlo Cadeia de Markov (MCMC). Essa inversa foi utilizada para predizer o valor genético dos indivíduos através de BLUP assistido por marcadores (MABLUP). O efeito dos cálculos de PDQ usando o algoritmo MCMC sobre a acurácia da MABLUP foi avaliado com base na resposta a seleção realizada, calculada para o pedigree simulado. Os resultados mostraram que

quando as PDQ's foram estimadas usando MCMC a perda em resposta devido ao uso da inversa aproximada pode ser reduzida em aproximadamente 20%, enquanto que em estudos anteriores essa redução foi de 50%. Ainda, quando quatro marcadores bi-alélicos foram utilizados a resposta para MABLUP foi maior e a perda em resposta devido a marcadores com informação perdida foi menor, quando comparadas à situação onde apenas dois marcadores bi-alélicos foram utilizados.

ABSTRACT

PITA, Fabiano Veraldo da Costa, D.S., Universidade Federal de Viçosa, September 2003. **Construction of the Gametic Covariance Matrix for Quantitative Trait Loci Analyses in Outbred Populations**. Adviser: Paulo Sávio Lopes. Committee members: Ricardo Frederico Euclydes and Simone Elisa Facioni Guimarães.

The application of Quantitative Trait Loci (QTL) analyses in outbred population is challenging because simplified assumptions do not hold for these populations (e.g., the QTL alleles cannot be assumed fixed in different families, the number of QTL alleles segregating is not known *a priori*, there is not gametic phase disequilibrium between a given genetic marker allele and a QTL allele). When the QTL genotypic effect is assumed random, the gametic covariance matrix must be calculated to perform QTL analyses in outbred populations. The accuracy of this matrix is important to obtain reliable estimates of QTL position or effect when applying QTL mapping, or QTL genotypic values when applying Marker Assisted Genetic Evaluation. The objective of the first study was to evaluate the different strategies already implemented in softwares (SOLAR, LOKI, ESIP and MATVEC) to calculate the matrix of identical by descent (IBD) coefficients, which is required for QTL mapping analysis in outbred populations. SOLAR uses a regression method, LOKI and ESIP are both based on reverse peeling, and the MATVEC sampler samples segregation indicators. A typical F2 pedigree was simulated with a small (2 offspring) or a large (20 offspring) F2 family, and the flanking markers were simulated 2 cM, 5 cM, or 10 cM apart, with the QTL located in the middle. The ability of these strategies to deal with missing genetic marker information was evaluated assuming one of the F2

parents with or without marker information. SOLAR failed to estimate the correct coefficients at almost all situations simulated, while LOKI showed problems when a large family was present in the pedigree. ESIP and MATVEC sampler performed well at all situations, providing IBD coefficients closed to the true ones. Therefore, ESIP and MATVEC are more indicated when genetic analysis are carried out on complex pedigree structures. The objective of the second study was to evaluate the effect of using a better approximation of the inverse of the gametic covariance matrix on the genetic evaluation of large livestock populations. Efficient algorithms, based on tracing the QTL alleles of an individual to its grandmother or grandfather (probability of descent a QTL - PDQ's), can be used to construct the inverse of the gametic covariance matrix directly. But this inverse is an approximation when incomplete marker information is available. Also, computing the exact PDQ's becomes difficult when marker information is incomplete. In this study, the inverse of the gametic covariance matrix for a simulated outbred pedigree was calculated using the efficient algorithm, but the PDQ's were calculated using a Markov chain Monte Carlo (MCMC) algorithm. This inverse was used to calculate the predicted genetic value of individuals through Marker Assisted Best Linear Unbiased Prediction (MABLUP). The effect of PDQ calculations using the MCMC algorithm on MABLUP accuracy was evaluated based on the realized response to selection for the simulated pedigree. The results showed that by estimating the PDQ's by MCMC the loss in response because of using an approximate inverse could be reduced to about 20%, while in previous studies this reduction was of 50%. Further, response to MABLUP was greater when four bi-allelic markers were used, and the loss in response due to missing markers was smaller in the case with four markers compared to when only two bi-allelic markers were used.

Chapter 1

Introduction

The molecular biology techniques had an extraordinary development in the last two decades. As a result of these improvements, specific genes affecting a Quantitative Trait Loci (QTL) now can be mapped and their inheritance followed on livestock populations.

Many approaches have been developed to associate variations in the genome of individuals with variations in their phenotype. Comparative and candidate gene approaches rely on the information of genes and gene functions from other species, mainly human and mouse, whose genome projects are better developed (Rothschild and Soller, 1999), while QTL mapping approach relies on the information of genetic markers spread on chromosomes with the aim to detect regions of the genome with putative QTL. The QTL mapping approach is applied either when no prior information is available for the specific species and economic traits (Rothschild, 1998) or to narrow down the space of search, allowing the use of the candidate gene approach in a specific portion of the genome. The QTL mapping methods assuming the putative

QTL as a random effect have been used for mapping QTL in outbred populations. In these methods, the detection power and the accuracy of the estimated parameters depend greatly on the accuracy of the gametic covariance matrix calculation, which models the covariance of the QTL genotypes among the individuals.

When a QTL mapping approach is applied in a population and a specific portion of a chromosome is identified to have a significant effect upon some economic trait, the genetic markers surrounding this specific portion are expected to be in gametic phase disequilibrium with the segregating QTL responsible by a fraction of the trait phenotypic variation. A QTL that is linked to a marker will be called the marked quantitative trait locus (MQTL). Thus, genetic markers can be used to trace the segregation of the MQTL in the population and to incorporate the MQTL effect in the ordinary genetic evaluation of livestock populations. The inclusion in the Henderson Mixed Model Equations (HMME) of the random MQTL effect using genetic marker information and the calculation of MQTL genotypic values using BLUP (Best Linear Unbiased Predictor) methodology was showed by Fernando and Grossman (Fernando and Grossman, 1989). The accuracy of the BLUP prediction for the MQTL genotypic values also depends on the accuracy of the gametic covariance matrix calculation.

The calculation of the gametic covariance matrix is not an easy task for outbred populations, where genetic markers and the MQTL are assumed under gametic phase equilibrium, the number of segregating MQTL alleles is not known *a priori* , and some individuals do not have genetic marker information. The objectives of this study were: 1) evaluate methods to construct the gametic covariance matrix for QTL mapping; 2) evaluate a Markov Chain Monte Carlo algorithm to construct the gametic covariance matrix for Marker Assisted Genetic Evaluation.

Chapter 2

Review of Literature

CALCULATION OF IBD PROBABILITIES

In Backcross or F2 designs, the QTL mapping analyses can be performed using either Maximum Likelihood (Lander and Botstein, 1989) or Ordinary Least Squares (Haley and Knott, 1992) methods, assuming trait divergent inbred lines as founders and gametic phase disequilibrium between genetic markers and the putative QTL. These experimental QTL mapping designs assume a heterozygositie of unity for both markers and segregating QTL in the F1 generation, which simplifies the analyses, but the results maybe cannot have immediate practical interest because of the poor performance of exotic breeds used as founders (Malek *et al.*, 2001).

Populations for QTL mapping built with commercial lines/breeds are closer to commercial crosses used in livestock production. In these crosses, a common assumption is that lines are fixed for different QTL alleles, but there is segregation of genetic marker alleles within each outbred line (Haley *et al.*, 1994). Haley et al. (Haley *et al.*,

1994) presented a method based on Ordinary Least Square analysis for mapping QTL in crosses between outbred lines, using multiple marker information. This method is computationally fast and allows to fit in the model of analysis many QTL together with additional parameters (fixed effects and covariates), but it does not make use of the information from the F2 sibs. Also, this method assumes a three generation fixed pedigree structure, all individuals are genotyped for all markers, and the QTL has different alleles fixed for each line. These assumptions are not expected to hold in outbred populations.

At outbred populations (e.g., pure breeds at breeding herds and human populations) much more sophisticated methods of QTL analysis should be applied. It is because at these populations there is information available over multiple generations, substantial amount of missing data is present, individuals are related across families, and the number of QTL and marker alleles segregating and their frequencies are not known *a priori* . In addition, populations that have not been formed recently and for a moderate-resolution marker map (e.g., a 10 cM map) the gametic phase disequilibrium ratios between markers and QTL must be expected to be zero (Hoeschele, 2001).

The variance component method for QTL mapping has been recognized as a very useful method for QTL mapping in outbred populations because its ability to deal with complex pedigrees (Grignola *et al.*, 1996; Hoeschele, 2001). In this method the QTL component is assumed random and its covariance structure must be provided. The gametic covariance matrix among relatives for a given MQTL can be constructed using pedigree and marker information (Chevalet *et al.*, 1984; Fernando and Grossman, 1989; Weller and Fernando, 1991). For a given locus, the genotypes of relatives

are similar because they share genes that are identical-by-descent (IBD). Two genes are said IBD if they are identical copies of a gene segregating from a common ancestor within the defined pedigree (Thompson, 2000). Thus, assuming additive genetic effect, the genetic covariance between an individual X with alleles i and j at locus k (G^k_{ij}) and an individual Y with alleles m and n at locus k (G^k_{mn}) can be calculated as:

$$\begin{aligned}
\mathrm{Cov}(G^k_{ij}, G^k_{mn}) &= E[(\alpha^k_i + \alpha^k_j)(\alpha^k_m + \alpha^k_n)] - E[(\alpha^k_i + \alpha^k_j)]E[(\alpha^k_m + \alpha^k_n)] \\
&= E(\alpha^k_i \alpha^k_m) + E(\alpha^k_i \alpha^k_n) + E(\alpha^k_j \alpha^k_m) + E(\alpha^k_j \alpha^k_n) \\
&= Pr(i \equiv m)E(\alpha_i^{k^2}) + Pr(i \equiv n)E(\alpha_i^{k^2}) + \\
&+ Pr(j \equiv m)E(\alpha_j^{k^2}) + Pr(j \equiv n)E(\alpha_j^{k^2}) \\
&= \left[\frac{Pr(i \equiv m) + Pr(i \equiv n) + Pr(j \equiv m) + Pr(j \equiv n)}{2}\right]\sigma^2_k
\end{aligned} \tag{2.1}$$

where: α^b_a is the average effect of allele a at locus b, $Pr(z \equiv w)$ is the probability of allele z being IBD to allele w, σ^2_k is the additive genetic variance for locus k, and $E(\alpha^k_i) = E(\alpha^k_j) = E(\alpha^k_m) = E(\alpha^k_n) = 0$.

The term $[Pr(i \equiv m) + Pr(i \equiv n) + Pr(j \equiv m) + Pr(j \equiv n)]/2$ is twice the Malecot's coefficient of relationship (Malecot, 1948) for a given locus. The set of IBD probabilities among all individuals in a pedigree is called the IBD probability matrix, and for a given QTL it is calculated based on the genetic markers linked to this QTL. But the calculation of IBD probabilities for the MQTL locus can be a difficulty task at outbred populations even when the marker information is available for all individuals and all markers. This difficulty arises because the order of the genotypes of individuals, and consequently the gametic linkage phase between genetic markers

(when more than one genetic marker locus is considered) cannot be inferred for sure in some situations. Also, in outbred populations some markers can be fixed for one allele, loops are present (mate of individuals more related than the average of the population), and the pedigree extends for many generations. When some individuals do not have genetic marker information for some loci, then it must be inferred and more difficulty to calculate IBD probabilities is expected.

Methods have been proposed to calculate IBD probabilities for a MQTL in outbred populations, given marker and pedigree information. Some of these methods applied a deterministic approach (Almasy and Blangero, 1998; Pong-Wong *et al.*, 2001; Sorensen *et al.*, 2002) that, in general, is fast but do not use all available information (information from loci and/or relative individuals). As a result, the IBD probabilities are approximations. Stochastic methods are also used for IBD calculations (Heath, 1997; Pérez-Enciso *et al.*, 2000; Fernández *et al.*, 2001; Stricker *et al.*, 2002). Markov chain Monte Carlo (MCMC) methods can use all information available, but are computing demanding and in some situations the chain may not be irreducible (Cannings and Sheehan, 2002).

MARKER ASSISTED GENETIC EVALUATION

When a MQTL is known to have an important effect upon a phenotype trait in a given population and genetic markers are available, the effect of this MQTL should be considered in genetic evaluations. Marker Assisted Genetic Evaluation by BLUP is expected to increase the accuracy of genetic value predictions and, as a result, improve selection responses (Weller and Fernando, 1991; Totir *et al.*, 2003).

Let a be the vector of unobservable genotypic values of candidates to selection and D represents all the information available to predict the unobservable genotypic values of the candidates. When D consists of pedigree relationships and trait phenotypes, and when the vector y of trait phenotypes and the vector a of unobservable genotypic values are assumed to have a multivariate normal distribution, the conditional mean of a is a linear function of y:

$$E(a|D) = \mu_a + CV^{-1}(y - \mu_y), \tag{2.2}$$

where, conditional on pedigree information (P), μ_a and μ_y are the expected values of a and y ($E(a|P)$ and $E(y|P)$, respectively), C is the covariance matrix between a and y, and V is the covariance matrix of y. Regardless of the joint distribution of a and y, the linear function of y on the right side of equation (2.2) gives the best linear predictor of a (Henderson, 1984).

Suppose y can be modeled as:

$$y = X\beta + Za + e \tag{2.3}$$

where X and Z are known incidence matrices, β is an unknown vector of fixed effects, a is a vector of additive effects of individuals, and e is a vector of residuals.Then, the expected value of y can be written as $\mu_y = X\beta$ and the covariance matrices C and V can be written as:

$$C = G_u Z' \tag{2.4}$$

and

$$V = Z'G_u Z + R, \tag{2.5}$$

where G_u is the conditional covariance matrix of a given P and R is the covariance matrix of e.

Now, consider a QTL closely linked to a marker. Suppose there is gametic phase equilibrium between the QTL and the marker, and this marker is not linked to any other trait locus. In this case:

$$E(a|P) = E(a|P, M) \tag{2.6}$$

and

$$E(y|P) = E(y|P, M), \tag{2.7}$$

where M is the marker information.

As there is gametic phase equilibrium between the QTL and the marker, the μ_a and μ_y are not affected by the marker information, but:

$$Var(a|P) \neq Var(a|P, M), \tag{2.8}$$

where $Var(a|P) = G_u$, and $Var(a|P, M) = G_g$ is the conditional covariance matrix of a given P and M.

The conditional covariance matrix of a given P can be constructed as:

$$G_u = A\sigma_u^2, \tag{2.9}$$

where A is the numerator of relationship matrix (Henderson, 1976), and σ_u^2 is the additive genetic variance for the polygenic effect.

The conditional covariance matrix of a given P and M can be constructed as:

$$G_g = \phi\sigma_k^2 + A\sigma_u^2, \tag{2.10}$$

where ϕ is the IBD probability matrix for a MQTL locus and $\sigma_k^2 = 2pq\alpha^2$ (Falconer and Mackay, 1996), assuming no dominance, where p and q are the MQTL allele frequencies and α is the genotypic value of the homozygous MQTL genotype.

Thus, C and V are affected by the marker information (Fernando and Totir, 2003). Using G_g in Henderson's mixed model equations (HMME) would give prediction of genotypic value of individuals under genetic evaluation. This, however, requires obtaining the inverse of G_g (G_g^{-1}), which is not sparse and cannot be inverted efficiently. As a result, genetic evaluation of real livestock populations with thousands of individuals using HMME is not feasible.

In order to use HMME for marker assisted BLUP, Fernando and Grossman (Fernando and Grossman, 1989) modeled the additive effect of each individual i (a_i) as:

$$a_i = v_i^m + v_i^p + u_i \tag{2.11}$$

where v_i^m and v_i^p are the additive effects of the maternal and paternal alleles, respectively, at the MQTL, and u_i is the additive effect of the remaining trait loci (polygenic effect).

For this model, the covariance among the maternal and paternal MQTL alleles is:

$$G_v = \Sigma_k \sigma_v^2 \tag{2.12}$$

where Σ_k is the IBD probability matrix for the MQTL alleles at locus k and $\sigma_v^2 = pq\alpha^2$ (Falconer and Mackay, 1996), assuming no dominance. G_v will be called hereafter as the gametic covariance matrix for the MQTL.

The vector a of n animals can be written as (Fernando and Grossman, 1989):

$$a = Kv + u, \tag{2.13}$$

where v is a vector of $2n$ gametic effects, each of the n individuals having a paternal MQTL effect (v_i^p), and a maternal MQTL effect (v_i^m); $K_{n\times 2n}$ is an incidence matrix relating v to a, the i'th row of K having ones for the elements corresponding to the gametic effects at the MQTL of individual i, and zeroes for the remaining elements. For the genetic model at (2.11), the mixed linear model becomes:

$$y = X\beta + ZKv + Zu + e, \tag{2.14}$$

and, if we define $W = ZK$, the HMME can be written as:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}W & X'R^{-1}Z \\ W'R^{-1}X & W'R^{-1}W + G_v^{-1} & W'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}W & Z'R^{-1}Z + G_u^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{v} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ W'R^{-1}y \\ Z'R^{-1}y \end{bmatrix} \tag{2.15}$$

Now, to obtain BLUP from the HMME given in (2.15), the inverses of G_v and G_u need to be calculated. Both G_v^{-1} and G_u^{-1}, however, have a sparse structure. This feature permits the development of algorithms to efficiently obtain these inverses, allowing the consideration of many individuals in the HMME. The inverse of G_u can be obtained using the same algorithms developed for BLUP with only relationship and phenotypic informations (Henderson, 1976; Quaas, 1988), while Fernando and Grossman (Fernando and Grossman, 1989) showed an efficient algorithm to obtain G_v^{-1} for situations where the parental origin of the genetic markers can be inferred.

But, the knowledge of the parental origin of the genetic markers is not common for the typical livestock pedigree structures. Wang et al.(Wang *et al.*, 1995) extended the

algorithm presented by Fernando and Grossman (Fernando and Grossman, 1989) to deal with unknown origin of marker alleles. Their method is based in the concept of probability of descent a QTL (PDQ). Supposing s and d are parents of an individual i, and j is not a direct descendant of i, the conditional covariance of the additive effects of QTL alleles $Q_i^{k_i}$ and $Q_j^{k_j}$ in individuals i and j, given the observed marker genotypes (G_{obs}), is:

$$Cov(v_i^{k_i}, v_j^{k_j}|G_{obs}) = Pr(Q_i^{k_i} \equiv Q_j^{k_j}|G_{obs})\sigma_v^2, \tag{2.16}$$

where k_i and k_j can be 1 or 2 (allele 1 or 2), $v_i^{k_i}$ and $v_j^{k_j}$ are the additive effects of $Q_i^{k_i}$ and $Q_j^{k_j}$, respectively, and $Pr(Q_i^{k_i} \equiv Q_j^{k_j}|G_{obs})$ is the conditional probability that $Q_i^{k_i}$ been IBD to $Q_j^{k_j}$ given G_{obs}.

Because individuals s and d are parents of i, $Q_i^{k_i}$ can be identical by descent to $Q_j^{k_j}$ in four ways (Fernando and Grossman, 1989), then the probability in (2.16) can be written as:

$$\begin{aligned} Pr(Q_i^{k_i} \equiv Q_j^{k_j}|\boldsymbol{G}_{obs}) = & Pr(Q_i^{k_i} \Leftarrow Q_s^1, Q_s^1 \equiv Q_j^{k_j}|G_{obs}) + \\ & Pr(Q_i^{k_i} \Leftarrow Q_s^2, Q_s^2 \equiv Q_j^{k_j}|G_{obs}) + \\ & Pr(Q_i^{k_i} \Leftarrow Q_d^1, Q_d^1 \equiv Q_j^{k_j}|G_{obs}) + \\ & Pr(Q_i^{k_i} \Leftarrow Q_d^2, Q_d^2 \equiv Q_j^{k_j}|G_{obs}) \end{aligned} \tag{2.17}$$

Because individual j is not a direct descendant of individual i, and marker genotypes of s and d are known, the conditional sampling of $Q_i^{k_i}$ from s and d is independent of alleles in j being identical by descent to alleles in s or d, given G_{obs}. In this

way the probability in (2.16) can be computed recursively as:

$$
\begin{aligned}
Pr(Q_i^{k_i} \equiv Q_j^{k_j} | G_{obs}) = & \\
& Pr(Q_i^{k_i} \Leftarrow Q_s^1 | G_{obs}) Pr(Q_s^1 \equiv Q_j^{k_j} | G_{obs}) + \\
& Pr(Q_i^{k_i} \Leftarrow Q_s^2 | G_{obs}) Pr(Q_s^2 \equiv Q_j^{k_j} | G_{obs}) + \\
& Pr(Q_i^{k_i} \Leftarrow Q_d^1 | G_{obs}) Pr(Q_d^1 \equiv Q_j^{k_j} | G_{obs}) + \\
& Pr(Q_i^{k_i} \Leftarrow Q_d^2 | G_{obs}) Pr(Q_d^2 \equiv Q_j^{k_j} | G_{obs})
\end{aligned}
\tag{2.18}
$$

The terms $Pr(Q_i^{k_i} \Leftarrow Q_p^{k_p} | G_{obs})$ in (2.18) are the conditional probability that allele $Q_i^{k_i}$ in offspring i descended from allele $Q_p^{k_p}$ in parent p (s or d) for k_i, k_p (1 or 2). These conditional probabilities are the PDQ's for a MQTL allele. Assuming a diallelic system, there are eight PDQ's for each individual. For an individual i these PDQ's can be grouped in a matrix:

$$
\mathbf{B_i} = \begin{bmatrix}
Pr(Q_i^1 \Leftarrow Q_s^1 | G_{obs}) & Pr(Q_i^1 \Leftarrow Q_s^2 | G_{obs}) & Pr(Q_i^1 \Leftarrow Q_d^1 | G_{obs}) & Pr(Q_i^1 \Leftarrow Q_d^2 | G_{obs}) \\
Pr(Q_i^2 \Leftarrow Q_s^1 | G_{obs}) & Pr(Q_i^2 \Leftarrow Q_s^2 | G_{obs}) & Pr(Q_i^2 \Leftarrow Q_d^1 | G_{obs}) & Pr(Q_i^2 \Leftarrow Q_d^2 | G_{obs})
\end{bmatrix}
\tag{2.19}
$$

Then, this matrix of PDQ's is used to calculate G_v^{-1} through a tabular method (Wang *et al.*, 1995). But, this tabular method just produces the exact variance-covariance matrix for the MQTL when there is not missing marker information. It is because only under this condition the assumption of independence of the events $Q_i^{k_i} \Leftarrow Q_s^1$ and $Q_s^1 \equiv Q_j^{k_j}$ made at (2.18) is true. Further, when multiple markers are considered (Goddard, 1992), to get exact results the recursive algorithm requires phase information in addition to the genotypes for all individuals (Fernando, 1998; Hoeschele, 2001).

MARKOV CHAIN MONTE CARLO METHODS

Both IBD's and PDQ's can be calculated using stochastic methods. Monte Carlo likelihood is becoming increasingly used where exact likelihood analysis is computationally infeasible. These likelihood arise at genetic mapping analysis when pedigrees are complex, many loci and alleles per locus are available, and phenotypic traits and marker information are observed for just a subset of individuals (Thompson, 1994). These features cause computational difficulties for likelihood calculation since the number of possible underlying configurations of genes on all the relevant individuals of the pedigree increases vastly. Monte Carlo likelihood estimation can replace exact computation of likelihood when it is difficult to calculate analytically.

The statistical problems involved in fitting genetic linkage models to trait data y on a set of related individuals may be viewed as latent variable or "missing data" problems (Thompson, 1994). Were the underlying multilocus genotypes (pair of haplotypes) of all individuals observable, likelihood computation and parameter estimation would be trivial, but only the trait data (phenotypes) and single locus marker genotypes of some individuals are observed. Let assume x being the underlying genotypes, recombination events and/or other unobserved indicators of the patterns of genes segregating in pedigrees. The observed trait and marker data will be denoted by y, and where necessary this will be separated into its trait T and marker M components. The vector θ will denote the complete set of parameters underlying a genetic model. The likelihood is (Thompson, 1994):

$$L(\theta) = P_\theta(y) = \sum_x P_\theta(y, x) = \sum_x P_\theta(y|x) P_\theta(x) \tag{2.20}$$

Although the summation may be infeasible, it is supposed that the latent variables

x are chosen in such a way that each term of the expression is easily computed. In this way equation (2.20) may also be written as:

$$P_\theta(y) = \sum_x P_\theta(y|x)P_\theta(x) = E_\theta P_\theta(y|x) \tag{2.21}$$

where the expectation is over x-values, with probabilities according to the "prior" $P_\theta(x)$.

The problem with these early Monte Carlo likelihoods is that sampling of genotypes x is not conditioned on the data y. Thus, on a large pedigree, the vast majority of realizations x provide minute (or even zero) likelihood contributions (Thompson, 2000).

This situation was changed dramatically by the explosion in use of Markov chain Monte Carlo (MCMC) methods, these provide for simulating from:

$$P_\theta(x|y) = \frac{P_\theta(y,x)}{P_\theta(y)} \tag{2.22}$$

These methods include the Gibbs sampler (Geman and Geman, 1984) and the Metropolis algorithm (Metropolis *et al.*, 1953). With a well-chosen latent variables x, the numerator of equation (2.22) is readily evaluated, but the denominator is:

$$L(\theta) = P_\theta(y) = \sum_x P_\theta(x,y) \tag{2.23}$$

and this summation is often infeasible. The denominator is, in fact, precisely the likelihood whose exact evaluation is impossible, necessitating the Monte Carlo estimation (Thompson, 1994). Metropolis-Hastings algorithms are Markov chain Monte Carlo methods designed to meet this need, providing realizations (approximately) from a distribution known up to a normalizing constant (Hastings, 1970). For each x

a "proposal distribution" $q(.,x)$ is defined. Then, if the process is now at x, the next value is generated as follows:

1- Generate x^* from the proposal distribution $q(.,x)$;

2- Compute the Hastings ratio:

$$h = \frac{q(x,x^*)P_\theta(x^*|y)}{q(x^*,x)P_\theta(x|y)} = \frac{q(x,x^*)P_\theta(y,x^*)}{q(x^*,x)P_\theta(y,x)} \tag{2.24}$$

Note that h can be computed without knowledge of $P_\theta(y)$;

3- With probability $h^* = min(1,h)$ the process moves to x^* and with probability $(1-h^*)$ it remains at x.

The distribution at equation (2.22) is an equilibrium distribution of the Markov chain just defined. Provided $q(.,.)$ is chosen so that the chain is ergodic, running the chain provides (after a sufficient number of steps) realizations from the distribution at equation (2.22) (Thompson, 1994).

The algorithm of Metropolis et al. (Metropolis *et al.*, 1953) is a special case where $q(x^*,x) = q(x,x^*)$. Then, the Hastings ratio reduces to the odds ratio of the proposal state x^* versus the current state x. The Gibbs sampler proposed by Geman and Geman (Geman and Geman, 1984) is also a special case, in which the proposal distribution is the conditional distribution for changing one element of x conditional on current values of the others, in which case the Hastings ratio reduces to unity and there is no rejection step. However, the fact of no rejection step is not necessarily advantageous, because the Gibbs sampler can make only small changes in x (Thompson, 1994).

An important issue at MCMC analysis is the choice of the latent variables x (Thompson, 1994). The most straightforward implementation in linkage analysis is

to use the multilocus genotypes of individuals as x. This implementation is present at MCMC samplers ESIP (Fernández *et al.*, 2001; Fernández *et al.*, 2002) and LOKI (Heath, 1997). Also, a MCMC sampler based on segregation indicators (Sobel and Lange, 1996; Stricker *et al.*, 2002) have shown to be efficient to calculate IBD probabilities, with good mixing under situations common in livestock pedigrees, like large family sizes and markers tightly linked (Stricker *et al.*, 2002).

The set of segregation indicators specifies the flow of alleles in a pedigree without specifying their states. In some situations, sampling of segregation indicators may be more efficient because of their smaller state space (Sobel and Lange, 1996). Sobel and Lange (Sobel and Lange, 1996) presented a Metropolis-Hastings (Hastings, 1970) algorithm to draw samples from the posterior distribution of segregation indicators for the marker loci given the marker and pedigree information.

LITERATURE CITED

ALMASY, L., and J. BLANGERO, 1998 Multipoint quantitative trait linkage analysis in general pedigrees. Am. J. Hum. Genet. **62**: 1198–1211.

CANNINGS, C., and N. A. SHEEHAN, 2002 On a misconception about irreducibility of the single-site gibbs sampler in a pedigree application. Genetics **162**: 993–996.

CHEVALET, C., M. GILLOIS, and J. VU TIEN KHANG, 1984 Conditional probabilities of identity of genes at a locus linked to a marker. Genet. Sel. Evol. **16**: 431–444.

FALCONER, D. S., and T. F. C. MACKAY, 1996 *Introduction to quantitative genetics.* Longman, Edinburgh.

FERNÁNDEZ, S. A., R. L. FERNANDO, B. GULDBRANDTSEN, C. STRICKER, M. SCHELLING, and A. L. CARRIQUIRY, 2002 Irreducibility and efficiency of esip to sample marker genotypes in large pedigrees with loops. Genet. Sel. Evol. **34**: 537–555.

FERNÁNDEZ, S. A., R. L. FERNANDO, B. GULDBRANDTSEN, L. R. TOTIR, and A. L. CARRIQUIRY, 2001 Sampling genotypes in large pedigrees with loops. Genet. Sel. Evol. **33**: 337–367.

FERNANDO, R. L., 1998 Genetic evaluation and selection using genotypic, phenotypic and pedigree information. *In:* 6th WCGALP, volume 26, p 329–336.

FERNANDO, R. L., and M. GROSSMAN, 1989 Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. **21**: 467–477.

FERNANDO, R. L., and R. L. TOTIR, 2003 *Poultry genetics, breeding and biotechnology*, chapter Incorporating molecular information in breeding programs: methodology, p 744. CABI publishing.

GEMAN, S., and D. GEMAN, 1984 Stochastic relaxation, gibbs distribution, and bayesian restoration of images. IEEE Tansactions on Pattern analysis and machine intelligence **6**: 721–741.

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

GRIGNOLA, F. E., I. HOESCHELE, and B. TIER, 1996 Mapping quantitative trait loci in outcross populations via residual maximum likelihood. i methodology. Genet. Sel. Evol. **28**: 479–490.

HALEY, C. S., and S. A. KNOTT, 1992 A simple regression model for interval mapping in line crosses **69**: 315–324.

HALEY, C. S., S. A. KNOTT, and J. M. ELSEN, 1994 Mapping quantitative trait loci in crosses between outbred lines using least squares **136**: 1195–1207.

HASTINGS, W. K., 1970 Monte carlo sampling methods using markov chain and their applications. Biometrika **57**: 97–109.

HEATH, S. C., 1997 Markov chain monte carlo segregation and linkage analysis for oligogenic models. Am. J. Hum. Genet. **61**: 748–760.

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

HENDERSON, C. R., 1984 Applications of linear models in animal breeding. University of Guelph.

HOESCHELE, I., 2001 *Handbook of statistical genetics*, chapter Mapping quantitative trait loci in outbred pedigrees, p 599–639. John Wiley & Sons Ltd.

LANDER, E. S., and D. BOTSTEIN, 1989 Mapping mendelian factors underlying quantitative trait using rflp linkage maps **121**: 185–199.

MALECOT, G., 1948 *Les mathématiques de l'hérédité*. Masson, Paris.

MALEK, M., J. C. M. DEKKERS, H. K. LEE, T. J. BAAS, and M. F. ROTHSCHILD, 2001 A molecular genome scan analysis to identify chromosomal regions influencing economic traits in the pig. I. growth and body composition. Mammalian Genome **12**: 630–636.

METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER, and E. TELLER, 1953 Equations of state calculations by fast computing machines. J. Chem. Phys. **21**: 1087–1092.

PÉREZ-ENCISO, M., L. VARONA, and M. F. ROTHSCHILD, 2000 Computation of identity by descent probabilities conditional on dna markers via monte carlo chain method. Genet. Sel. Evol. **32**: 467–482.

PONG-WONG, R., A. W. GEORGE, J. A. WOOLLIAMS, and C. S. HALEY, 2001 A simple and rapid method for calculating identity-by-descent matrices using multiple markers. Genet. Sel. Evol. **33**: 453–471.

QUAAS, R. L., 1988 Additive genetic model with groups and relationships. J. Dairy Sci. **71**: 1338–1345.

ROTHSCHILD, M. F., 1998 Identification of quantitative trait loci and interesting candidate genes in the pig: Progress and prospects. *In:* 6th WCGALP, volume 26, p 403–409.

ROTHSCHILD, M. F., and M. SOLLER, 1999 Candidate gene analysis to detect genes controlling traits of economic importance in domestic livestock. *In:* LOPES, P. S., R. F. EUCLYDES, R. A. TORRES, and S. E. F. GUIMARAES,, (Eds.), International symposium on animal breeding and genetics, p 219–242, Viqsa, Brazil.

SOBEL, E., and K. LANGE, 1996 Descent graphs in pedigree analysis: Applications to haplotyping location scores, and marker-sharing statistics. Am. J. Hum. Genet. **58**: 1323–1337.

SORENSEN, A. C., R. PONG-WONG, J. J. WINDIG, and J. A. WOOLLIAMS, 2002 Precision of methods for calculating identity-by-descent matrices using multiple markers. Genet. Sel. Evol. **34**: 557–579.

STRICKER, C., M. SCHELLING, F. DU, I. HOESCHELE, S. A. FERNÁNDEZ, and R. L. FERNANDO, 2002 A comparison of efficient genotype samplers for complex pedigrees and multiple linked loci. *In:* 7th WCGALP. CD ROM communication n.21-12.

THOMPSON, E. A., 1994 Monte carlo likelihood in genetic mapping. Statistical Science **9**: 355–366.

THOMPSON, E. A., 2000 Statistical inferences from genetic data on pedigrees. NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics, Beachwood, OH.

Totir, L. R., R. L. Fernando, J. C. M. Dekkers, S. A. Fernández, and B. Guldbrandtsen, 2003 Effect of using approximate gametic variance covariance matrices on marker assisted selection by blup. Genet Sel Evol (under submission).

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

Weller, J. I., and R. L. Fernando, 1991 *Gene-mapping techniques and applications*, chapter Strategies for the improvement of animal production using marker-assisted selection, p 305–328. Marcel Dekker Inc.

Chapter 3

A comparison of public domain programs for computing identical by descent coefficients.

A comparison of public domain programs for computing identical by descent coefficients.

Article submitted to the scientific journal "Genetics Selection Evolution".

ABSTRACT

The matrix of IBD coefficients among individuals must be calculated when variance component approaches are applied for Quantitative Trait Locus detection and/or Marker Assisted Selection. Many methods are available to calculate these coefficients. At this work we compared the performance, under horizontal and vertical dependence effects, of the methods implemented at SOLAR, which uses a regression method, LOKI, ESIP, both based on reverse peeling, and the MATVEC sampler, a segregation indicator sampler. SOLAR failed to estimate the correct coefficients at almost all situations simulated, while LOKI showed problems when vertical dependence was present in the pedigree. ESIP and MATVEC sampler performed well at all situations, therefore they are more indicated when genetic analysis are carried out on complex pedigree structures.

IBD / MCMC / sampling methodology / segregation indicator

RESUMO

A matriz de coeficientes IBD entre indivíduos para um QTL deve ser calculada quando procedimentos de componentes de variância são aplicados para a detecção de QTL e/ou seleção marcador assistida. Vários métodos estão disponíveis para o cálculo desses coeficientes, os quais diferem em sua habilidade em lidar com dependências horizontal e vertical entre os dados de marcador. Neste estudo, nós comparamos a habilidade em lidar com dependências horizontal e vertical dos métodos implementados no programa computacional SOLAR, que utiliza um método baseado em regressão linear, nos programas LOKI e ESIP, ambos baseados em "reverse peeling", e no amostrador implementado em MATVEC, um amostrador de indicadores de segregação. SOLAR falhou em estimar os coeficientes corretos em quase todas as situações simuladas, enquanto LOKI mostrou problemas quando dependência vertical estava presente no pedigree. ESIP e MATVEC apresentaram bons resultados para todas as situações simuladas, portanto ambos são mais indicados quando análises genéticas são conduzidas em pedigrees com estruturas complexas.

INTRODUCTION

Often, estimation of genetic parameters is based on modeling the covariance between relatives in terms of the unknown genetic parameters (Falconer and Mackay, 1996). This approach can also be used to estimate QTL parameters by modeling the genetic covariance between relatives at a QTL given pedigree and marker information (Chevalet *et al.*, 1984; Weller and Fernando, 1991; Grignola *et al.*, 1996). This covariance between two relatives, i and j, is proportional to Malecot's coefficient of relationship specific to the QTL, which is the conditional probability that a random QTL allele from individual i is identical-by-descent (IBD) to a random QTL allele from individual j given pedigree and marker information (Chevalet *et al.*, 1984; Fernando and Grossman, 1989).

When the marker information is limited to a single locus and genotypes are known for all parents, an efficient algorithm is available to recursively compute the QTL-specific IBD probabilities (Wang *et al.*, 1995). When marker genotypes are missing, this recursive algorithm does not give exact results (Wang *et al.*, 1995; Fernando, 1998). Further, when multiple markers are considered, to get exact results the recursive algorithm requires phase information in addition to genotypes for all individuals (Fernando, 1998; Hoeschele, 2001). In most QTL-mapping experiments, multiple markers are used, and missing genotypes and missing phase information are almost inevitable. Thus, alternative strategies have been proposed to compute these IBD probabilities. The most widely used of these have been developed in human genetics, where typically, family size is smaller and pedigrees are less complex than in livestock. It is expected that the performance of these methods will depend on the

pedigree structure. Thus, the objective of this paper is to examine the performance of some of these strategies for computing IBD probabilities in livestock pedigrees.

METHODS

Almasy and Blangero (Almasy and Blangero, 1998) proposed a deterministic method to compute QTL-specific IBD probabilities, and this approach has been implemented in SOLAR (Blangero *et al.*, 2000), a computer package for analysis of genetic data. In this approach, multiple regression is used to predict the IBD probability at a QTL using the IBD probabilities at markers as the independent variables. The IBD probability between two individuals at a marker is calculated using a pairwise likelihood based method when pedigrees are simple (Almasy and Blangero, 1998). When the pedigree is not simple, but marker genotypes are not missing, a recursive method (Davis *et al.*, 1996; Hoeschele, 2001) seems to be used to compute the exact IBD probability. When there are missing marker genotypes in a complex pedigree, an unspecified Monte Carlo method is used to obtain many samples of the missing marker genotypes, conditional on the observed markers and pedigree information. Then, exact IBD probabilities are calculated given each sample of the genotypes, and these probabilities are averaged, using the likelihood of the sampled genotypes as weights.

Markov Chain Monte-Carlo (MCMC) samplers are also being used to compute QTL-specific IBD probabilities. LOKI (Heath, 1997; Heath, 2002) is a very general package for analysis of genetic data. This package has a program to estimate QTL-

specific IBD probabilities by MCMC. LOKI uses reverse peeling (Heath, 1997; Kong, 1991) to draw samples from the joint distribution of genotypes of all members of the pedigree at each locus conditional on the pedigree and the observed genotypes at the current locus and the sampled genotypes at all the other loci. Thus, LOKI requires that the pedigree can be peeled (Heath, 1997).

ESIP (Fernández *et al.*, 2001) is a MCMC algorithm to draw samples from the joint posterior distribution of genotypes of all members of a large pedigree with loops. It has been used to estimate QTL-specific IBD probabilities by jointly sampling from the entire pedigree the genotypes at a QTL and the missing genotypes at flanking markers, conditional on the observed marker genotypes (Totir *et al.*, 2003). Although ESIP also uses reverse peeling, it does not require that the pedigree can be peeled. When peeling is not computationally feasible, ESIP draws a candidate sample from a modified pedigree that can be peeled (Wang *et al.*, 1996). Then, the candidate sample is accepted or rejected according to the Metropolis-Hastings algorithm (Hastings, 1970).

MATVEC (Kachman and Fernando, 2002) is a library of C++ classes that contains an MCMC algorithm to sample segregation indicators (Stricker *et al.*, 2002). The idea of sampling segregation indicators instead of genotypes was first proposed by Thompson (Thompson, 1994). The set of segregation indicators specifies the flow of alleles in a pedigree without specifying their states. Thus, Thompson argued that samplers of segregation indicators would be more efficient because of their smaller state space (Sobel and Lange, 1996). Sobel and Lange (Sobel and Lange, 1996) gave a Metropolis-Hastings (Hastings, 1970) algorithm to draw samples from the posterior distribution of segregation indicators for the marker loci given the marker and

pedigree information. The proposal they use samples segregation indicators for a randomly chosen individual at a randomly chosen locus. This algorithm has mixing problems when marker loci are closely linked (Stricker *et al.*, 2002). To address this problem, Stricker et al. (Stricker *et al.*, 2002) have proposed two modifications (Haplotype-sampler and Cascading Origin-sampler) to the sampler proposed by Sobel and Lange (SL Sampler) (Sobel and Lange, 1996). Samples are drawn using the original and the two modified samplers (Stricker *et al.*, 2002). In the Haplotype-sampler, the SL sampler is first used to sample some segregation indicators at a randomly chosen locus. Each of these segregation indicators specifies the grand-maternal or grand-paternal origin of an allele. Conditional on the origin of this first allele, the origin of the remaining alleles of each haplotype are sampled according the map distance between markers. In Cascanding Origin-sampler also, the SL sampler is first used to sample some segregation indicators at a randomly chosen locus. If a sampled segregation indicator switches the origin of an allele from one grand parent to the other, the origin of the remaining alleles are also switched according to the map distance between markers.

SIMULATIONS

It is well known that an MCMC sampler that is irreducible can give rise to a chain that is reducible "in practice" because of poor (slow) mixing. Poor mixing can occur due to horizontal or vertical dependence. Horizontal dependence refers to the dependence between genotypes at tightly linked loci within an individual, and the

vertical dependence refers to the dependence between genotypes at the same locus between individuals in large families (Heath, 1997).

To examine the effect of horizontal dependence on the IBD probabilities computed by the alternative algorithms, data were simulated for a QTL flanked by two markers either 2 cM, 5 cM or 10 cM apart. To examine the effect of vertical dependence on the IBD probabilities, the simulated pedigree (Pedigree 1) consisted of a typical F2 family with either 2 or 20 F2 offspring (Table I). To examine the consequences of missing marker genotypes, simulations were performed with or without genotype information for individual 5 (mother of the F2 offspring).

The scalar-Gibbs sampler, which is widely used in animal breeding, may be reducible when a marker locus with missing genotypes has more than two alleles (Sheehan and Thomas, 1993). However, it has been claimed that if at least one of the parents of each family is genotyped the scalar-Gibbs sampler would be irreducible. Recently this claim has been shown to be incorrect, using a counter example consisting of simple unlooped pedigree with three generations (Cannings and Sheehan, 2002). A pedigree similar to this counter example (Pedigree 2) was also simulated, with the flanking markers either 2 cM, 5 cM or 10 cM apart, to examine the consequences of missing genotypes (Table II).

Because of the simplicity of Pedigree 1 (no loops and marker allele segregation just for the maternal meioses of F2 generation), we were able to analytically calculate the exact IBD probabilities for all situations simulated. Because Pedigree 2 is not as simple as Pedigree 1, IBD probabilities estimated using 50 million samples from ESIP were treated as exact. The IBD probabilities provided by the alternative algorithms were compared to the exact IBD probabilities by calculating the average, the maxi-

Table I. Pedigree relationships, marker and QTL genotypes of the simulated F2 pedigrees (small and large F2 family).

Generation	ID	Mother	Father	M_1	M_2	QTL
Parent (breed A)	1	0	0	1,2	1,2	AA
Parent (breed A)	2	0	0	3,3	3,3	AA
Parent (breed B)	3	0	0	1,2	1,2	BB
Parent (breed B)	4	0	0	3,3	3,3	BB
F1 Cross	5	1	3	1,2	1,2	AB
F1 Cross	6	2	4	3,3	3,3	AB
F2	7	5	6	1,3	1,3	??
F2	8	5	6	1,3	2,3	??
F2	9	5	6	1,3	1,3	??
F2	10	5	6	1,3	1,3	??
F2	11	5	6	1,3	1,3	??
F2	12	5	6	1,3	1,3	??
F2	13	5	6	2,3	1,3	??
F2	14	5	6	2,3	2,3	??
F2	15	5	6	2,3	2,3	??
F2	16	5	6	2,3	2,3	??
F2	17	5	6	2,3	2,3	??
F2	18	5	6	2,3	2,3	??
F2	19	5	6	1,3	1,3	??
F2	20	5	6	1,3	1,3	??
F2	21	5	6	1,3	1,3	??
F2	22	5	6	1,3	1,3	??
F2	23	5	6	2,3	2,3	??
F2	24	5	6	2,3	2,3	??
F2	25	5	6	2,3	2,3	??
F2	26	5	6	2,3	2,3	??

Table II. Three generation pedigree where at least one in every parent pair has marker information (Pedigree 2).

Generation	ID	Mother	Father	M_1	M_2
0	1	0	0	1,1	1,2
0	2	0	0	-,-	1,2
1	3	0	0	2,3	1,2
1	4	2	1	-,-	2,2
2	5	4	3	2,3	1,2
2	6	4	3	1,3	1,2
2	7	4	3	1,3	1,2
2	8	4	3	1,3	1,2
2	9	4	3	1,3	1,2
2	10	4	3	1,3	1,2

mum, and the standard deviation of the absolute deviations among all individuals in the pedigree. For the three MCMC algorithms, the IBD probabilities were estimated from all the samples that could be generated in 20 minutes.

RESULTS AND DISCUSSION

Results for Pedigree 1 with all individuals genotyped for the markers are given in Table III. Although the marker genotypes were known for all individuals, the linkage phase between markers was not known in the F1 individuals. When the F2 family had only 2 offspring, all the algorithms performed well, and in this case, the distance between the genetic markers (2, 5 or 10 cM apart) had no effect on the IBD probabilities. However, when the F2 family had 20 offspring, SOLAR and LOKI gave incorrect results for some IBD probabilities. The largest deviation for SOLAR happened for the IBD probability calculation between the full sibs 8 and 13 (.31), while for LOKI it was among animals 1 and 3 (maternal grandparents) and the non-recombinants full sibs (7, 9, 10, 11, 12, 14,..., 26)(IBD deviations equal .25). In this case also, the distance between the genetic markers had no effect on the IBD probabilities.

As discussed below, the incorrect results from SOLAR may be an indication that the pairwise likelihood method used to calculate IBD probabilities for genetic markers does not capture the linkage-phase information from the other individuals in the pedigree. In the pedigree with 20 F2 offspring, the F1 mother (Individual 5) transmitted marker alleles A_1 and B_1 to nine offspring, A_2 and B_2 to another nine, A_1 and B_2 to

Table III. Average absolute deviation (standard deviation), and maximum deviation between exact IBD and those estimated by the algorithms for Pedigree 1, without missing marker information.

		Distance between genetic markers					
F2 generation size		2 cM		5 cM		10 cM	
		average(std)	maximum	average(std)	maximum	average(std)	maximum
2 full sibs	SOLAR	.0000(.0000)	.0000	.0000(.0000)	.0000	.0000(.0000)	.0000
	LOKI	.0000(.0000)	.0002	.0000(.0000)	.0001	.0000(.0000)	.0002
	ESIP	.0008(.0008)	.0018	.0008(.0009)	.0025	.0008(.0008)	.0022
	MATVEC	.0003(.0011)	.0063	.0003(.0007)	.0024	.0002(.0004)	.0012
20 full sibs	SOLAR	.0367(.0553)	.3117	.0335(.0507)	.3044	.0279(.0428)	.2907
	LOKI	.0257(.0802)	.2499	.0257(.0801)	.2497	.0255(.0798)	.2488
	ESIP	.0010(.0011)	.0029	.0009(.0010)	.0029	.0009(.0010)	.0028
	MATVEC	.0014(.0023)	.0105	.0015(.0021)	.0065	.0020(.0031)	.0135

one offspring (Individual 8), and A_2 and B_2 to another (Individual 13). Considering all this information jointly, individual 5 has linkage phase $\frac{A_1B_1}{A_2B_2}$ with probability of 99.9%. Thus, individuals 8 and 13 are recombinants with high probability. Given the QTL is at the center of these marker loci, the probability that Individuals 8 and 13 receive the same QTL allele from their mother is 0.5. But considering only the information from individuals 8 and 13, the phase of individual 5 is $\frac{A_1B_2}{A_2B_1}$ with high probability.

Thus, in the F2 pedigree used here, the information from the whole family was important to infer correctly the linkage phase of the mother (individual 5). However, SOLAR, which uses pairwise information, would get incorrect phase information for individual 5 when computing the IBD between the two recombinant offspring, individuals 8 and 13. When the family consists of only two individuals, the pairwise method uses all the available information and gives correct results (Table III).

For this same pedigree, but without marker information for individual 5, when the F2 family had only 2 offspring, all the algorithms except SOLAR performed well (Table IV). These results indicate that the Monte-Carlo method used by SOLAR to infer IBD probabilities when some marker genotypes are missing may give incorrect results.

When the F2 family had 20 offspring, ESIP and MATVEC continued to perform well. Here, SOLAR and LOKI gave incorrect IBD probabilities for the same pairs of relatives as in the case without missing genotypes, but the results were not affected by the distance between flanking markers.

Both LOKI and ESIP draw samples from the joint distribution of genotypes of all members of the pedigree. However, LOKI samples genotypes at each locus conditional

Table IV. Average absolute deviation (standard deviation), and maximum deviation between exact IBD and those estimated by the algorithms for Pedigree 1, with missing marker information.

		Distance between genetic markers					
F2 generation size		2 cM		5 cM		10 cM	
		average(std)	maximum	average(std)	maximum	average(std)	maximum
2 full sibs	SOLAR	.0044(.0263)	.1577	.0041(.0243)	.1461	.0036(.0219)	.1313
	LOKI	.0003(.0016)	.0096	.0006(.0036)	.0216	.0010(.0062)	.0372
	ESIP	.0009(.0018)	.0103	.0012(.0037)	.0220	.0017(.0063)	.0379
	MATVEC	.0000(.0002)	.0007	.0001(.0003)	.0018	.0004(.0016)	.0096
20 full sibs	SOLAR	.0367(.0553)	.3117	.0335(.0507)	.3044	.0279(.0428)	.2906
	LOKI	.0257(.0802)	.2499	.0256(.0801)	.2497	.0255(.0798)	.2488
	ESIP	.0009(.0011)	.0029	.0009(.0010)	.0029	.0009(.0010)	.0028
	MATVEC	.0026(.0051)	.0185	.0012(.0017)	.0064	.0019(.0029)	.0102

on the sampled genotypes at the other loci, while ESIP jointly samples the genotypes at the flanking markers and the QTL. Thus, results from LOKI can be affected by horizontal dependence. The incorrect results from LOKI for Pedigree 1 when the F2 family had 20 offspring disappeared when the interval between markers was increased to 40 cM (results not shown). It is worth noting that horizontal dependence was not a problem when the family size was small. Thus, this shows that large families can contribute to horizontal dependence.

For Pedigree 2 (Table II), where at least one in every parent pair had marker information, SOLAR gave incorrect results for IBD probabilities between the grandparents (individuals 1 and 2) and grand-offspring, and between individuals of generation 2 (Table V). LOKI and MATVEC produced good estimates of IBD probabilities, independently of the distance between markers (Table V).

For all situations studied here, MATVEC produced good results. The results from pedigrees with large families indicates that the composite transition rules proposed by Sobel and Lange (Sobel and Lange, 1996) were effective in avoiding vertical dependence due to large families. Further, the results with tightly linked markers indicates that the extensions proposed by Stricker (Stricker *et al.*, 2002) were effective in avoiding horizontal dependence.

CONCLUSIONS

Although SOLAR is very fast, it seems to be inappropriate for genetic analysis of complex pedigrees like those commonly found at livestock populations. These

Table V. Average absolute deviation (standard deviation), and maximum deviation between IBD probabilities calculated by ESIP and those estimated by the algorithms, for Pedigree 2.

Distance between markers		average(std)	maximum
2 cM	SOLAR	.0477(.0721)	.1202
	LOKI	.0001(.0000)	.0002
	MATVEC	.0005(.0008)	.0022
5 cM	SOLAR	.0444(.0663)	.1131
	LOKI	.0001(.0000)	.0002
	MATVEC	.0021(.0032)	.0076
10 cM	SOLAR	.0395(.0580)	.1019
	LOKI	.0000(.0000)	.0002
	MATVEC	.0028(.0042)	.0101

pedigrees typically have large families (e.g., swine and poultry populations or MOET nucleus) and may provide valuable information to compute IBD probabilities. Also, in these populations, missing marker information is common, and thus it is essential to use methods that can accommodate missing marker information. The poor results from SOLAR for the Pedigree 2, which has a typical structure found it livestock populations, where the sires are fully typed for markers but the dams are not, indicates that it may not be appropriate for livestock populations.

For the pedigrees studied here, the only difference between LOKI and ESIP is due to horizontal dependence. Because LOKI samples genotypes at each locus conditional on the genotypes at other loci, its results are subject to horizontal dependence. ESIP samples all genotypes jointly and is not affected by horizontal dependence. We have seen here that large families can contribute to horizontal dependence. Thus, samplers, such as that in LOKI, that are subject to horizontal dependence may have problems in livestock pedigrees where large families are common.

Although ESIP did not have any problems with the pedigrees studied here, sampling genotypes jointly will not be feasible with many loci in large pedigrees with many missing genotypes. Even in such situations, the sampler of segregation indicators implemented in MATVEC can be used to compute IBD probabilities. It should be noted, however, that this sampler requires a good initial sample of segregation indicators (Sobel and Lange, 1996). In this study, ESIP was used to obtain this initial sample.

LITERATURE CITED

ALMASY, L., and J. BLANGERO, 1998 Multipoint quantitative trait linkage analysis in general pedigrees. Am. J. Hum. Genet. **62**: 1198–1211.

BLANGERO, J., K. LANGE, L. ALMASY, J. WILLIAMS, T. DYER, and C. PETERSON, 2000 *Solar Manual.*

CANNINGS, C., and N. A. SHEEHAN, 2002 On a misconception about irreducibility of the single-site gibbs sampler in a pedigree application. Genetics **162**: 993–996.

CHEVALET, C., M. GILLOIS, and J. VU TIEN KHANG, 1984 Conditional probabilities of identity of genes at a locus linked to a marker. Genet. Sel. Evol. **16**: 431–444.

DAVIS, S., M. SCHROEDER, L. R. GOLDIN, and D. E. WEEKS, 1996 Nonparametric simulation-based statistics for detecting linkage in general pedigrees. Am. J. Hum. Genet. **58**: 867–880.

FALCONER, D. S., and T. F. C. MACKAY, 1996 *Introduction to quantitative genetics.* Longman, Edinburgh.

FERNÁNDEZ, S. A., R. L. FERNANDO, B. GULDBRANDTSEN, L. R. TOTIR, and A. L. CARRIQUIRY, 2001 Sampling genotypes in large pedigrees with loops. Genet. Sel. Evol. **33**: 337–367.

FERNANDO, R. L., 1998 Genetic evaluation and selection using genotypic, phenotypic and pedigree information. *In:* 6th WCGALP, volume 26, p 329–336.

FERNANDO, R. L., and M. GROSSMAN, 1989 Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. **21**: 467–477.

GRIGNOLA, F. E., I. HOESCHELE, and B. TIER, 1996 Mapping quantitative trait loci in outcross populations via residual maximum likelihood. i methodology. Genet. Sel. Evol. **28**: 479–490.

HASTINGS, W. K., 1970 Monte carlo sampling methods using markov chain and their applications. Biometrika **57**: 97–109.

HEATH, S. C., 1997 Markov chain monte carlo segregation and linkage analysis for oligogenic models. Am. J. Hum. Genet. **61**: 748–760.

HEATH, S. C., 2002 *A package for multipoint linkage analysis on large pedigree using reversible jump markov chain monte carlo.*

HOESCHELE, I., 2001 *Handbook of statistical genetics*, chapter Mapping quantitative trait loci in outbred pedigrees, p 599–639. John Wiley & Sons Ltd.

KACHMAN, S. D., and R. L. FERNANDO, 2002 Analysis of generalized linear mixed models with matvec. *In:* 7th WCGALP. CD ROM communication n.28-4.

KONG, A., 1991 Analysis of pedigree data using methods combining peeling and gibbs sampling. *In:* 23rd Symposium on the Interface, p 379–385.

SHEEHAN, N., and A. THOMAS, 1993 On the irreducibility of a markov chain defined on a space of genotype configurations by a sampling scheme. Biometrics **49**: 163–175.

SOBEL, E., and K. LANGE, 1996 Descent graphs in pedigree analysis: Applications to haplotyping location scores, and marker-sharing statistics. Am. J. Hum. Genet. **58**: 1323–1337.

Stricker, C., M. Schelling, F. Du, I. Hoeschele, S. A. Fernández, and R. L. Fernando, 2002 A comparison of efficient genotype samplers for complex pedigrees and multiple linked loci. *In:* 7th WCGALP. CD ROM communication n.21-12.

Thompson, E. A., 1994 Monte carlo likelihood in genetic mapping. Statistical Science **9**: 355–366.

Totir, L. R., R. L. Fernando, J. C. M. Dekkers, S. A. Fernández, and B. Guldbrandtsen, 2003 Effect of using approximate gametic variance covariance matrices on marker assisted selection by blup. Genet Sel Evol (under submission).

Wang, T., R. L. Fernando, C. Stricker, and R. C. Elston, 1996 An approximation to the likelihood for a pedigree with loops. Theor. Appl. Genet. **93**: 1299–1309.

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

Weller, J. I., and R. L. Fernando, 1991 *Gene-mapping techniques and applications*, chapter Strategies for the improvement of animal production using marker-assisted selection, p 305–328. Marcel Dekker Inc.

Chapter 4

An improved approximation of the gametic covariance matrix for marker assisted genetic evaluation by BLUP.

An improved approximation of the gametic covariance matrix for marker assisted genetic evaluation by BLUP.

Article submitted to the scientific journal "Genetics Selection Evolution".

ABSTRACT

Marker Assisted Best Linear Unbiased Prediction (MABLUP) with large livestock pedigrees depends critically on efficient algorithms to invert the gametic covariance matrix for the marked QTL (G_v). These algorithms are based on tracing the QTL alleles of an individual to its grandmother or grandfather. When marker information is complete, there are simple rules to compute the probabilities (PDQ's) of these events. When marker information is incomplete, even if the PDQ's are computed exactly, the efficient algorithms would be approximate. Further, computing the exact PDQ's becomes difficult when marker information is incomplete. In this study, we examined the effect of estimating the PDQ's by MCMC on response to MABLUP. In a previous study, use of an analytical approximation of PDQ's was shown to result in a loss of about 50% of the response to MABLUP. The results of this study show that by estimating the PDQ's by MCMC this loss could be reduced to about 20%. Further, response to MABLUP was greater when four bi-allelic markers were used, and the loss in response due to missing markers was smaller in the case with four markers compared to when only two bi-allelic markers were used.

marker assisted BLUP / MCMC sampler / gametic covariance matrix / response to selection

RESUMO

A melhor predição linear não-viesada assistida por marcadores (MABLUP) em grandes populações de animais domésticos depende muito de um algoritmo eficiente para inverter a matriz de covariância gamética para um QTL marcado (G_v). Esses algoritmos são baseados no rastreamento dos alelos QTL do indivíduo em relação aos de seus avós. Quando a informação de marcador é completa existem regras simples para o cálculo da probabilidades desses eventos (PDQ's). Quando a informação de marcador é incompleta, até mesmo se as PDQ's são calculadas de maneira exata os algoritmos eficientes ainda serão uma aproximação. Ainda, o cálculo exato das PDQ's torna-se difícil quando a informação de marcador é incompleta. Nesse estudo, nós examinamos o efeito da estimação das PDQ's por MCMC sobre a resposta a MABLUP. Em um estudo anterior, o uso de uma aproximação analítica para o cálculo das PDQ's resultou em uma perda de aproximadamente 50% da reposta à seleção obtida por meio de MABLUP. Os resultados desse estudo mostram que estimando-se as PDQ's por MCMC essa perda pode ser reduzida a aproximadamente 20%. Ainda, a resposta por MABLUP foi maior e a perda em resposta devido ao uso de marcadores sem informação foi menor quando quatro marcadores bi-alélicos foram usados, em comparação quando utilizou-se apenas dois marcadores bi-alélicos.

INTRODUCTION

Until recently, phenotypic and pedigree data were the only sources of genetic information for Best Linear Unbiased Prediction (BLUP) in livestock populations. But, due to advances in molecular biology over the last two decades, genetic marker data have become an additional source of information for BLUP. Genetic markers can be used to trace the inheritance of alleles at Quantitative Trait Loci (QTL) (De Koning *et al.*, 2001; Malek *et al.*, 2001). Fernando and Grossman (Fernando and Grossman, 1989) and Goddard (Goddard, 1992) showed how to incorporate marker information in BLUP. BLUP using marker information, in addition to phenotypic and pedigree information, will be referred to as MABLUP. Solving Henderson's Mixed Model Equations (HMME) (Henderson, 1984) for MABLUP gives BLUP of the gametic values for the MQTL and of the additive genotypic value for the remaining QTL (RQTL). It is expected that MABLUP increases the accuracy of prediction, and, as a result, improves response to selection (Weller and Fernando, 1991).

To construct Henderson's Mixed Model equations (HMME) for MABLUP it is necessary to invert the gametic covariance matrix, and efficient algorithms have been developed for this purpose (Fernando and Grossman, 1989; Goddard, 1992; van Arendonk *et al.*, 1994; Wang *et al.*, 1995; Abdel-Azim and Freeman, 2001). Unfortunately, these algorithms require complete marker information, and in large livestock pedigrees incomplete marker information is almost unavoidable. This is especially a problem when flanking markers are used for MABLUP. In this case, it is not sufficient to know the genotypes of the flanking markers but the linkage phase between the markers also needs to be known (Hoeschele, 2001). The efficient algorithms to invert the gametic

covariance matrix are based on a recursive equation, equation 4 in Fernando and Grossman (1989) or equation 3 in Wang *et al.* (1995), for computing the conditional probability that MQTL alleles are identical by descent (IBD) given marker information. For these recursive calculations it is necessary to compute the probability that a QTL allele originated in the grandmother or grandfather conditional on the observed marker information. In the MABLUP literature, these probabilities are referred to as probabilities of descent for QTL alleles (PDQ).

When complete marker information is available at a single marker, Wang et al. (Wang *et al.*, 1995) gave simple rules to compute the PDQ's. When marker information is missing, computing the exact PDQ's becomes difficult, and approximations have been used (Pong-Wong *et al.*, 2001; Totir *et al.*, 2003). The consequence of two such approximations on response to selection have been studied, and results show that using an approximate inverse reduced the additional response due to MABLUP by as much as 100% for one of the methods (Method A) and as much as 50% for the other (Method B) (Totir *et al.*, 2003). In both these approximations the recursive equation for IBD probabilities was used. Even if exact PDQ's are used in this recursive equation, if marker information is not complete, the inverse will not be exact. Further, in methods A and B the PDQ's were also approximated. In Method A, marker information was used only when the linkage phase between the markers was known, otherwise the marker information was ignored and the PDQ's were set to 0.5. In Method B, if the linkage phase between the flanking marker was unknown, the genotype information at just one of the two markers was used to compute the PDQ's. Finally, when both genotypes at both markers were unobserved, the PDQ's were set to 0.5. Thus, in methods A and B, PDQ's were based only on marker information of

the individual and its parents; information from relatives was not used to compute the PDQ's.

Markov Chain Monte Carlo (MCMC) methods can be used to estimate PDQ's using all the marker information in the pedigree. Even in this case, as explained above, if marker information is not complete the inverse will still not be exact. The objective of this paper is to examine the effect of using PDQ's estimated by MCMC on the additional response to selection by MABLUP. It is hoped that because the PDQ's are estimated using all available information, the inverse obtained using the efficient algorithm would be more accurate and the loss in the additional response to MABLUP would be reduced.

USE OF MCMC TO ESTIMATE PDQ'S

Consider the paternal MQTL allele Q_i^p of individual i. This allele is either the paternal allele Q_s^p or the maternal allele Q_s^m of its sire s. The segregation indicator S_i^p for MQTL allele Q_i^p is a variable that specifies its origin: if Q_i^p was inherited from Q_s^p then $S_i^p = 1$, and if Q_i^p was inherited from Q_s^m then $S_i^p = 0$. Thus, the PDQ's for Q_i^p are $\Pr(S_i^p = 1)$ and $\Pr(S_i^p = 0)$. Suppose we can draw samples from the conditional distribution of the segregation indicators for the MQTL alleles given the observed marker data. Then, the PDQ's can be estimated from these samples. Based on an idea proposed by Thompson (Thompson, 1994), Sobel and Lange (Sobel and Lange, 1996) developed an MCMC algorithm to draw samples of segregation indicators for marker loci from their conditional distribution given the observed marker data; this

will be referred to as the SL sampler. Segregation indicators for the MQTL can be sampled from its conditional distribution given marker data at linked loci by including the MQTL in the algorithm as a marker locus without any genotype information.

Although the SL sampler is known to produce an irreducible chain, when loci are tightly linked results may not be reliable due to mixing problems (Stricker *et al.*, 2002). In order to overcome this problem, Stricker et al. (Stricker *et al.*, 2002) proposed two modifications to the SL sampler; these were called the Haplotype sampler and the Cascading-Origin sampler. In this paper, we used a combination of the SL sampler and the two modified samplers to estimate PDQ's. In the following sections we describe these samplers.

The SL sampler

This sampler uses a Metropolis algorithm to sample the segregation indicators. In the Metropolis algorithm, a candidate sample is drawn from some proposal distribution, and this candidate is accepted or rejected according to the Metropolis acceptance probability (Thompson, 1994). The proposal used in the SL sampler chooses a candidate sample that is "close" to the current sample in the space of segregation indicators by modifying a few of the segregation indicators in the current sample. Sobel and Lange (Sobel and Lange, 1996) provided four transition rules to choose these candidate samples given the current sample.

In all four rules, the transition is applied to a randomly chosen locus. The first rule (T_0) involves a single randomly chosen individual. In this individual, at the ran-

domly chosen locus there are two segregation indicators: the paternal and maternal indicators. One of these is randomly chosen and its value is switched to its complement, i.e., if the segregation indicator has value 0 it is switched to 1 and if it has value 1 it is switched to 0. The remaining three are composite rules (T_1, $T2_a$ and $T2_b$) involving several related individuals. Rule T_1 involves all the offspring of a randomly chosen parent. If this parent is a father, the paternal segregation indicator is switched to its complement in all its offspring, and if this parent is a mother, the maternal segregation indicator is switched to its complement in all its offspring. Rules $T2_a$ and $T2_b$ involve a pair of randomly chosen parents (i and j), their common offspring and grand-offspring. In rule $T2_a$, for each offspring, if the two segregation indicators do not have the same value, they are switched to their complements. In $T2_b$, for each offspring, if the two segregation indicators have the same value, they are switched to their complements. Further, in both $T2_a$ and $T2_b$, any segregation indicator in the grand-offspring that traces back to i or j is switched to its complement. Sobel and Lange (Sobel and Lange, 1996) showed that if the proposal consists of only a single transition, the resulting chain may not be irreducible. To overcome this problem, their proposal consists of a random number of transitions.

The Haplotype Sampler

In the SL-sampler, the candidate sample is obtained by modifying the segregation indicators at a single locus. This leads to poor mixing when loci are tightly linked. To avoid this problem, the proposal used in the Haplotype sampler modifies segrega-

tion indicators at all the loci. The first step in the Haplotype sampler is to modify the current sample by applying the proposal of the SL sampler. For each haplotype where a segregation indicator was modified in the first step, the remaining segregation indicators are also modified as explained below. Suppose for example, in some haplotype, the segregation indicator at locus j was modified to the value 0. Then, the segregation indicator at locus $j+1$ will also be assigned a value of 0 with probability $(1-r_j)$ or a value 1 with probability r_j, where r_j is the probability of a recombination between loci j and $j+1$. Similarly, conditional on the value assigned to the segregation indicator for locus $j+1$, a value is assigned to the segregation indicator at locus $j+2$. This process is repeated to modify all the segregation indicators on the haplotype. Unfortunately, the resulting proposal is not symmetric as in the SL sampler, and thus we use the Metropolis-Hastings acceptance probability to accept or reject the candidate sample.

The Cascading-Origin Sampler

The Haplotype sampler may also show poor mixing for tightly linked loci when recombinant haplotypes are present in the data (Stricker *et al.*, 2002). The reason is that when the Haplotype sampler modifies the segregation indicators for a recombinant haplotype, the recombinations will be removed with high probability. Such a candidate will be rejected with high probability, and the chain will not move. The goal of the Cascading-Origin sampler is to sample all the segregation indicators on a haplotype while preserving recombinations. As in the Haplotype sampler, the first

step here is to modify the current sample by applying the proposal of the SL sampler. For each haplotype where a segregation indicator was modified in the first step, the remaining segregation indicators are also modified as explained below. Suppose for example, in some haplotype, the segregation indicator at locus j was switched to its complement. Then, the segregation indicator at locus $j+1$ will also be switched to its complement with probability $(1-r_j)$ or not switched with probability r_j, where r_j is the probability of a recombination between loci j and $j+1$. Similarly, the segregation indicator at locus $j+2$ will be switched or remain unswitched conditional on if the segregation indicator for locus $j+1$ was switched or not. This process is repeated for all the segregation indicators on the haplotype. In this case the proposal is symmetric and the Metropolis acceptance probability is used to accept or reject the candidate sample.

The combination of these three samplers (hereafter named HT-CO sampler) is expected to be more efficient with tightly linked loci than the original approach proposed by Sobel and Lange (Sobel and Lange, 1996). Thus, in this study, for each set of 30 samples the first 10 were obtained using the SL sampler, the next 10 using the Haplotype sampler, and the last 10 using the Cascading-Origin sampler.

Simulations

The pedigree (Figure 4.1) used by Totir et al. (2003) to study the effect of PDQ approximations A and B on MABLUP response will also be used here to study the effect of PDQ's estimated by MCMC on MABLUP. This pedigree consisted of 96

individuals from four generations, and each nuclear family in the pedigree had 10 offspring. The simulated trait had a heritability of 10%, where 2.85% of the genetic variance was due to a bi-allelic MQTL and 7.15% was due to the RQTL. Two or four bi-allelic markers were simulated flanking the MQTL. A recombination rate of 5% was assumed among all three or five loci (the two flanking markers and the MQTL or the four flanking markers and the MQTL). In order to evaluate the effect of missing marker genotypes upon PDQ calculations, individuals 1, 2, 3, 14, and 15 were simulated with or without marker information. Five thousand replications of the simulation were performed for each of the four combinations of these two experimental factors. For each replication, PDQ's were calculated from 50 000 HT-CO samples. Then these PDQ's were used to compute efficiently the inverse of the gametic covariance matrix. This inverse was used in HMME to obtain MABLUP. Individuals 47 to 96 were simulated without trait phenotypes and five of these were selected according to their predicted genotypic values by BLUP or MABLUP.

Because the pedigree in this study was small and it did not have complex loops, samples could be drawn from the conditional distribution of the MQTL genotypes given the observed marker information. These samples could be used to estimate the gametic covariance matrix to any degree of approximation. In this study, the gametic covariance matrix was estimated from 15 000 independent samples of the MQTL genotypes obtained using ESIP (Fernández *et al.*, 2001). For this small pedigree this matrix could be inverted directly. When this inverse of the gametic covariance matrix is used in the HMME, the maximum possible response to selection from MABLUP is expected.

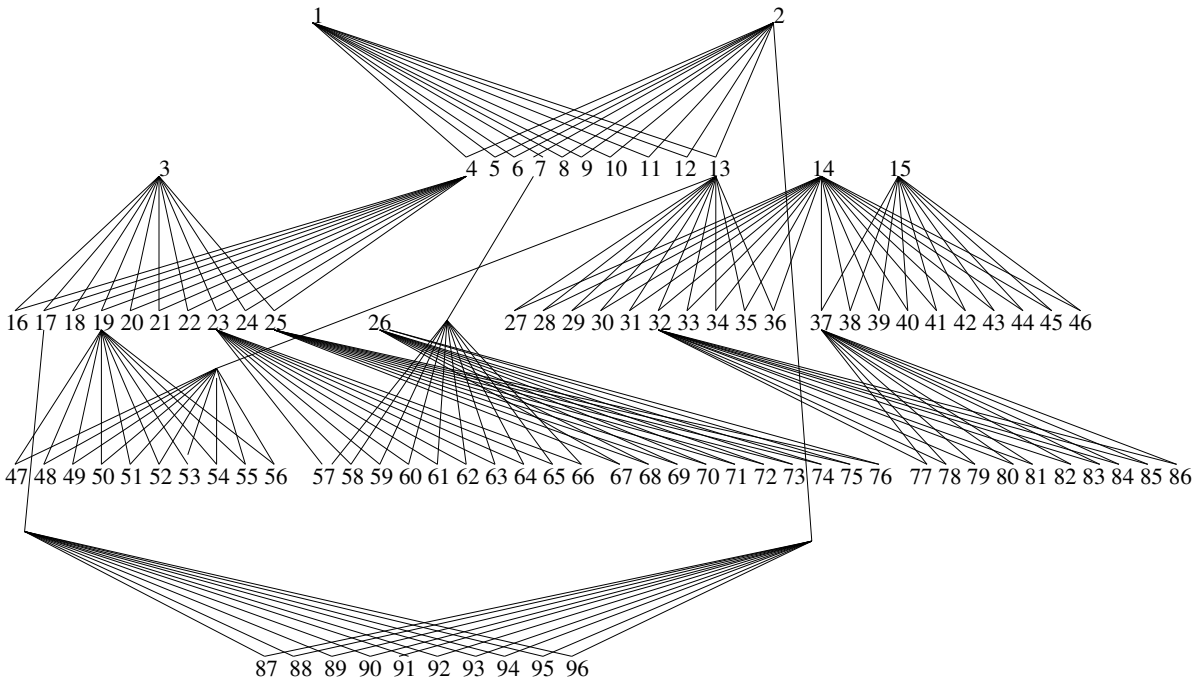

Fig. 4.1. Pedigree simulated.

The superiority of MABLUP was calculated as

$$\frac{R_{MABLUP} - R_{BLUP}}{R_{BLUP}} \times 100$$

where, R_{MABLUP} is the realized response to selection using MABLUP and R_{BLUP} is the realized response to selection using BLUP. This superiority was calculated when the inverse of the gametic covariance was constructed using the efficient algorithm based on PDQ's and when the gametic covariance was inverted directly. The effect of PDQ's estimated by MCMC on MABLUP was studied by comparing these calculations.

RESULTS AND DISCUSSION

Results for the pedigree without missing marker genotypes are given in Figure 4.2. When the inverse of the gametic covariance matrix conditional on two flanking markers was constructed using the efficient algorithm based on estimated PDQ's, response to selection was 19.2% smaller than when the inverse was obtained directly. Here, the PDQ's used in the efficient algorithm were estimated by MCMC using all the marker information in the pedigree. For this same simulation, when the inverse was obtained efficiently based on an analytical approximation of the PDQ's, response to selection was 50% smaller than when the inverse was obtained directly (Totir *et al.*, 2003). Totir et al. (Totir *et al.*, 2003) observed that this loss in response is due to 1) the approximation in the PDQ's and 2) the efficient algorithm used to invert the gametic covariance matrix is based on a recursive equation, which is exact only with complete marker data. Because the PDQ's in this study were estimated accurately by MCMC using all available marker information, the approximation in the inverse

is almost entirely due to use of the efficient algorithm. The small loss in response (19.2%) that we observed here indicates that the approximation introduced through the recursive equation is small.

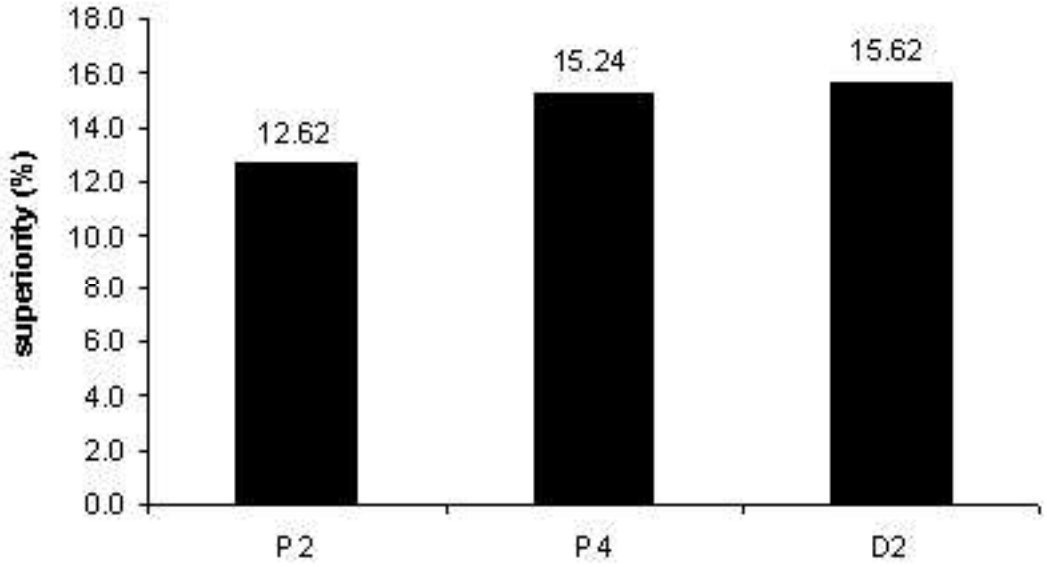

Fig. 4.2. Superiority of MABLUP over BLUP without missing marker genotypes in the pedigree. P2 and P4 are the results when PDQ's conditional on 2 or 4 flanking markers were used to compute efficiently the inverse of the gametic covariance matrix. D2 is the result when the gametic covariance matrix conditional on 2 markers was inverted directly.

When more marker loci were considered in the analysis (four bi-allelic flanking markers), response to selection using the efficient algorithm was only 2.4% smaller than with direct inversion with two markers. Unfortunately, at present, ESIP can

sample only three loci jointly (two flanking markers and one MQTL locus), and so we could not calculate the response to selection with direct inversion conditional on four marker loci. Although direct inversion could be used for MABLUP with the small pedigree used in this study, in most real applications this would not be possible. However, even with a big pedigree, the efficient algorithm can be used to invert the gametic covariance matrix conditional on four markers, and this gave almost the same response as direct inversion with two loci.

Results for the pedigree with missing marker genotypes are given in Figure 4.3. Comparing these results with those in Figure 4.2 shows that the effect of the missing genotypes was largest when the inverse was constructed using the efficient algorithm based on estimated PDQ's conditional on two markers (reduction of 3.2% over the response observed without missing marker genotypes). This effect was smallest when the inverse was constructed directly (reduction of 0.4% over the response observed without missing marker genotypes). When four markers were used, response to selection using the efficient algorithm was 7.4% smaller than with direct inversion with two markers. It has been shown that use of markers that are more polymorphic results in greater response to MABLUP (Totir *et al.*, 2003). This study shows that a similar result can be achieved by using more bi-allelic loci.

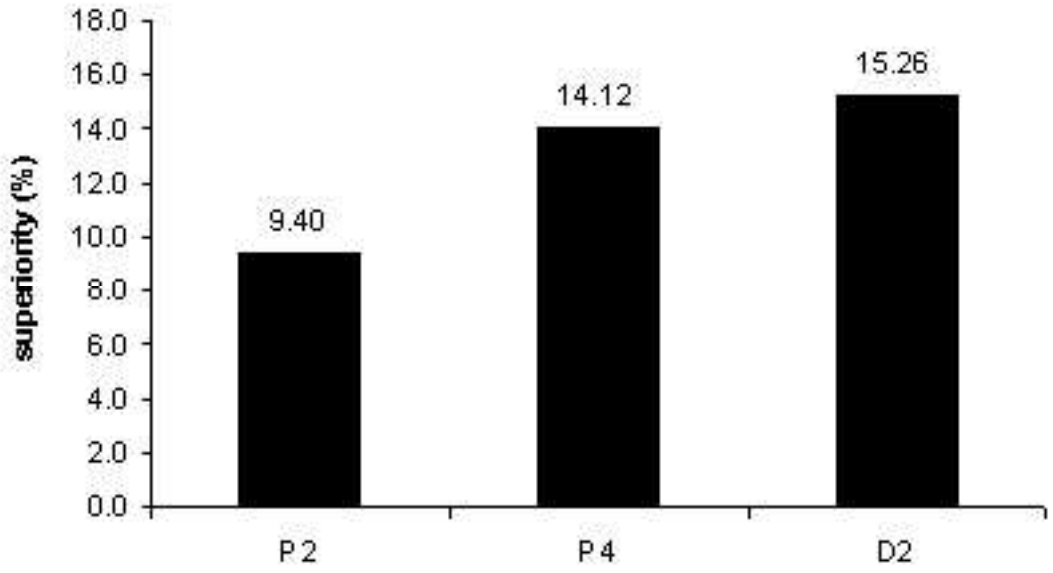

Fig. 4.3. Superiority of MABLUP over BLUP with missing marker genotypes in the pedigree.P2 and P4 are the results when PDQ's conditional on 2 or 4 flanking markers were used to compute efficiently the inverse of the gametic covariance matrix. D2 is the result when the gametic covariance matrix conditional on 2 markers was inverted directly.

CONCLUSIONS

The efficient algorithm for inverting the gametic covariance matrix gives good results when the PDQ's are well approximated by MCMC. Therefore, this algorithm can be applied to efficiently perform MABLUP in large livestock populations.

LITERATURE CITED

ABDEL-AZIM, G., and A. E. FREEMAN, 2001 A rapid method for computing the inverse of the gametic covariance matrix between relatives for a marked quantitative trait locus. Genet. Sel. Evol. **33**: 153–173.

DE KONING, D., L. L. G. JANSS, A. P. RATTINK, P. A. M. VAN OERS, B. J. DE VRIES, and ET AL., 2001 Detection of quantitative trait loci for backfat thickness and intramuscular fat content in pigs (*Sus Scrofa*). Genetics **152**: 1679–1690.

FERNÁNDEZ, S. A., R. L. FERNANDO, B. GULDBRANDTSEN, L. R. TOTIR, and A. L. CARRIQUIRY, 2001 Sampling genotypes in large pedigrees with loops. Genet. Sel. Evol. **33**: 337–367.

FERNANDO, R. L., and M. GROSSMAN, 1989 Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. **21**: 467–477.

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

HENDERSON, C. R., 1984 Applications of linear models in animal breeding. University of Guelph.

HOESCHELE, I., 2001 *Handbook of statistical genetics*, chapter Mapping quantitative trait loci in outbred pedigrees, p 599–639. John Wiley & Sons Ltd.

MALEK, M., J. C. M. DEKKERS, H. K. LEE, T. J. BAAS, and M. F. ROTHSCHILD, 2001 A molecular genome scan analysis to identify chromosomal regions

influencing economic traits in the pig. i.growth and body composition. Mammalian Genome **12**: 630–636.

Pong-Wong, R .and George, A. W., J. A. Woolliams, and C. S. Haley, 2001 A simple and rapid method for calculating identity-by-descent matrices using multiple markers. Genet. Sel. Evol. **33**: 453–471.

Sobel, E., and K. Lange, 1996 Descent graphs in pedigree analysis: Applications to haplotyping location scores, and marker-sharing statistics. Am. J. Hum. Genet. **58**: 1323–1337.

Stricker, C., M. Schelling, F. Du, I. Hoeschele, S. A. Fernández, and R. L. Fernando, 2002 A comparison of efficient genotype samplers for complex pedigrees and multiple linked loci. *In:* 7th WCGALP. CD ROM communication n.21-12.

Thompson, E. A., 1994 Monte carlo likelihood in genetic mapping. Statistical Science **9**: 355–366.

Totir, L. R., R. L. Fernando, J. C. M. Dekkers, S. A. Fernández, and B. Guldbrandtsen, 2003 Effect of using approximate gametic variance covariance matrices on marker assisted selection by blup. Genet Sel Evol (under submission).

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

Wang, T., R. L. Fernando, S. Van der Beek, M. Grossman, and J. A. M. Van Arendonk, 1995 Covariance between relatives for a marked quantitative

locus. Genet. Sel. Evol. **27**: 251–274.

WELLER, J. I., and R. L. FERNANDO, 1991 *Gene-mapping techniques and applications*, chapter Strategies for the improvement of animal production using marker-assisted selection, p 305–328. Marcel Dekker Inc.