

# Mapping Quantitative Trait Loci in Complex Pedigrees: A Two-Step Variance Component Approach

Andrew W. George,\* Peter M. Visscher<sup>†</sup> and Chris S. Haley\*

\*Roslin Institute, Midlothian EH25 9PS, United Kingdom and <sup>†</sup>Institute of Cell, Animal and Population Biology, Edinburgh EH9 3JG, Scotland

Manuscript received May 12, 2000  
Accepted for publication July 27, 2000

## ABSTRACT

There is a growing need for the development of statistical techniques capable of mapping quantitative trait loci (QTL) in general outbred animal populations. Presently used variance component methods, which correctly account for the complex relationships that may exist between individuals, are challenged by the difficulties incurred through unknown marker genotypes, inbred individuals, partially or unknown marker phases, and multigenerational data. In this article, a two-step variance component approach that enables practitioners to routinely map QTL in populations with the aforementioned difficulties is explored. The performance of the QTL mapping methodology is assessed via its application to simulated data. The capacity of the technique to accurately estimate parameters is examined for a range of scenarios.

WITH the widespread usage of genetic markers in helping to detect and localize quantitative trait loci (QTL), marker data are becoming available on human and livestock populations with increasingly complex pedigree structures. QTL analysis in such populations is challenging because the number of alleles segregating at the QTL is unknown, the marker phases may be unknown or only partially known, the marker and QTL allele frequencies must be estimated from the data, inbreeding loops may exist in the pedigree, and markers may be noninformative or ungenotyped. Although it is possible to simplify the analysis of complex pedigree data by fragmenting the pedigree into smaller component families, methods that fully account for complex relationships between individuals are expected to provide greater power to detect QTL (ALMASY and BLANGERO 1998).

Literature surrounding the mapping of QTL in populations with complex pedigrees can be classified according to the allelic assumptions associated with the QTL. Mapping methods either assume the QTL is a fixed effect with a finite number of alleles or a random effect with an infinite number of alleles.

Analysis of statistical models that treat the QTL as a fixed effect range from simple regression-based methodologies (KNOTT *et al.* 1996) to complex statistical analyses involving Markov chain Monte Carlo (MCMC) methods within frequentist (HEATH 1997; JANSEN *et al.* 1998) and Bayesian (UIMARI and HOESCHELE 1997; GEORGE *et al.* 2000) paradigms. The statistical models are mixture distributions, where the number of component densi-

ties is determined by the number of QTL genotypes, and assumptions regarding the total number of segregating alleles have a profound effect on the formulation of the statistical model.

Random effects models, based upon the simple premise that individuals of like phenotype are more likely to share genes identical-by-descent (IBD), offer a less parameterized statistical environment in which to map QTL. This environment is obtained by assuming the QTL effects are normally distributed—an assumption that circumvents the estimation of QTL allele frequencies and is robust to violation (HOESCHELE *et al.* 1997).

Random effects models have long been utilized by human geneticists interested in partitioning the genetic variance of quantitative traits into effects due to specific chromosomal regions. As early as the 1970s, variance component approaches (*i.e.*, analytical methods that estimate the parameters of random effects models) were being used to detect QTL in phase-known pedigrees (JAYAKAR 1970). Since then, the development of increasingly sophisticated variance component methods has enabled QTL to be mapped in increasingly general pedigrees (AMOS 1994; ALMASY and BLANGERO 1998).

In contrast to the long association human geneticists have had with random effects models, animal geneticists' acceptance of QTL as random effects is relatively recent. FERNANDO and GROSSMAN (1989), HOESCHELE (1993), and VAN ARENDONK *et al.* (1994) began by assuming the QTL variance and location, among other parameters, were known. These parameters were later estimated with a single-marker single-QTL model (GRIGNOLA *et al.* 1996a,b) and multiple linked markers and QTL model (GRIGNOLA *et al.* 1997). To date, QTL mapping in outbred animal populations has been confined to experimental populations (*e.g.*, daughter and

Corresponding author: Andrew W. George, Division of Genetics and Biometry, Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, United Kingdom. E-mail: andrew.george@bbsrc.ac.uk

granddaughter designs). This can be attributed to the availability of data and complexities associated with calculating (co)variance matrices for QTL effects given multigenerational pedigrees.

The aim of this article is to present to the animal genetics community a new two-step variance component method that is capable of routinely mapping QTL in populations with considerable missing marker information and complex pedigree structures. The methodology is based upon an interval mapping procedure and begins by utilizing available marker data and pedigree information to calculate the (co)variance matrices associated with a QTL at a particular position on the genome. Once the (co)variance matrix is calculated, the mixed linear model is constructed and parameter estimates are obtained. This two-step process of calculating the (co)variance matrix and estimating the parameters of the mixed linear model is repeated for each position on the genome. A test statistic measuring QTL presence is then obtained from which position and size can also be determined. The ability of this method to analyze complex pedigree data is owed to the recently upgraded and freely available software package Loki (HEATH 1997). Loki enables the IBD probabilities at a QTL to be calculated between all pairs of individuals given considerable missing information and pedigree complexities. These IBD probabilities are used to construct the QTL's (co)variance matrix.

## MATERIALS AND METHODS

**Mixed linear models:** When constructing a mixed linear model that accounts for a QTL, the quantitative trait is commonly assumed to be controlled by a linear combination of fixed effects, putative QTL, and additive residual (polygenic) effects. The polygenic effects account for the cumulative result of all loci affecting the quantitative trait that are unlinked to the QTL. Mixed linear models can be constructed at the animal or gametic level. In this article, an animal model is presented, which, in matrix notation, is defined as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{Z}\mathbf{v} + \mathbf{e}, \quad (1)$$

where  $\mathbf{y}$  is an  $(m \times 1)$  vector of phenotypes,  $\mathbf{X}$  is an  $(m \times s)$  design matrix,  $\boldsymbol{\beta}$  is a  $(s \times 1)$  vector of fixed effects,  $\mathbf{Z}$  is an  $(m \times q)$  incidence matrix relating animals to phenotypes,  $\mathbf{u}$  is a  $(q \times 1)$  vector of additive polygenic effects,  $\mathbf{v}$  is a  $(q \times 1)$  vector of additive QTL effects, and  $\mathbf{e}$  is a residual vector.

The random effects  $\mathbf{u}$ ,  $\mathbf{v}$ , and  $\mathbf{e}$  are assumed to be uncorrelated and distributed as multivariate normal densities as follows:  $\mathbf{u} \sim \mathbf{N}_q(\mathbf{0}, \mathbf{A}\sigma_u^2)$ ,  $\mathbf{v} \sim \mathbf{N}_q(\mathbf{0}, \mathbf{G}\sigma_v^2)$ , and  $\mathbf{e} \sim \mathbf{N}_m(\mathbf{0}, \mathbf{R}\sigma_e^2)$ , where the scalar variances  $\sigma_u^2$ ,  $\sigma_v^2$ , and  $\sigma_e^2$  are the polygenic variance, the additive variance of the QTL, and the residual variance, respectively;  $\mathbf{A}$  is the standard additive genetic relationship matrix;  $\mathbf{G}$  is the  $(q \times q)$  (co)variance matrix for the additive effects of the QTL conditional on marker information; and  $\mathbf{R}$  is a known  $(m \times m)$  diagonal matrix. If individuals  $i$  and  $j$  are noninbred, then  $g_{ij} \in \mathbf{G}$  represents the proportion of alleles IBD at the QTL. However, if  $i$  and  $j$  are inbred,  $g_{ij}$  is interpreted as twice the coefficient of coancestry for the QTL. The coefficient of coancestry is defined as the probability that two randomly drawn alleles from individuals  $i$  and  $j$  are

IBD (MALÉCOT 1948). See XIE *et al.* (1998) for a detailed explanation of the interpretation of  $g_{ij}$  given inbred and noninbred individuals.

When no QTL is assumed to be segregating in the population, the mixed linear model in matrix notation becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \quad (2)$$

with  $\mathbf{u} \sim \mathbf{N}_q(\mathbf{0}, \mathbf{A}\sigma_u^2)$  and  $\mathbf{e} \sim \mathbf{N}_m(\mathbf{0}, \mathbf{R}\sigma_e^2)$ .

**Calculating the IBD probabilities for the  $\mathbf{G}$  matrix:** In practice, QTL genotypes are unobservable. Instead, linked markers are genotyped and used to infer QTL IBD status. The marker information in complex pedigrees is often incomplete. Unknown linkage phases, noninformative markers, and/or missing marker genotypes complicate the calculation of  $\mathbf{G}$ . Several methods for calculating IBD probabilities in complex pedigrees have been developed. These methods fall into one of three classes—recursive algorithms, correlation-based algorithms, or simulation-based algorithms.

**Recursive algorithms:** Recursive algorithms to calculate IBD probabilities for a QTL's gametic relationship matrix were developed by VAN ARENDONK *et al.* (1994) and WANG *et al.* (1995). These algorithms can also be used to construct  $\mathbf{G}$  since a simple linear relationship exists between the (co)variance matrix used in animal QTL models and the gametic relationship matrix. That is,  $g_{ij} = 0.5 \sum_{s=m,p} \sum_{t=m,p} g_{s,t} g_{i,j}$ , where  $s, t \in \{\text{maternal } (m), \text{paternal } (p)\}$  and  $g_{s,t}$  represents the probability of the  $s$ th parental gamete inherited from individual  $i$  being IBD to the  $t$ th parental gamete inherited from individual  $j$ . The calculation of the gametic IBD probabilities is based upon information from a single fully genotyped marker linked to a QTL. Extensions to linked phase-known marker data were made by GRIGNOLA *et al.* (1996a).

Recursive algorithms are an effective and economical way of calculating IBD probabilities given the availability of full marker information; however, this requirement is difficult to guarantee for complex pedigrees. WANG *et al.* (1995) discussed a nonstochastic approach to handling missing marker information while maintaining the recursive integrity of the algorithm; however, large amounts of missing marker information render the algorithm intractable. Furthermore, recursive algorithms follow a "top-down" strategy beginning with the calculation of IBD probabilities for the parents and using these estimates to infer the IBD probabilities of the offspring. Missing information on individuals early in the pedigree introduces estimation errors that propagate throughout the pedigree because recursive algorithms are incapable of utilizing information that is not otherwise passed down through the parents.

**Correlation-based algorithms:** ALMASY and BLANGERO (1998) developed an alternate approach for IBD probability calculation. Their methodology espouses the IBD correlation relationships of AMOS (1994), who, in matrix notation, showed

$$\mathbf{G} = \mathbf{A} + \mathbf{B}(r, \theta) \otimes (\mathbf{G}_M - \mathbf{A}), \quad (3)$$

where  $\mathbf{B}(r, \theta)$  is the correlation matrix between the proportion of alleles shared IBD at the fully genotyped marker and a putative QTL,  $r$  denotes the  $n$ th kinship relationship,  $\otimes$  represents the Hadamard product, and  $\mathbf{G}_M$  is the  $(q \times q)$  (co)variance matrix conditioned on and calculated at the marker  $M$ . ALMASY and BLANGERO (1998) used the averaging method of FULKER *et al.* (1995) to extend (3) to allow the calculation of  $\mathbf{G}$  to be conditional on all available marker information.

ALMASY and BLANGERO (1998) have made a significant contribution to the advancement of correlation-based algorithms; however, little attention is paid to the difficulties of calculating  $\mathbf{G}_M$  given missing marker information. The authors suggest Monte Carlo methods to impute the missing marker genotypes but irreducibility (*i.e.*, the ability of a sampler to visit any

consistent state in a parameter space with positive probability) of the chains is difficult to assess and guarantee. Issues relating to the use of Monte Carlo methods to infer missing marker genotypes are discussed in further detail below.

*Simulation-based algorithms:* For pedigrees with incomplete marker information, direct application of recursive or correlation-based IBD algorithms is impossible. In this situation,  $\mathbf{G}$  is often replaced by its expectation conditioned on the observed marker data ( $M_{\text{obs}}$ ) such that

$$E(\mathbf{G}|M_{\text{obs}}) = \sum_{\omega \in \Omega} \mathbf{G}_{\omega} \Pr(\omega|M_{\text{obs}}), \quad (4)$$

where  $\omega$  is a single phase-known marker configuration for the pedigree from the set of all possible marker configurations ( $\Omega$ ),  $\mathbf{G}_{\omega}$  is the (co)variance matrix for the QTL conditional on  $\omega$ , and  $\Pr(\omega|M_{\text{obs}})$  is the conditional probability of the complete marker configuration  $\omega$  given the observed data  $M_{\text{obs}}$ . The (co)variance matrix  $\mathbf{G}_{\omega}$  can now be estimated via one of the above IBD algorithms as if full marker data are available.

Calculating the expectation of  $\mathbf{G}$  for pedigrees containing substantial missing data presents two computational challenges. First, the number of configurations in  $\Omega$  is potentially large, thus the order of the summation in (4) makes the calculation infeasible. In practice, a Monte Carlo approximation is used (see GRIGNOLA *et al.* 1996a). Second, the exact calculation of  $\Pr(\omega|M_{\text{obs}})$  is intractable. Exact methods such as the ELSTON and STEWART (1971) algorithm and peeling algorithms (CANNINGS *et al.* 1978) are exponential in pedigree complexity and marker polymorphicity. Instead practitioners rely on simulation techniques, namely MCMC methods.

A plethora of MCMC algorithms have been developed for the exploration of  $\Omega$  and thus approximation of  $\Pr(\omega|M_{\text{obs}})$ . Among the simplest are the “single-site” approaches (SHEEHAN 1990), which update each locus for each individual separately. The individual’s genotype is updated, conditioned upon the individual’s phenotype and the current genotypes of the parents, spouses, and offspring. Unfortunately, single-site samplers can possess poor “mixing” qualities for complex pedigrees and irreducibility of the chains can be ensured only for biallelic loci (LIN *et al.* 1994). Difficulties in exploring  $\Omega$  stem from the observed marker data constraining the set of missing marker configurations. Not all marker configurations are consistent with Mendelian inheritance rules. Several more complex samplers (LIN *et al.* 1993; GEYER and THOMPSON 1995; LUND and JENSEN 1998) that reportedly ensure irreducibility have been suggested; however, irreducibility of the chains can still not always be guaranteed as discovered by JENSEN and SHEEHAN (1998).

These difficulties prompted THOMPSON (1994) to devise an alternate sampling strategy which can be used for a variety of tasks including the estimation of  $\mathbf{G}$ . It has long been recognized that segregation events (*i.e.*, the separation of alleles at a locus during meiosis) govern the inheritance of genetic material from parent to offspring. In fact, marker genotypes are merely the observed results of segregations. THOMPSON (1994) developed a sampler, based upon segregation indicators that are binary variables modeling segregations, to explore the set of possible segregation configurations ( $\Lambda$ ). This then allowed  $\mathbf{G}_{\mathbf{s}}$ , the (co)variance matrix for a QTL conditioned on the segregation indicators, and  $\Pr(\mathbf{s}|M_{\text{obs}})$  to be estimated where  $\mathbf{s} \in \Lambda$ . Also the expectation of  $\mathbf{G}$  can be easily calculated as  $E(\mathbf{G}|M_{\text{obs}}) = \sum_{\mathbf{s} \in \Lambda} \mathbf{G}_{\mathbf{s}} \Pr(\mathbf{s}|M_{\text{obs}})$ . The space of segregation indicators is far less constrained than the space of missing marker genotypes, culminating in Monte Carlo chains with improved convergence and irreducibility properties. A single-site Metropolis-Hastings sampler was developed by THOMPSON (1994)

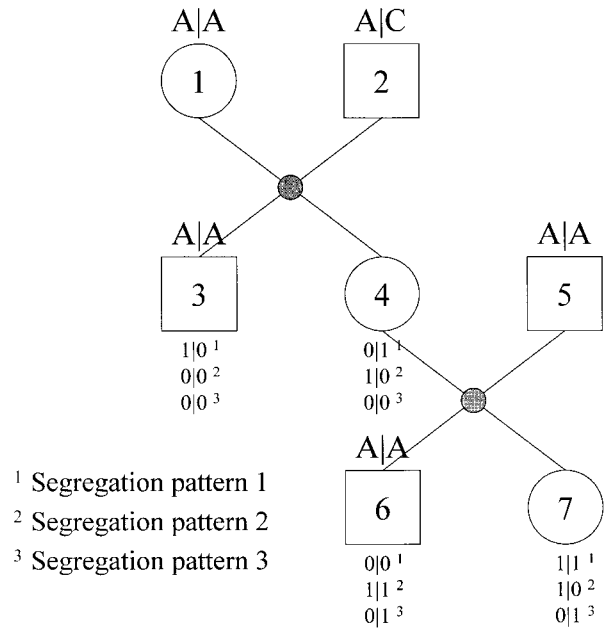


FIGURE 1.—A simple pedigree illustrating the relationship between marker genotypes (*e.g.*, A|C) and segregation indicators (*e.g.*, 1|0):  $\circ$ , a female;  $\square$ , a male; and  $\bullet$ , the mating of two individuals. Individuals 4 and 7 have missing marker genotypes. The marker genotypes are ordered such that  $x|y$  signifies that allele  $x$  has been inherited from the maternal parent and allele  $y$  has been inherited from the paternal parent. Three segregation patterns (*i.e.*,  $\mathbf{s}^1$ ,  $\mathbf{s}^2$ ,  $\mathbf{s}^3$ ) consistent with the marker data are shown. These segregation patterns are vectors of segregation indicators and they give the possible allelic pathways through the pedigree. For example, the first segregation pattern for individual 3 infers that this individual inherited its mother’s paternal allele and its father’s maternal allele. Note that 1|1 is not a valid set of segregation indicators for individual 3 since the paternal marker allele of 2 is not passed on to individual 3.

and later extended to the simultaneous updating of multiple sites in THOMPSON and HEATH (1999).

**Segregation indicators and their use in estimating  $\mathbf{G}$ :** Using notation consistent with THOMPSON and HEATH (1999), the segregation indicator ( $S_{ij}$ ) equals 0 if the inherited allele at the  $i$ th segregation and the  $j$ th locus is the parent’s maternal allele. Alternately,  $S_{ij} = 1$  if the inherited allele at the  $i$ th segregation and the  $j$ th locus is the parent’s paternal allele. The set of segregation indicators for the  $m$  segregations in the pedigree and the  $n$  loci where these loci may be marker loci and/or QTL is represented by  $\mathbf{s} = \{S_{ij}; i = 1, \dots, m; j = 1, \dots, n\}$ .

Consider the pedigree depicted in Figure 1, where ordered marker information is recorded on a single locus (*i.e.*,  $x|y$  implies  $x$  is the allele inherited from the maternal parent and  $y$  is the allele inherited from the paternal parent). Shown are three different sets of segregation indicators consistent with the observed marker data and pedigree structure. These segregation indicators give possible allelic pathways through the pedigree. Since segregation events are not directly observable, several segregation patterns may be consistent for the same set of marker data.

By obtaining a large number of  $\mathbf{s}$  with probability  $\Pr(\mathbf{s}|M_{\text{obs}})$ , these segregation indicators can be used to estimate IBD probabilities between any pair of individuals in the pedigree. For example, in the first and third segregation patterns in Figure

TABLE 1  
Features of the four setups considered in the simulation study

Setup	Description	No. of alleles/marker	$h_v^2$	Pedigree structure	Marker information
Benchmark	Standard setup	8	0.10	Sheep	Complete
A	Pedigree structure altered	8	0.10	Pig	Complete
B	Partial marker information	8	0.10	Sheep	Partial
C	Change in marker polymorphism	3	0.10	Sheep	Complete

1, the maternal allele of individual 6 originates from (or is IBD to) the maternal allele of individual 1, while in the second segregation pattern, the maternal allele of individual 6 originates from (or is IBD to) the maternal allele of individual 2. Therefore, based upon these realizations,  $\Pr(6_m \equiv 2_m | \mathbf{M}_{\text{obs}}) = \frac{1}{3}$  and  $\Pr(6_m \equiv 1_m | \mathbf{M}_{\text{obs}}) = \frac{2}{3}$ , where  $\equiv$  represents IBD.

**Multiple-site segregation sampler:** A brief introduction to the multiple-site segregation sampler, as developed by THOMPSON and HEATH (1999) and employed in this article, is now presented. Readers who wish to pursue a more rigorous derivation are invited to read THOMPSON and HEATH (1999).

Very simply, the multiple-site segregation sampler is a cleverly designed Gibbs sampler (GEMAN and GEMAN 1984) with batch updating, which allows IBD probabilities to be calculated in pedigrees with unknown marker genotypes and unknown marker phases. Exploration of the joint density  $\Pr(\mathbf{s} | \mathbf{M}_{\text{obs}})$ , which may be of high dimension when the pedigree is large, is facilitated through the sampling of  $m$  simpler  $n$ -dimensional conditional distributions such that

$$\Pr(\mathbf{s} | \mathbf{M}_{\text{obs}}) = \prod_{i=1}^m \Pr(\mathbf{s}_{i \cdot} | \{\mathbf{s}_{j \cdot}; j \neq i\}, \mathbf{M}_{\text{obs}}), \quad (5)$$

where  $\mathbf{s}_{i \cdot} = \{S_{il}; l = 1, \dots, n\}$  and  $\Pr(\mathbf{s}_{i \cdot} | \{\mathbf{s}_{j \cdot}; j \neq i\}, \mathbf{M}_{\text{obs}})$  is the probability of the segregation indicators across  $n$  loci at the  $i$ th segregation conditional on all other segregation indicators and observed marker data. Since the number of loci may be large, drawing realizations directly from the conditional distributions in (5) remains challenging. Therefore, THOMPSON and HEATH (1999) devised a two-step strategy to sample  $\mathbf{s}_{i \cdot}$  from its  $n$ -dimensional conditional distribution.

The first step involves moving through the genome, calculating locus by locus, cumulative probabilities for  $S_{ij}$ . These probabilities are relatively easy to calculate recursively. Once all  $n$  cumulative probabilities have been obtained, the second step involves moving back down the genome, sampling  $S_{ij}$  from a univariate density that is a function of the associated cumulative probability, the previous sampled segregation indicator ( $S_{i,j+1}$ ), and the recombination rate between loci  $j$  and  $j+1$ . In this way,  $\mathbf{s}_{i \cdot}$  can be sampled from its conditional distribution. By repeating these two steps for  $i = 1, \dots, m$ , a realization from (5) is obtained.

**Implementation of the multiple-site segregation sampler:** Implementation of the multiple-site segregation sampler is via an adapted version of the QTL mapping software Loki. Loki was originally designed for multipoint linkage analysis in general pedigrees using MCMC methods; however, it has since been modified for IBD probability calculation. The user supplies Loki with the pedigree structure, marker genotypes, marker positions, and QTL positions for which the IBD matrices are to be calculated. Dependent chains of IBD probabilities are then obtained for each QTL position. Convergence is determined by monitoring the IBD probabilities over the itera-

tion number. Once the probabilities stabilize, the sampler is deemed to have reached convergence.

**Two-step variance component approach:** The variance component approach to map QTL in complex pedigrees is composed of two distinct steps:

- Step 1. For each QTL position on the chromosomal segment, the (co)variance matrix for the QTL (*i.e.*,  $\mathbf{G}$ ) is calculated.
- Step 2. For each position considered in step 1, construct the mixed linear models (1) and (2), obtain estimates of the parameters, and test for the presence of a QTL.

These steps are common to all interval mapping-based variance component methods; however, their implementation differs greatly among practitioners. For example, there are various approaches to calculating the  $\mathbf{G}$  matrix that have already been discussed and there are numerous analytical and simulation techniques for estimating the parameters of a mixed linear model.

With regard to the implementation strategy adopted in this article, in step 1 the IBD probabilities for the  $\mathbf{G}$  matrix are obtained via the multiple-site segregation sampler. In step 2 ASREML (GILMOUR *et al.* 1998) provides restricted maximum-likelihood (REML) estimates of (1) and (2). ASREML was chosen over other available REML packages due to its ability to handle large user-defined (co)variance matrices. To test for the presence of a QTL against no QTL at a particular chromosomal position, the test statistic  $\log \text{LR} = -2 \ln(L_0(H_0, \text{no QTL present}) - L_1(H_1, \text{QTL present}))$  is constructed, where  $L_1$  and  $L_0$  represent the respective likelihood values of (1) and (2) evaluated at the REML solutions.

**Distribution of the test statistic:** Statistical theory states that  $\log \text{LR}$  follows a  $\chi^2$  distribution with the degrees of freedom equal to the number of parameters being tested (WILKS 1938). However, in the context of interval mapping, the asymptotic behavior of  $\log \text{LR}$  is under nonstandard conditions since the null hypothesis places parameters on the boundary of the parameter space defined by the alternative hypothesis (STRAM and LEE 1994). Furthermore, the distribution of  $\log \text{LR}$  under  $H_0$  is influenced by the chromosomal segment length, the degree of missing marker data, and the distributional properties of the trait.

When a single chromosomal position is being tested,  $\log \text{LR}$  follows a 50:50 mixture distribution, where one mixture component is a peak at 0 and the other component is a  $\chi^2_1$  distribution (SELF and LIANG 1987). When a chromosomal interval is being tested, XU and ATCHLEY (1995) found the empirical distribution of  $\log \text{LR}$  follows a  $\chi^2$  distribution with between 1 and 2 d.f. QTL detection, however, is often carried out over large chromosomal regions and even the entire genome. For these situations, the distribution of  $\log \text{LR}$  under  $H_0$  is unclear.

Since this article deals with simulated data, it is possible

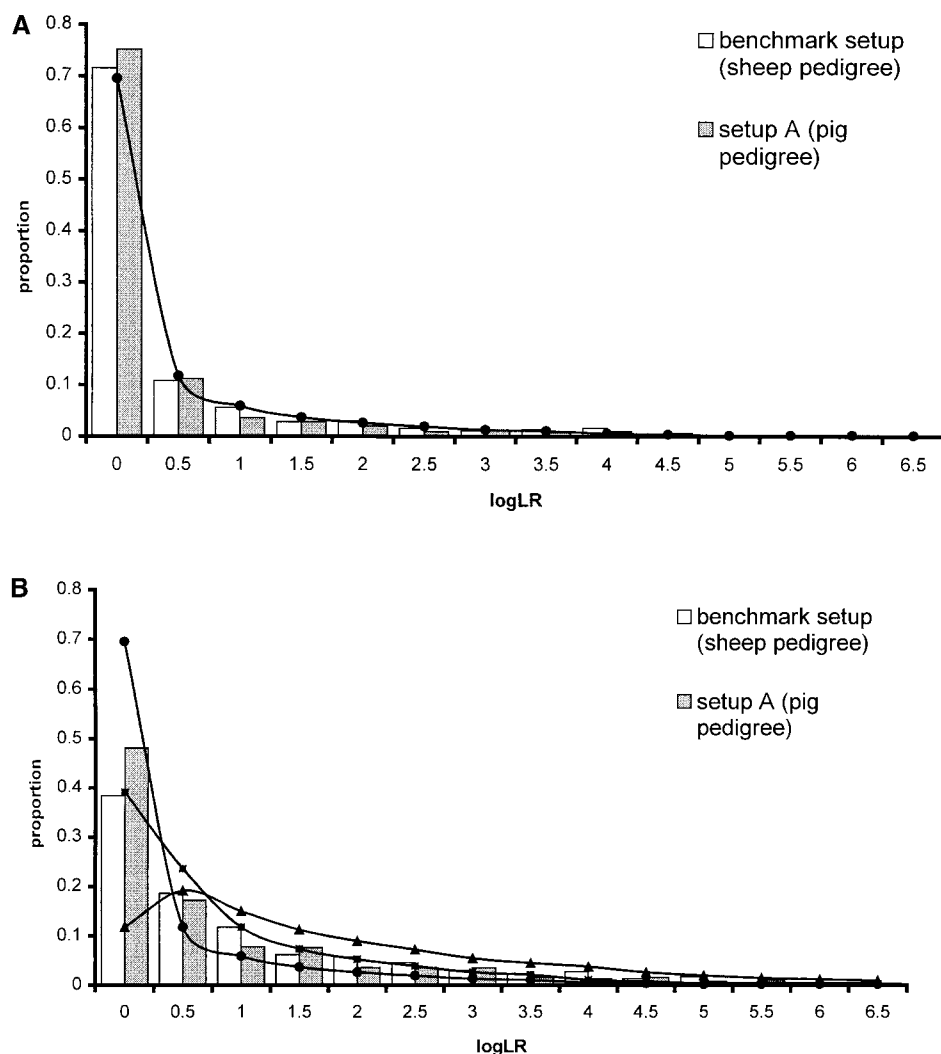


FIGURE 2.—The empirical distributions of log LR (from 500 replicates) under  $H_0$  for the benchmark setup (*i.e.*, a simulated sheep population) and setup A (*i.e.*, a simulated pig population). (A) Locus-wide test statistic: the distribution of the test statistic when a single point on the chromosome is being tested (*i.e.*, 35 cM). (B) Chromosome-wide test statistic: the distribution of log LR for a chromosome-wide test. The possible theoretical distributions of log LR are also shown: ■,  $\chi^2_1$ ; ▲,  $\chi^2_2$ ; and ●, the 50:50 mixture distribution of SELF and LIANG (1987).

to replicate data under the null hypothesis, construct the empirical distribution of log LR, and derive empirical threshold values in which to determine QTL presence. For real data, permutation methods (CHURCHILL and DOERGE 1994) have been suggested. The large number of required analyses, though, limits the methodology to relatively small pedigrees. Furthermore, it is not clear how the data should be permuted given populations with complex pedigree structures.

#### SIMULATION STUDY

The simulation study begins with the analysis of data generated under  $H_0$ . By constructing a histogram of the test statistic over replicates, an empirical distribution of log LR is obtained. QTL presence at a chromosomal position is then determined in subsequent analyses by comparing the respective test statistic to the empirical threshold.

To investigate the performance of the two-step variance component method for mapping QTL in complex pedigrees, data are generated under four simulation setups. The first setup, referred to as the “benchmark”

setup, involves the generation of fully genotyped, highly polymorphic marker data. A biallelic QTL that explains 10% of the total variation (*i.e.*,  $h_v^2 = 0.1$ ) is segregating in the population. Setups A, B, and C then change a single feature of the benchmark setup, enabling the effect on the variance component method’s performance to be assessed. The four setups considered in this study are summarized in Table 1 and are discussed in detail below together with the generation of data under  $H_0$ .

**Generation of data under  $H_0$ :** Replicates are generated according to the benchmark setup and setup A (which are described below) but without a segregating QTL in the population. These two setups are equivalent except the benchmark setup is based upon a sheep pedigree where setup A is based upon a pig pedigree. This allows the sensitivity of the test statistic’s distribution to changes in pedigree structure to be assessed.

**Benchmark setup:** The pedigree structure is based upon a real pedigree created to explore copper deficiency in a selected sheep population. The original ex-

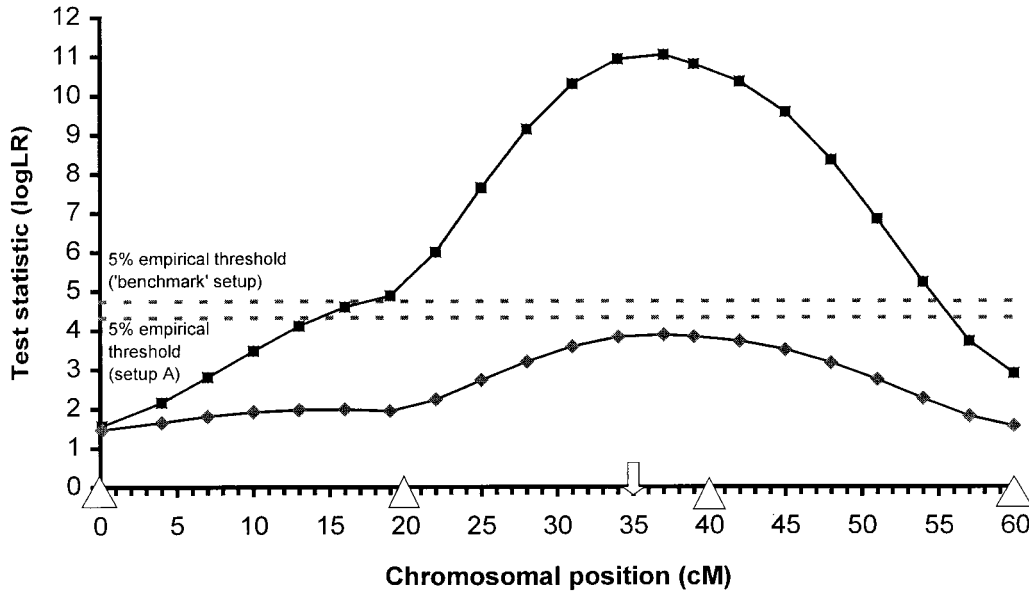


FIGURE 3.—The mean log LR over 50 replications of data generated under the benchmark setup (*i.e.*, sheep pedigree, biallelic QTL,  $h_v^2 = 0.1$ , eight alleles per marker, complete marker information) and setup A (*i.e.*, pig pedigree, biallelic QTL,  $h_v^2 = 0.1$ , eight alleles per marker, complete marker information).  $\blacklozenge$  and  $\blacksquare$ , the mean log LR profile for the benchmark setup and setup A, respectively;  $\triangle$ , marker location;  $\Downarrow$ , the simulated position of the QTL.

periment contained over 2000 individuals; however, for the purposes of demonstrating the methodology, a subset of 500 individuals is selected. In reducing the pedigree's size, careful attention is given to maintaining the structure's original complex nature. The reduced pedigree consists of 269 related families spanning four generations with 1.8 offspring on average per mating. The pedigree structure contains no inbreeding.

The marker information consists of four polymorphic markers segregating with eight equally frequent alleles and placed on a chromosomal segment of length 60 cM at positions 0, 20, 40, and 60 cM. A biallelic QTL with alleles  $Q$  and  $q$  segregating at equal frequencies is then placed between the second and third markers at position 35 cM. If an individual inherits  $QQ$  from its parents, its phenotypic contribution due to the QTL is  $v_i = m + a$ , where  $m = 0$  and  $a = 13.5$ . If the individual's QTL genotype is heterozygous or  $qq$ , the individual's phenotypic contribution is  $v_i = m + d$  or  $v_i = m - a$ , respectively, where  $d = 0$ . FALCONER (1989) defines  $m$

as the midhomozygote value,  $a$  as the additive effect, and  $d$  as the dominance effect.

The polygenic contribution made by an individual is dependent upon the polygenic contributions of its parents and Mendelian sampling. Since the complete parentage of every individual in the pedigree is not available,  $u_i$  is generated according to the number of known parents. If both parents are unknown (*i.e.*, the individual is a founder), then  $u_i \sim N(0, \sigma_u^2)$ . Suppose, however, the sire of the  $i$ th individual ( $s_i$ ) is in the pedigree. Then  $u_i \sim N(0.5u_{s_i}, (0.75 - 0.25f_{s_i})\sigma_u^2)$ , where  $f_{s_i}$  is the inbreeding coefficient of  $s_i$ . A similar distribution is used to generate  $u_i$  when only the individual's dam is in the pedigree. If both parents are present,  $u_i \sim N(0.5(u_{s_i} + u_{d_i}), 0.5(1 - 0.5(f_{s_i} + f_{d_i}))\sigma_u^2)$ , where  $f_{d_i}$  is the inbreeding coefficient for the dam of individual  $i$ . The environmental error term  $e_i$  is generated from  $N(0, \sigma_e^2)$ . Fixed effects are not generated; thus  $\mathbf{X}\beta$  in (1) and (2) equals  $\mu$ , the overall mean.

The value of  $\sigma_v^2$  is dependent upon  $m$ ,  $a$ ,  $d$ , and  $p_Q$

TABLE 2

Results from the analysis of data generated under the benchmark setup

Parameters	Simulated value	Mean estimate	SD	$\overline{\text{BRV}}$	Mean bias
$h_v^2$	0.10	0.145	0.091	0.86 <sup>a</sup>	0.040
$h_u^2$	0.34	0.296	0.127	1.03 <sup>a</sup>	-0.039
$\sigma_e^2$	500	488.59	65.76	0.33	-4.03
$d_Q$	35	34.48	15.64	4.31	-0.91
Log LR	—	4.96	4.40	0.001	—

The parameter estimates' mean, standard deviation (SD), mean between-run variance ( $\overline{\text{BRV}}$ ), and mean biases for  $h_v^2$ ,  $h_u^2$ ,  $\sigma_e^2$ , and  $d_Q$  are based upon 50 replicates. The mean estimate, SD, and BRV of the peak log LR are also shown.

<sup>a</sup> Multiply values by  $10^{-5}$ .

TABLE 3  
Results from the analysis of data generated under setup A

Parameters	Simulated value	Mean estimate	SD	$\overline{\text{BRV}}$	Mean bias
$h_v^2$	0.10	0.113	0.051	0.23 <sup>a</sup>	0.009
$h_u^2$	0.34	0.324	0.119	0.19 <sup>a</sup>	-0.022
$\sigma_e^2$	500	496.51	56.99	1.43	6.52
$d_Q$	35	34.23	9.22	2.05	-1.09
Log LR	—	12.10	7.34	0.073	—

See Table 2 for description.

<sup>a</sup> Multiply values by  $10^{-5}$ .

(the  $Q$  allele frequency), where  $\sigma_v^2 = 2p_Q(1 - p_Q)a^2$  when  $d = 0$  (FALCONER 1989). Although altering  $m$ ,  $a$ ,  $d$ , and  $p_Q$  affects  $\sigma_v^2$ , the size of the QTL is characterized by the amount of total variation explained by the QTL. To obtain a QTL explaining 10% of the total variation,  $\sigma_u^2$  and  $\sigma_e^2$  are set to 300.0 and 500.0, respectively. For these values, the heritability of the trait is 44%.

**Setup A:** The ability of the variance component method to map QTL in a pedigree with large numbers of offspring per mating and inbreeding is investigated. The pedigree used in this study is again based upon a real structure originating from a Meishan pig experiment. The initial experiment contained  $\sim 2500$  related individuals, but for meaningful comparisons to be made with the benchmark analysis, 500 related individuals are selected. The average number of offspring per family is 14.3 across five generations of matings, consisting of 35 related families. The average inbreeding coefficient is 4.5% with a maximum inbreeding coefficient of 17%.

**Setup B:** Complex pedigrees often contain individuals with missing marker genotypes. This missing information introduces uncertainty into the analysis. To better understand the ability of the methodology to cope with this uncertainty, two approaches to removing the marker information generated according to the benchmark setup are explored. The first approach is where 50% of the marker genotypes are randomly removed. The second approach removes only the marker informa-

tion of offspring that are not themselves parents. This results in a 53% loss in marker information.

**Setup C:** The final setup investigates the effect a reduction in marker informativeness has upon the analysis. Data are generated according to the benchmark setup except three alleles as opposed to eight alleles are segregating at the markers.

## RESULTS

Results from the application of the two-step variance component method to replicated data generated under the above-described simulation study are now reported. Due to the analyses being computationally demanding, only every third centimorgan is tested for the presence of a QTL. A single analysis across the chromosome can take up to 56 min on a Compaq Professional Workstation XP1000 utilizing a single Alpha 21264 processor running at 500 MHz. Four parallel analyses per replicate are performed, where Loki and ASREML runs begin from different well-dispersed starting values. The empirical distributions of the test statistic, however, are created from the analysis of 500 replicated data sets; therefore only a single run is performed per replicate due to obvious computational constraints.

**Construction of the empirical distribution of log LR under  $H_0$ :** Figure 2A reveals close agreement between the empirical and theoretical (*i.e.*, 50:50 mixture where one component mixture is a peak at 0 and the other is a  $\chi_1^2$ ) distributions of log LR when a single position on the chromosome is tested for QTL presence (*i.e.*, the 35-cM position). This was found to hold regardless of the position being tested.

In Figure 2B, when a chromosome-wide QTL search is performed, the empirical distributions appear to follow a  $\chi_1^2$ . However, the 5% threshold value obtained from  $\chi_1^2$  is less conservative than the empirical thresholds. The 5% empirical threshold values are 4.74 and 4.32 for the benchmark setup and setup A, respectively, where  $\chi_{1,0.05}^2 = 3.84$ . Empirical thresholds are used throughout this article. It is interesting to note that the empirical distributions are relatively unaffected by changes in pedigree structure.

TABLE 4

The proportion of individuals in the pedigree with 0, 1, 2, 3, and 4 missing marker genotypes for patterns 1–5 in setup B

Pattern of missing marker genotypes	No. of missing marker genotypes				
	0	1	2	3	4
1	0.05	0.26	0.38	0.24	0.07
2	0.08	0.23	0.41	0.24	0.04
3	0.07	0.26	0.37	0.23	0.06
4	0.08	0.28	0.34	0.22	0.08
5	0.0	0.0	0.0	0.0	0.53

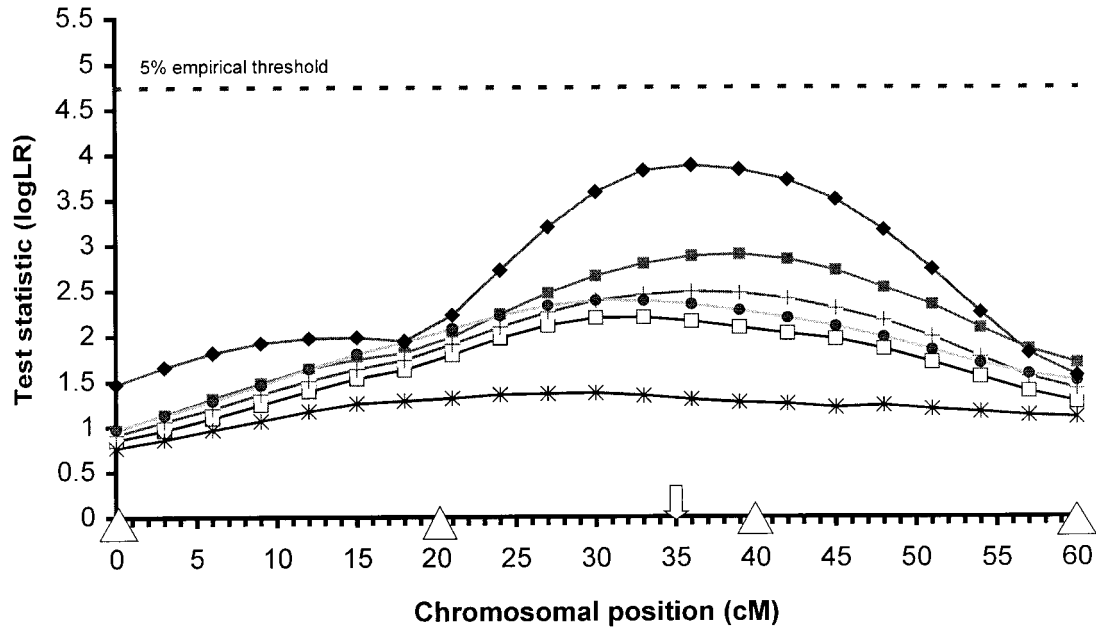


FIGURE 4.—The mean log LR over 50 replications of data generated under setup B (*i.e.*, sheep pedigree, biallelic QTL,  $h_v^2 = 0.1$ , eight alleles per marker, partial marker information).  $\square$ ,  $\blacksquare$ , +,  $\bullet$ , \*, patterns 1, 2, 3, 4, and 5 respectively. For comparison, the mean log LR of data generated under the benchmark setup is provided ( $\blacklozenge$ ).  $\triangle$ , marker location;  $\ddagger$ , the simulated position of the QTL.

**Benchmark setup:** The mean log LR profile over 50 replications of data generated under the benchmark setup is shown in Figure 3. The profile peaks between markers 2 and 3 at the position of the simulated QTL (*i.e.*, 35 cM). The mean peak is well below the 5% empirical threshold; however, this result is slightly misleading. The peak of the mean profile is biased downward because the estimated position, and thus corresponding peak of the profile, varies across replicates. In fact, 48% of the analyses yield a test statistic along the chromosomal segment in excess of the 5% threshold.

The ability of the methodology to accurately estimate the parameters of interest can be gauged from the results presented in Table 2. The mean parameter estimates of  $h_v^2$ ,  $h_{us}^2$ ,  $\sigma_e^2$ , and  $d_Q$ , where  $d_Q$  represents the location of the QTL in centimorgans, are close to the simulated values; the parameter estimates' standard deviations (SD) are reasonable; and the mean between-run variance is small. However, a more appropriate measure of accuracy is the mean bias. The mean bias,  $E(\hat{\theta} - \theta)$ , is defined as the expected value of the difference between the estimator ( $\hat{\theta}$ ) and the parameter's true value ( $\theta$ ). For example, the mean bias of the estimate of  $h_v^2$  is  $E(\hat{h}_v^2 - h_v^2) = \sum_{i=1}^{50} \sum_{j=1}^4 [(h_{v_i}^2)_{ij} - (h_{v_i}^2)_i] / 200$ , where  $(h_{v_i}^2)_{ij}$  represents the REML estimate of  $h_v^2$  from the analysis of the  $i$ th replicate and the  $j$ th parallel run, and  $(h_{v_i}^2)_i$  represents the true parameter value of  $h_v^2$ .

For the parameters  $h_{us}^2$ ,  $\sigma_e^2$ , and  $d_Q$  in Table 2 the mean bias is small and negative, implying a slight underestimating of the parameters. For  $h_v^2$ , the mean bias of  $\sim 0.04$  suggests that the method tends to overestimate the

amount of variation explained by the QTL. A similar result is evident in GRIGNOLA *et al.* (1996b) from their simulation study. Also, the mean of the maximum log LR test statistic is 4.96, considerably larger than the peak of the mean profile given in Figure 3. This substantiates the previous claim of the mean profile being downwardly biased.

**Setup A:** Increasing the average family size has an obvious effect on the performance of the methodology as evidenced in Figure 3. With larger families, the peak of the log LR profile (based upon 50 replicates) increases from 3.9 to 11.0, where 82% of the analyses yield a log LR in excess of the 5% empirical threshold. Once again, the parameters are well estimated (see Table 3), with mean biases slightly smaller than the biases obtained under the benchmark setup.

**Setup B:** In setup B, five patterns of missing marker data are analyzed. Patterns 1–4 correspond to the random removal of 50% of the marker genotypes while pattern 5 is obtained by only genotyping the parents, which constitutes a 53% loss in marker genotypes. The proportions of individuals in the pedigree with 0, 1, 2, 3, and 4 missing marker genotypes, for each pattern, are given in Table 4. Each pattern consists of 50 replicates. These replicates, before the marker data are removed, are the same as those replicates generated under the benchmark setup. Thus, differences between the results obtained under setup B and the benchmark setup can be directly attributed to the effect of missing marker information.

The mean log LR profiles for patterns 1–5 together



TABLE 5

Results are based upon the analysis of five different patterns of missing marker data generated under setup B

Parameters	Simulated value	Pattern no.	Mean estimate	SD	$\overline{\text{BRV}}$	Mean bias
$h_v^2$	0.10	1	0.148	0.097	2.56 <sup>a</sup>	0.044
		2	0.151	0.094	1.06 <sup>a</sup>	0.046
		3	0.152	0.100	2.15 <sup>a</sup>	0.047
		4	0.164	0.106	2.73 <sup>a</sup>	0.059
		5	0.240	0.156	1.52 <sup>a</sup>	0.134
$h_u^2$	0.34	1	0.292	0.112	2.59 <sup>a</sup>	-0.053
		2	0.291	0.129	1.17 <sup>a</sup>	-0.049
		3	0.296	0.156	2.40 <sup>a</sup>	-0.044
		4	0.272	0.111	2.56 <sup>a</sup>	-0.072
		5	0.211	0.162	1.34 <sup>a</sup>	-0.134
$\sigma_e^2$	500	1	494.51	61.40	0.26	1.99
		2	491.82	54.66	0.19	5.18
		3	485.64	91.46	0.52	-13.39
		4	497.73	60.38	1.60	4.63
		5	488.07	63.76	6.16	-4.44
$d_Q$	35	1	28.32	19.03	2.21	-7.40
		2	33.46	18.49	1.32	-2.40
		3	30.18	17.86	1.25	-5.45
		4	32.40	18.93	1.36	-3.32
		5	29.98	21.13	4.98	-5.72
log LR	—	1	2.95	2.62	0.006	—
		2	3.93	3.74	0.005	
		3	3.12	2.94	0.037	
		4	3.28	2.81	0.019	
		5	2.57	2.51	0.019	

See Table 2 for description.

<sup>a</sup> Multiply values by  $10^{-5}$ .

with the mean log LR profile for the benchmark setup are shown in Figure 4. There are two points of interest to note with respect to this figure. First, the mean log LR profile for data with partially genotyped markers lies below the profile obtained with complete marker information. Less marker information introduces extra uncertainty into the analysis and the method's ability to detect QTL decreases. In fact, the percentages of analyses yielding a log LR in excess of the 5% empirical threshold are only 24, 36, 24, 28, and 20% for patterns 1–5, respectively, well below the 48% achieved when the same data contain completely genotyped individuals.

Second, no real difference exists between the performance of the method across patterns 1–4. However, the mean profile for pattern 5, where only the parents are genotyped, does appear to differ from the other log LR profiles.

The difference in the method's performance across the five patterns of missing marker data is further evidenced in Table 5. The SD, average between-run variance, and mean bias are marginally higher for patterns 1–4 than they are under the benchmark stepup (given in Table 2). For pattern 5, though, the method struggles to obtain reasonable parameter estimates, with the

mean bias for  $h_v^2$  being more than double the mean bias obtained under the other patterns.

**Setup C:** The impact of less informative markers on the ability of the variance component method to detect QTL is evident from Figure 5. The mean log LR profile is well below the 5% threshold with only 26% of the analyses yielding a significant peak log LR. The parameter estimates (shown in Table 6), however, are similar to the estimates obtained under the benchmark setup with highly informative markers. Thus, the use of less polymorphic markers imparts greater uncertainty into the detection of QTL as opposed to the estimation of QTL.

## DISCUSSION

To date, several statistical approaches have been developed to map QTL in outbred livestock populations; however, these methods focus on granddaughter or half-sib designs and are not easily extended to more challenging pedigree structures. In this article, a two-step variance component method is presented that is capable of detecting and estimating QTL in populations with complex pedigrees and considerable missing

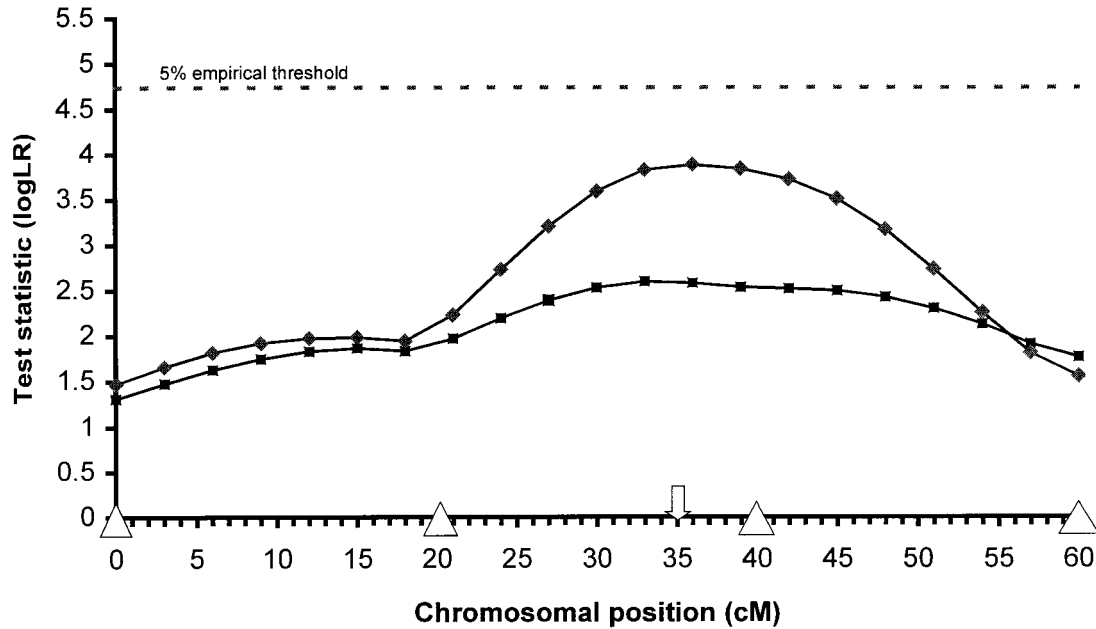


FIGURE 5.—The mean log LR over 50 replications of data generated under the setup C (*i.e.*, sheep pedigree, biallelic QTL,  $h_v^2 = 0.1$ , three alleles per marker, complete marker information). ■, the mean log LR profile for setup C. For comparison, the mean log LR of data generated under the benchmark setup is provided (◆). Δ, marker location; and ↓, the simulated position of the QTL.

marker information. The methodology is illustrated through its application to simulated sheep and pig populations.

By formulating the QTL mapping problem within a mixed linear model framework, a less parameterized statistical environment is obtained, reducing the computational burden of the analysis. The complex relationships that may exist between individuals are included within the model, leading to more accurate parameter inferences, and additional fixed and random effects can be easily incorporated into the analysis with minimal adjustment to the methodology.

For example, to simultaneously map two linked QTL, the mixed linear model becomes  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{Z}\mathbf{v}_1 + \mathbf{Z}\mathbf{v}_2 + \mathbf{e}$ , where  $\mathbf{v}_1$  and  $\mathbf{v}_2$  are the additive effects of the linked QTL. Analogous to the two-step process of mapping a single QTL, Loki is used to calculate the

(co)variance matrices of  $\mathbf{v}_1$  and  $\mathbf{v}_2$  at two separate test positions along the chromosome. Estimates of the parameters are then obtained via ASREML and the test statistic for the presence of two linked QTL is constructed. This process is repeated for each pair of test positions on the chromosome, enabling multiple QTL to be detected and localized. Note, when two QTL are being mapped, the QTL profile is a two-dimensional surface.

A two-step process to estimating the variance components of a mixed linear model is not new *per se*. FERNANDO and GROSSMAN (1989), VAN ARENDONK *et al.* (1994), and WANG *et al.* (1995) are but a few who first calculate the IBD probabilities for the QTL's (co)variance matrix and then estimate the parameters of the mixed linear models using standard statistical techniques. Difficulties in determining marker phase and

TABLE 6  
Results from the analysis of data generated under setup C

Parameters	Simulated value	Mean estimate	SD	$\overline{\text{BRV}}$	Mean bias
$h_v^2$	0.10	0.137	0.088	1.49 <sup>a</sup>	0.032
$h_u^2$	0.34	0.287	0.133	1.70 <sup>a</sup>	-0.048
$\sigma_r^2$	500	493.54	60.81	2.13	0.910
$d_Q$	35	32.21	19.66	4.29	-2.94
log LR	—	3.58	3.79	0.023	—

See Table 2 for description.

<sup>a</sup> Multiply values by  $10^{-5}$ .

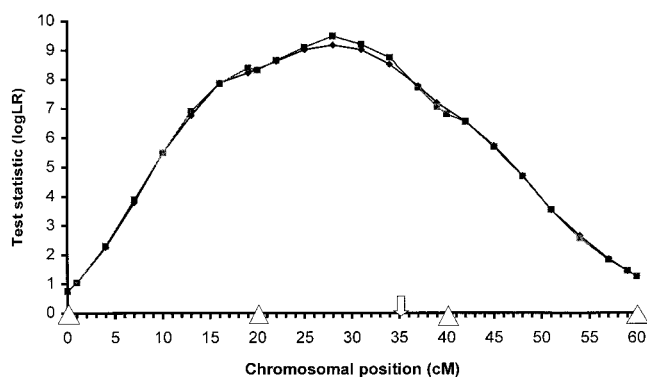


FIGURE 6.—The log LR profiles for a single replicate from setup A using two different approaches to calculating the REML estimates. The test statistics at nonmarker positions vary slightly due to the REML packages being run on computers with differing machine architectures and precisions. ◆, the log LR profile obtained with ASREML; ■, the log LR profile obtained with the derivative-free REML package of Visscher; △, marker location; ↓, the simulated position of the QTL.

unknown marker genotypes, however, mean these methods have limited application to populations with general pedigree structures. Via the multiple-site segregation sampler, opportunities now exist to analyze data with considerably complex pedigrees.

A variety of algorithms are available for the calculation of REML estimates. Standard algorithms such as ASREML require the inverse of the QTL's (co)variance matrix, which is singular at marker loci. In this article,  $\mathbf{G}$  and thus the test statistic are calculated at a position slightly to the right or left of the marker, an approach also adopted by I. HOESCHELE (personal communication). VISSCHER *et al.* (1999) instead use a derivative-free algorithm to calculate REML estimates that does not require  $\mathbf{G}^{-1}$  but  $\mathbf{V}^{-1}$ , where  $\mathbf{V}$  represents the complete (co)variance matrix for the likelihood. The complete (co)variance matrix is always nonsingular, allowing the test statistic to be calculated at marker positions. The two approaches give almost identical results. To illustrate this, a single replicate from setup A was analyzed using ASREML and the derivative-free algorithm of VISSCHER *et al.* (1999). The resulting QTL profiles are shown in Figure 6. Clearly, there is little difference between the two REML strategies; however, the present version of the derivative-free approach that calculates  $\mathbf{V}^{-1}$  is considerably more computer intensive due to its reliance upon calculating  $\mathbf{V}^{-1}$  for every convergence iterate.

Pivotal to the success of any interval mapping procedure is the calculation of an appropriate threshold value in which QTL are declared significant. The threshold value is dependent upon the distribution of the test statistic, which is known for a single point [*i.e.*, 50:50 mixture where one component mixture is a peak at 0 and the other is a  $\chi^2_1$  (STRAM and LEE 1994)], but only

approximately known for chromosome (or genome)-wide scans. Questions of how missing marker genotypes, unknown marker phase, pedigree size, map density, and QTL size influence the distribution of the test statistic remain unanswered. One solution is, given  $N$  independent tests over the chromosome (genome), calculate the  $(1 - \alpha)\%$  threshold value using the distribution of the test statistic at a single point but with the level of confidence adjusted to  $\alpha/N\%$  (*i.e.*, the Bonferroni correction for multiple testing). See LANDER and KRUGLYAK (1995) for a discussion relating to the calculation of  $N$ . An alternate solution is to further develop permutation testing (CHURCHILL and DOERGE 1994) so that the trait is reshuffled in a way that destroys the association between QTL and trait but retains the association between polygenic effect and trait. It is not yet clear how this can be accomplished for complex pedigrees.

Problems also surround the construction of confidence intervals for QTL position estimates. These problems are not unexpected given that the construction of such intervals is challenging in even simple pedigrees. For an approximate confidence interval, the LOD drop-off method could be employed and more accurate confidence intervals obtained under parametric and/or nonparametric bootstrapping methods. However, as with permutation testing, resampling for nonparametric bootstrapping methods may be difficult. Clearly, further research is needed to resolve these issues.

This article has been catalytic to initiating work in three further areas of research. First, the simulation study suggests partial marker information on most individuals is to be desired over having a mixture of fully genotyped and completely ungenotyped individuals. This is currently being explored in greater detail for a range of missing marker scenarios. Second, a new recursive algorithm to calculate IBD probability in complex pedigrees has been developed and is currently being tested. Third, the methodology is to be applied to the analysis of real sheep and beef cattle data.

The authors thank Simon Heath for his many useful comments and fine-tuning of Loki. This work was partly supported by a Biotechnology and Biological Sciences Research Council award.

#### LITERATURE CITED

- ALMASY, L., and J. BLANGERO, 1998 Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **62**: 1198–1211.
- AMOS, C. I., 1994 Robust variance-components approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.* **54**: 535–543.
- CANNINGS, C., E. A. THOMPSON and M. H. SKOLNICK, 1978 Probability functions on complex pedigrees. *Adv. Appl. Prob.* **10**: 26–61.
- CHURCHILL, G., and R. DOERGE, 1994 Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 963–971.
- ELSTON, R. C., and J. STEWART, 1971 A general model for the genetic analysis of pedigree data. *Hum. Hered.* **21**: 523–542.
- FALCONER, D. S., 1989 *Introduction to Quantitative Genetics*, Ed. 3. Longman Scientific and Technical, Essex, UK.
- FERNANDO, R. L., and M. GROSSMAN, 1989 Marker assisted selection

- using best linear unbiased prediction. *Genet. Sel. Evol.* **21**: 467–477.
- FULKER, D. W., S. S. CHERNY and L. R. CARDON, 1995 Multipoint interval mapping of quantitative trait loci, using sib pairs. *Am. J. Hum. Genet.* **56**: 1224–1233.
- GEMAN, S., and D. GEMAN, 1984 Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6**: 721–741.
- GEORGE, A. W., K. L. MENGERSEN and G. P. DAVIS, 2000 Localisation of a quantitative trait locus via a Bayesian approach. *Biometrics* **56**: 40–51.
- GEYER, C. J., and E. A. THOMPSON, 1995 Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Am. Stat. Assoc.* **90**: 909–920.
- GILMOUR, A. R., B. R. CULLIS, S. J. WELHAM and R. THOMPSON, 1998 *ASREML. Program User Manual*. Orange Agricultural Institute, New South Wales, Australia.
- GRIGNOLA, F. E., I. HOESCHELE and B. TIER, 1996a Mapping quantitative trait loci in outcross populations via residual maximum likelihood. I. *Methodology. Genet. Sel. Evol.* **28**: 479–490.
- GRIGNOLA, F. E., I. HOESCHELE, Q. ZHANG and G. THALLER, 1996b Mapping quantitative trait loci in outcross populations via residual maximum likelihood. I. A simulation study. *Genet. Sel. Evol.* **28**: 491–504.
- GRIGNOLA, F. E., Q. ZHANG and I. HOESCHELE, 1997 Mapping linked quantitative trait loci via residual maximum likelihood. *Genet. Sel. Evol.* **29**: 529–544.
- HEATH, S. C., 1997 Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am. J. Hum. Genet.* **61**: 748–760.
- HOESCHELE, I., 1993 Elimination of quantitative trait loci equations in an animal model incorporating genetic marker data. *J. Dairy Sci.* **76**: 1693–1713.
- HOESCHELE, I., P. UIMARI, F. E. GRIGNOLA, Q. ZHANG and K. M. GAGE, 1997 Advances in statistical methods to map quantitative trait loci in outbred populations. *Genetics* **147**: 1445–1457.
- JANSEN, R. C., D. L. JOHNSON and J. A. M. VAN ARENDONK, 1998 A mixture model approach to the mapping of quantitative trait loci in complex populations with an application to multiple cattle families. *Genetics* **148**: 391–399.
- JAYAKAR, S. D., 1970 On the detection and estimation of linkage between a locus influencing a character and a marker locus. *Biometrics* **26**: 451–464.
- JENSEN, C. S., and H. SHEEHAN, 1998 Problems with determination of noncommunicating classes for Monte Carlo Markov chain applications in pedigree analysis. *Biometrics* **54**: 416–425.
- KNOTT, S. A., J. M. ELSÉN and C. S. HALEY, 1996 Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. *Theor. Appl. Genet.* **93**: 71–80.
- LANDER, E., and L. KRUGLYAK, 1995 Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* **11**: 241–247.
- LIN, S., E. THOMPSON and E. WIJSMAN, 1993 Achieving irreducibility of the Markov chain Monte Carlo method applied to pedigree data. *IMA J. Math. Appl. Biol.* **10**: 1–17.
- LIN, S., E. THOMPSON and E. WIJSMAN, 1994 Finding noncommunicating sets for Markov chain Monte Carlo estimations on pedigrees. *Am. J. Hum. Genet.* **54**: 695–704.
- LUND, M., and C. S. JENSEN, 1998 Multivariate updating of genotypes in a Gibbs sampling algorithm in the mixed inheritance model. *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production Science*, Vol. 25. Armidale, Australia, pp. 521–524.
- MALÉCOT, G., 1948 *Les Mathématiques de l'Hérédité*. Masson et Cie, Paris.
- SELF, S. G., and K. Y. LIANG, 1987 Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.* **82**: 605–610.
- SHEEHAN, N., 1990 Genetic restoration on complex pedigrees. Ph.D. Thesis, University of Washington, Seattle.
- STRAM, D. O., and J. W. LEE, 1994 Variance component testing in the longitudinal mixed effects model. *Biometrics* **50**: 1171–1177.
- THOMPSON, E. A., 1994 Monte Carlo estimation of multilocus autozygosity probabilities. *Proceedings of the 1994 Interface Conference*, Fairfax Station, VA, pp. 498–506.
- THOMPSON, E. A., and S. C. HEATH, 1999 Estimation of conditional multilocus gene identity among relatives, pp. 95–113 in *Statistics in Molecular Biology*, edited by F. SEILLER-MOSEWITZ, P. DONNELLY and M. WATERMAN. Springer-Verlag IMS lecture note series, New York.
- UIMARI, P., and I. HOESCHELE, 1997 Mapping two linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. *Genetics* **146**: 735–743.
- 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.
- VISSCHER, P. M., C. S. HALEY, S. C. HEATH, W. J. MUIR and D. H. R. BLACKWOOD, 1999 Detecting QTLs for uni and bipolar disorder using a variance component method. *Psych. Genet.* **9**: 75–84.
- 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 trait locus. *Genet. Sel. Evol.* **27**: 251–274.
- WILKS, S. S., 1938 The large sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9**: 60–62.
- XIE, C., D. G. GESSLER and S. XU, 1998 Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. *Genetics* **149**: 1139–1146.
- XU, S., and W. R. ATCHLEY, 1995 A random model approach to interval mapping of quantitative trait loci. *Genetics* **141**: 1189–1197.

Communicating editor: T. F. C. MACKAY