The Estimation of Heritability Using Inferred Relationships

Stuart C. Thomas

Ph.D.

University of Edinburgh

2001

Abstract

Estimates of variance parameters, such as heritability, describing quantitative genetic traits in natural populations are of scientific interest in a number of fields such as conservation and evolutionary biology. Central to the estimation of heritability is the covariance of the trait between individuals of known relationship. In natural populations, however, exact relationships may be unknown. In situations where molecular marker data are available, information can be inferred about the relationships without reference to an exact pedigree, and this information may be used to estimate the heritability.

Two existing estimators use inferred relationship information on a pair-wise level: regression of the phenotypic similarity of a pair of individuals on an estimate of their relationship and a likelihood procedure that maximises the probability of their genotypic and phenotypic observations. Computer simulation was used to compare the behaviour of these approaches. Bias in estimates of heritability decreased with increasing sample size, decreased simulated heritability, increasing relatedness and increasing sample size. The regression approach showed less bias than the likelihood approach, but much larger sampling variance. A modified form of the likelihood technique, requiring fewer initial assumptions about population parameters was developed, which showed lower bias in its estimates of heritability than the likelihood technique originally proposed.

An alternative approach in which marker-information was used to reconstruct sibships through relationship assignment within a single generation using Markov chain Monte Carlo (MCMC) techniques was developed. The reconstructed sibships were assumed correct and analysed using restricted maximum likelihood under an animal model. Simulations to compare the properties of estimates with those made using existing techniques indicated that sibship reconstruction was, in many cases, superior to earlier methods, regaining family-specific weighting lost through pair-wise analysis, having lower mean squared errors and showing only slight downwards bias, provided that there was sufficient marker information. Equations appropriate for MCMC analysis of half-sib, full-sib and hierarchical sib-ship structures are presented.

The approaches were extended so that information from other types of marker loci, for example mitochondrial or dominant loci, known maternal information and additional variance parameters can be incorporated into the analysis.

Analysis using the technique was made of a feral population of Soay sheep, with body weight being used as an example trait. Results indicated that the Soay population has a low level of relatedness and so heritability estimates were not reliable, unless inferred relationship data was used only to augment an existing set of known relationships.

In conclusion, the described methods show considerable promise, but are restricted by the need for large family sizes or, equivalently, a large variance of relationship in the sample. In addition they require that about ten polymorphic marker loci be typed per individual before estimated heritabilities become reliable, unless known relationships are also included in the analysis. In consequence they will not be appropriate for all natural populations of interest.

Acknowledgements

I would like to thank the following for their help in the preparation of this thesis.

- My supervisors, Professor Bill Hill and Doctor Josephine Pemberton for their comments, discussions and advice. In particular, I would like to thank Bill for his patience during impromptu statistics lessons and Josephine for her handling of matters bureaucratic.

- Professor Nick Barton for suggestions regarding Chapter 4; his words, if memory serves, "If we move someone from here to here [hand gestures are left as an exercise for the reader] and see how the likelihood of the parameters changes."

- My officemates for screening my work (both scientific and random), before it faced the scrutiny of my supervisors.

- My family and Katherine, for their support and encouragement over the past three years, and their distractions when distractions were required.

- Finally, this acknowledgement thanks Bill for his tireless pursuit of the passive voice.

This work was funded by a Biotechnology and Biological Sciences Research Council Ph.D. studentship.

Table of Contents.

Chapter 1

Introduction.

1.1 Variance components in natural populations.

The variance components that describe quantitative genetic traits have been of interest to animal and plant breeders for many years. They are used to calculate the expected change in the mean value of a trait between parent and offspring, given that only a particular proportion of the parental generation is selected to reproduce (Falconer and Mackay, 1996). In addition they are used in the formation of indices to determine the relative value of individuals within a selection scheme. Selection indices are used to maximise (in terms of profit) the response in an economically important trait, such as milk yield, or in a set of correlated traits (Dekkers and Gibson, 1998). At the same time selection indices must minimise any undesirable correlated responses in other traits, such as fertility (Rauw *et al.*, 1998). Knowledge of variance components has led to a dramatic improvement in animal and plant production. Considerable effort has gone into the estimation of variance components and a number of procedures for their estimation have been developed (Falconer and Mackay, 1996; Lynch and Walsh, 1998).

In more recent years there has been increasing interest in estimating genetic variance components in natural populations, with heritabilities being estimated in hundreds of studies (see meta-analyses of: Mousseau and Roff, 1987; Roff and Mousseau, 1987; Weigensberg and Roff, 1996). The most extensive use of estimates has been to address the questions posed by evolutionary biologists. In evolutionary studies, estimates are important in the understanding of patterns of short-term evolution, the reconstruction of historical patterns of natural selection (Lande, 1979)

and the prediction of genetic responses to selection. Comparison of the variance components describing the same traits within sub-populations of the same species allows inferences to be made about the selection pressures specific to each sub-population and the underlying causes of clinal variation. In addition they provide information on the target of selection in a set of correlated characters. For example, investigations on the fruit fly, *Drosophila melanogaster,* indicated that both wing length and bristle number showed clinal patterns with latitude (Coyne and Beecham, 1987). The observed clinal patterns were either the result of natural selection on both traits, or selection on one trait causing a correlated response in the other. This was tested by comparing the observed slope of the regression line of one trait (the correlated trait) against latitude, with the expected slope based upon the genetic covariance with the other (selected) trait. The expected slope was calculated as the slope of the regression line of the selected trait against latitude multiplied by the genetic regression of the correlated trait against the selected trait, determined from the estimates of the variance components. The observed patterns of variation were consistent with selection on wing length causing a response in bristle number, but not vice versa.

In conservation studies variance components provide information on the number of individuals required in order to maintain a viable population, and so are required for the management of captive populations (Storfer, 1996). Loss of genetic variation is a restricting factor in a species' ability to respond to natural selection, and hence a limitation on its potential to evolve (Lande, 1982; Falconer and Mackay 1996; Lande and Shannon, 1996; Mousseau and Roff, 1987). Variation is therefore critical for maintenance of species within a changing environment.

Whether variance components are sought for evolutionary insight or conservation biology, standard estimation methods (see below) are often difficult or impossible to follow in the wild due to their requirement for known pedigrees. Many estimates have therefore been made under laboratory conditions. But are heritabilities measured in the laboratory the same as in the wild? The more constant environment of the laboratory is expected to reduce trait variation due to the environment, and increase the proportion explained by other causes, thus inflating estimates of the

heritability. Although a meta-analysis of studies failed to support this idea (Weigensberg and Roff, 1996), the preference for such studies 'in the wild' remains.

1.2 Estimating variance components.

1.2.1 Pedigree-based approaches.

Central to the estimation of the variance parameters is the determination of the covariance of a trait between groups of known relationship, and hence knowledge of the relationships or pedigree is required. A number of approaches to estimate this covariance have been adopted in natural populations, including parent-offspring regression, analysis of variance for full-sib or half-sib groups and restricted maximum likelihood (REML).

In parent-offspring regression, the mean phenotypic value of the offspring is regressed against the mid-parent value for the same trait. The slope of the regression line provides a direct estimate of the heritability (Falconer and Mackay, 1996; Lynch and Walsh, 1998). Alternatively, the population may be broken down into groups of different sex, with male offspring and female offspring regressed separately against both male and female parents, allowing more detailed partitioning of variance, with maternal effects, for example, being included into the model. Parent-offspring regression has been used extensively in the estimation of trait heritabilities in natural populations (Mousseau and Roff, 1987).

The basic principles of parent-offspring regression may be extended. For example, lab-reared offspring may be regressed against wild-caught parents to estimate the 'natural' heritability, without the need to raise both generations in the laboratory or extensive observation of the natural population. The previously mentioned study of Coyne and Beecham (1987), on wing length and bristle number in *Drosophila melanogaster*, adopted this approach to parent-offspring regression in order to estimate 'natural' heritabilities. They compared estimates against those made from parent-offspring groups raised solely in the laboratory. The results showed that the two estimates were similar for bristle number, but the 'natural' heritability of wing length was significantly lower than the laboratory based estimate.

A second extension to parent-offspring regression is cross fostering in birds, in which eggs are swapped between nests in order to estimate common environmental effects (e.g. Boag and Grant, 1978; Dhont, 1982). Calculation of the regressions of 'offspring on parents' and 'offspring on foster-parents' allows the estimation of both the heritability and the environmental covariance of full-sibs. This type of study, however, requires the assumption that nest mates are full-sibs and that the adult male bird associated with each nest is the father. Molecular techniques, however, have revealed that this is not always the case (Birkhead and Møller, 1992).

As an alternative to parent-offspring regression, variance component analysis may be carried out on full-sib or half-sib groups. Here, variance is partitioned into between and within group variation using analysis of variance (ANOVA) methodology. The genetic parameters to be estimated are simple functions of the between and within components (Falconer and Mackay, 1996; Lynch and Walsh, 1998). Again, numerous studies have adopted a full-sib approach to variance component estimation (Mousseau and Roff, 1987). For example, full-sib analysis was used to compare laboratory and field estimates of the heritability of wing dimorphism, a threshold trait, in the cricket *Gryllus pennsylvanicus* (Roff and Simon, 1997). In the study each full-sib family was divided in half, with one half raised in the constant conditions of the laboratory and the other half raised in cages outside. The results showed that laboratory based heritability estimates were much larger than in the field (0.70 verses 0.21) and thus supported the theory that laboratory based estimates of heritability are in general larger. The estimate of the genetic correlation between the environments was high indicating that responses to selection in the laboratory would reflect responses to selection in the field.

A drawback of only analysing full-sib groups, is that it is not possible to partition the environmental covariance of full-sibs from the additive genetic variation, possibly introducing upwards bias into estimates of the heritability. The study of Mousseau and Roff (1987) indicated that on average estimates obtained using full-sib correlation were higher than estimates made on the same trait, in the same populations, obtained using by parent-offspring regression, although the difference was not significant.

More recently, sophisticated REML procedures (Patterson and Thompson, 1971; Searle *et al.*, 1992; Lynch and Walsh, 1998), which have been extensively studied and developed in the fields of plant and animal breeding, have been used for variance component estimation in natural populations (e.g. Kruuk *et al.*, 2000; Milner *et al.*, 2000). REML accommodates unbalanced population structures, optimally weighting unequally sized families through the use of a relationship matrix, and thereby making efficient use of the available information. In previous approaches it was difficult or impossible to combine all the data from multiple generations, multiple relationships and uneven family sizes, resulting in the loss of potentially valuable information. As an added appeal, REML techniques are readily expandable, allowing the simple inclusion of additional variance components, such as the environmental covariance of full-sibs, into the model. They may also be expanded to allow the study of longitudinal data, such as the change body weight heritability with age. This type of study, however, requires substantial data sets, and a number of temporal readings on individuals, making them less useful in the study of natural populations, although in principle the techniques may be used.

1.2.2 'Pedigree-free' approaches.

The techniques described in the previous section require that familial relationships be known exactly, or nearly so. In unmanaged populations, detailed knowledge of the relationships is seldom available, unless the population has been intensively studied. Even then the relationship information may be incomplete, or based predominantly on one type of relationship only, e.g. mother-offspring relationships. Information determined from molecular marker loci, however, provides a means to infer information on the relationships without reference to an exact pedigree.

Two general approaches have been developed that allow the estimation of variance components without the need to reference an exact pedigree. The first of these is a regression-based technique (Ritland, 1996b; Lynch and Walsh 1998). In the regression approach, the covariance between estimates of pair-wise relatedness and phenotypic similarity is regressed against an estimate of the variance of the relationship, thereby determining to what extent the phenotypic similarity is

explained by the relationship. The second approach is based upon likelihood techniques, and works by placing pairs into relationship classes of predetermined structure, according to the probability of observing their genotype and phenotype (Mousseau *et al.*, 1998). The likelihood of the observed phenotype is a function of the desired variance components, and so maximising the likelihood function with respect to those parameters provides estimates of them. These techniques have the added appeal that sampling need only occur once, with phenotypic measurements being taken 'on location' and a tissue sample collected for later analysis. Detailed descriptions of the regression and likelihood approaches are provided in Chapters 2 and 3 respectively.

1.3 Inferring relationship data using molecular markers.

The most fundamental issue underlying the techniques investigated in this thesis is the determination of relationship data using molecular marker loci. Additive genetic relationship, r, is defined as twice the probability of drawing an allele from each individual that is identical by descent, and is measured on a scale of zero to one in an outbred population. For example, in unrelated individuals $r = 0$, in half-sibs $r = ¼$ and in full-sibs or parent-offspring pairs $r = ½$. All methods to infer relationship data are therefore based upon individuals sharing co-dominant alleles at autosomal marker loci. The level of similarity due to identity by descent must, however, be distinguished from similarity due to the chance of sampling identical alleles from a finite number of allele-types. It is the purpose of this section to outline the two conceptual approaches to inferring relationship information, while the actual techniques for estimation are detailed within the relevant chapters.

1.3.1 Relatedness.

The first type of measure, termed relatedness, is defined in this thesis as an estimate of the genetic distance between two individuals. The relatedness of a pair is distinct

from their relationship, since it is a measure of their allelic similarity rather than a measure of their exact relationship. Relatedness is therefore an estimate of the actual parameter, the relationship and is measured on a continuous scale. Estimates of relatedness are useful in the study of genetic structure and cooperative behaviour within natural populations (for example: Packer *et al.*, 1991; Taylor *et al.*, 1997; Fjerdingstad *et al.*, 1998), and in examination of the average relatedness between sub-populations (Bernardo, 1993).

Relatedness is estimated by regressing the observed allelic similarity between a pair against some reference value, either one of the individuals of the pair or a population value, while at the same time accounting for the allelic similarity due to chance. One notable feature of relatedness is that the expectation of the average pair-wise relatedness within a sample is zero, when allele frequencies have also been estimated from the sample. Hence negative estimates of pair-wise relatedness must arise, and these are directly attributable to the correction for the allelic similarity due to chance. Despite this, the expectation of relatedness for a given pair is approximately equal to their genetic relationship. This is because the actual average relatedness in a large population is likely to be close to zero, due to the large number of unrelated pairs versus related pairs. Therefore, as the amount of marker information increases the estimate of pair-wise relatedness will approach the true value of the relationship. A number of estimators have been derived (Lynch, 1988; Queller and Goodnight, 1989; Ritland, 1996a; Lynch and Ritland 1999), which use different measures of the genotypic similarity and account for the similarity due to chance in different ways (Chapter 2). In addition, they weight the marker information in alternative manners and thus show differences in their ability to estimate relationship information.

1.3.2 Likelihood approaches.

Alternatively, likelihood techniques may be adopted (Edwards, 1972; Weir, 1996). In likelihood approaches the marginal probabilities that a pair fall into a number of candidate relationship classes are calculated (Thompson, 1975), rather than calculation of a single value for the relationship. A pair can then either be assigned

the relationship with the highest probability, or the probability information for each relationship can be carried forward into subsequent analysis. Likelihood procedures are therefore particularly suited to situations where specific questions are being asked about the relationships, such as is X a full-sib of Y, or is X the father of Y. This feature makes them ideal for use in studies involving particular relationship categories, for example paternity assignment (Meagher, 1986). Confidence levels for each paternity may also be determined by simulation (Marshall *et al.*,1998; Slate *et al.*, 2000; Coltman *et al.*, 1999; Pemberton *et al.*, 1999). Alternatively they may be used to reconstruct pedigrees through identification of close relationships, e.g. parent-offspring and full-sib, on a pair-wise or triplet-wise level.

Likelihoods are calculated by examining the probability of observing the genotype data given that the pair share a particular relationship. The marginal probability that a pair shares a given relationship is then calculated as the likelihood for that relationship divided by the sum of the likelihoods for the candidate relationships. An attractive feature of likelihood techniques is that they may be easily modified to include extra information. For example they may be extended to include information from dominant loci (Thompson, 1975), or include prior information on the population structure. For example in the likelihood technique for estimating variance components (Chapter 3; Mousseau *et al.*, 1998) the prior probability that any given pair falls into a particular class of relationship is assumed known. The genotypic and phenotypic information is then used to update the prior information, resulting in the estimation of the posterior probabilities of the pair fall into each of classes of relationship examined.

Likelihood techniques may also be used to calculate relatedness (Thompson, 1975; Ritland, 1996a; Lynch and Ritland, 1999), however, they require a lot of data in order to be unbiased (a feature of likelihood techniques). Lynch and Ritland (1999) noted that the likelihood approach analogous to their estimator of relatedness became stable only when about 70 diallelic loci were simulated.

1.4 The objectives of this thesis.

Biologists studying natural populations now have a choice when it comes to estimating variance components. Is it better to use marker data to establish pedigrees and use traditional approaches to partition variance components, or should the new marker-based approaches be adopted? To address this question properly, analyses of the properties of the marker-based systems are required. The objectives of this thesis are to:

i). Examine the properties of the existing marker-based approaches, with respect to the amount of marker information, the sample size, the actual heritability, the problem of inaccurate or missing marker data and the population structure (Chapters 2, 3 and 5).

ii). Examine which method of relatedness estimation is the most appropriate for use with the regression approach for variance component estimation (Chapter 2).

iii). Present modified forms of the likelihood approach that require fewer initial assumptions about population parameters and to compare the modified forms against the existing approach (Chapter 3).

iv). Develop and investigate new methods for estimating the variance components using relationship data inferred from marker information and to compare the new methods against the existing approaches (Chapters 2-5).

v). Expand existing and new methodology to allow the inclusion of data from known relatives, information from mitochondrial loci and marker loci with dominant alleles, and to allow the inclusion of the covariance of full-sibs into variance component analysis (Chapter 5).

vi). Examine the performance of different methods on an example data set for the feral Soay sheep population on St. Kilda (Chapter 6).

vii). Discuss the problem of study design (Chapter 7).

The primary interest of this thesis is the estimation of additive genetic variance and environmental variance. These are relatively simple quantities to estimate and it is reasoned that the marker-based approaches will be unable to determine more complex components accurately if they can not first estimate the simpler ones.

Chapter 2

Approaches based upon measures of Pair-wise Relatedness.

2.1 Introduction

All the marker-based systems of variance component estimation work on the same basic principle. On average, relatives share more of their DNA than non-relatives, and the expected percentage of DNA shared is dependent upon the level of the relationship (Falconer and Mackay, 1996; Lynch and Walsh 1998). Since genotype is one of the controlling factors of phenotype, close relatives are expected to be more similar in phenotype than distant relatives. Thus molecular marker data, and relationship measures derived from them, give an indication of the proportion of the DNA shared and are correlated with measures of phenotypic similarity. Equating the relationship information derived from the markers with the phenotypic information therefore allows inferences to be made about the genetic parameters describing the phenotype.

This chapter is designed as an introduction to the regression-based estimation procedure, which was the first of the marker-based systems for estimating variance components introduced (Ritland, 1996b; Lynch and Walsh, 1998). The main aim of this chapter is to present a description of the approach and to use simulation of full-sib families to assess its basic properties with respect to the level of marker information, the sample size, the mean family size and the actual value of the heritability.

There are a number of estimators of pair-wise relatedness available (Lynch, 1988; Queller and Goodnight, 1989; Li *et al.*, 1993; Ritland, 1996a; Lynch and Ritland, 1999), and as a secondary objective, this chapter addresses the question

"which estimator of relatedness yields the most accurate estimates of the heritability?" Despite the differences in sampling variances observed in previous studies of the estimators of relatedness (Lynch and Ritland, 1999) it remains unsure *a priori* which of the estimators of relatedness will yield the best estimates of the heritability. Firstly, the regression-based procedure relies on the estimation of the actual variance of the relationships in the population using a weighted ANOVA (Ritland, 1996b), and as a result small differences in the sampling variance of relatedness estimates should be eliminated. Secondly, the sampling variances of the different measures of relatedness are not ranked in the same order for different levels of relationship, or with different allelic distributions (Lynch and Ritland, 1999). For example, Queller and Goodnight's (1989) estimator has smallest sampling variance with unrelated pairs, but has largest sampling variances when full sibs are considered (Lynch and Ritland, 1999). However, the similarity index (Lynch, 1988; Li *et al.*, 1993) has the third largest (of the four examined) sampling variances with unrelated pairs, but the smallest with full-sibs. This information suggests that, in the case of a sample comprised of full-sib families, the similarity index (Lynch, 1988; Li *et al.*, 1993) is the most appropriate, since there are comparatively small numbers of full-sib pairs in a full-sib family design compared to the number of unrelated pairs, and information about the heritability comes from the difference between the full-sib and unrelated data. However in practice such reasoning may be flawed, and the estimator with the lowest sampling variance over all the estimates of relatedness required in sample analysis might be the most appropriate.

As a final objective this chapter examines the use of estimates of pair-wise relatedness directly in a relationship matrix suitable for use with restricted maximum likelihood (REML). REML techniques make better use of the data through more efficient weighting of information from different relationships and different family sizes (Lynch and Walsh, 1998). The regression-based procedure operates on a pair-wise basis and weights family data according to the number of pairs within which that family appears, rather than weighting by the information content of the family. Using the measures of relatedness in the more efficient REML machinery may help in part to regain appropriate weights lost through pair-wise analysis. Simulation is used in this chapter to compare marker-based REML estimates of heritability with

estimates made using the same estimator of relatedness within the regression framework, and with estimates derived using the simulated relationship matrix.

2.2 Statistical methods.

Throughout the methods *R* has been used to indicate an estimated relationship, and *r* has been used to indicate the actual relationship.

2.2.1 The estimators of pair-wise relatedness.

All estimators of relatedness are based upon the sharing of alleles at marker loci. However sharing due to a common source for the allele (identity by descent) must be distinguished from sharing due to chance. Here four estimators of pair-wise relatedness are examined: The similarity index (SI) (Lynch, 1988; Li *et al.*, 1993), the correlation estimator (CO) (Ritland, 1996a), the regression estimator (RE) (Lynch and Ritland, 1999) and Queller and Goodnight's estimator (QG) (Queller and Goodnight, 1989). The four estimators share the same basic form, with relatedness being estimated by regressing a measure of pair-wise allelic similarity against some reference point, either one of the individuals in the pair or a population value. However the similarity and reference values must be corrected to account for the similarity due to chance. The general form is described by:

$$\text{Relationship} = \frac{(\text{Similarity measure}) - (\text{Correction factor})}{(\text{Reference}) - (\text{Correction factor})}.$$

There are a few basic differences between the estimators. SI and CO use a reference point based upon the population, and are therefore symmetrical estimators, producing identical relatedness measures when X is compared to Y and when Y is compared to X. RE and QG compare one individual (the *proband*) against the other (the reference) and are therefore asymmetrical; X compared to Y does not always equal Y compared to X. With asymmetrical measures, some average measure of the two estimates may be used as the relatedness value. Since RE and QG may give negative

relatedness estimates, the arithmetic average provides the least problematic composite measure.

In addition, the estimators are weighted in different manners: SI is weighted at the locus level, CO is weighted at both the allele and locus level, and RE and QG are weighted at the locus level according to the genotype of the reference individual.

In all cases any population allele frequencies used in the estimator must be recalculated excluding the information from the pair under investigation. Inclusion of this information results in a small positive covariance between the pair-wise and population allele frequencies, introducing bias into estimates. Inclusion of relatives of the individual within the sample also introduces positive covariance, although this is less easily addressed. Chapter 4 outlines an approach that addresses the problem of allele frequency estimation.

Let X and Y be two individuals from a population. X has the genotype (a, b) and Y the genotype (c, d) at a locus where a, b, c and d denote alleles that need not be mutually exclusive. S_{ab} will denote an index variable describing allele identity and is one when a is identical to b and zero otherwise.

2.2.1.1 The similarity index.

The similarity index described here was first introduced by Li *et al.* (1993) and is a modified form of the index proposed by Lynch (1988). SI is estimated from the fraction of identical alleles in X and Y, corrected by the fraction of alleles identical due to chance in an unrelated pair. If X contains an allele of type i the probability of Y containing a similar allele given that Y is unrelated is $p_i^2 + 2p_i(1-p_i)$ where p_i is the frequency of i. Summed over all alleles the expected similarity of X and Y at a single locus due to chance, denoted $S_{0,}$ is

$$\sum_i p_i\left(p_i^2 + 2p_i(1-p_i)\right) = \sum_i \left(2p_i^2 - p_i^3\right).$$

The observed similarity at a single locus is calculated as

$$S_{XY} = 0.25 \cdot (S_{ac} + S_{bc} + S_{ad} + S_{bd}) \cdot \left((1+S_{ab})^{-1} + (1+S_{cd})^{-1}\right).$$

The reference value in this case is one, the population value obtained for a pair having identical genotypes. The estimate for an individual locus, *l*, is therefore expressed as:

$$R_{XYl} = (S_{XY} - S_0) \cdot (1 - S_0)^{-1}. \qquad (2.1)$$

where R_{XYl} is the estimate of the relationship between X and Y for locus *l*.

Since different loci give different amounts of information about the relationship because they have different allelic distributions, multi-locus estimates are obtained using a weighted average of the single locus estimates. However optimal weights are dependent upon the actual relationship between X and Y, which is unknown. Ritland (1996a) argued that since the average actual relationship in a sample is likely to be close to zero, then effective weights can be calculated assuming zero actual relationship. In cases where the actual relationship is not zero the relatedness measure has increased standard error but is unbiased. Locus-specific weights are equal to the inverse of the sample variance of the locus. Sampling variance is readily calculated at the locus level by summation over all pair-wise allelic combinations at that locus, given that the pair have zero relationship. The overall estimator is equal to:

$$R_{XY} = \frac{\sum_l w_l R_{XYl}}{\sum_l w_l} \qquad (2.2)$$

where w_l is the weight for locus *l*.

The estimator takes no account of the frequency of the allele that is identical in the two individuals and is therefore not the most efficient estimator achievable, because rare alleles provide more information on the relationship than common alleles. However, since SI examines all the types of allele simultaneously when calculating S_{XY}, allele-specific weights are difficult to incorporate. CO, described in section 2.2.1.2 below, sums over both allele and locus, and so incorporates both allele-specific and locus-specific weights.

Equation 2.1 can yield negative estimates for the relatedness, when X and Y share no common alleles, but is bounded above by one. Estimates of relatedness made from few marker loci are therefore unreliable.

2.2.1.2 The Correlation estimator.

The probability of identity between a pair for a particular allele type is conditional upon the relationship (Falconer and Mackay, 1996):

$s_{XY} = 0.5 r_{XY} p_i + (1 - 0.5 r_{XY}) p_i^2$,

where s_{XY} is the probability of identity and r_{XY} is the actual relationship between X and Y. The factor of 0.5 appears since genetic relationship is defined as twice the probability of drawing an allele from each individual that is identical by descent. Rearrangement therefore yields an allele-specific estimator for the relatedness:

$$R_{XYIi} = 2 \cdot (S_i - p_i^2) \cdot (p_i - p_i^2)^{-1} \qquad (2.3)$$

where S_i is the probability of drawing two alleles of type i, one from each individual in the pair. e.g. $S_i = 1$ for X = *ii* and Y = *ii*, $S_i = 0.5$ for X = *ij* and Y = *ii*, etc. For this estimator p_i^2 is the correction factor, and is equal to the probability of drawing an allele of type i from two unrelated individuals. The reference value is p_i which in this case represents the probability of drawing an allele of type i from a second individual, i given that the first individual contains an allele i and the pair are unrelated.

Ritland (1996a) calculated allele-specific weights by minimising the variance of the weighted sum of the allele-specific estimates of relatedness conditional upon the pair being unrelated. Efficient allele-specific weights were found to be equal to $(1 - p_i)/(n-1)$ where n was the number of alleles at that locus. Locus-specific weights are then calculated as proportional to the inverse of the sampling variance of the locus-specific estimates of relatedness; which is equal to (n-1) regardless of the allele distribution. The complete estimator is equal to:

$$R_{XY} = 2 \cdot \frac{\sum_{il}\left(S_i - p_i^2\right)}{\sum_{l}(n_l - 1)}. \qquad (2.4)$$

Equation 2.4 can yield negative estimates for the relatedness, and may also give estimates that are greater than one.

2.2.1.3 The regression estimator.

RE is also a method of moments procedure. As described above RE uses one of the individuals in the pair as a reference (the other individual being the proband) and so is an asymmetric estimator of the relatedness. The estimator is equal to

$$R_{XYll} = \frac{0.5\left(p_a\left(S_{bc} + S_{bd}\right) + p_b\left(S_{ac} + S_{ad}\right)\right) - 2p_a p_b}{0.5\left(1 + S_{ab}\right)\left(p_a + p_b\right) - 2p_a p_b}, \qquad (2.5)$$

for a single locus given that X is the reference and Y is the proband. In this case the similarity measure is weighted by the population frequency of the two alleles of X. The reference measure is equal to the similarity measure which would be obtained if the proband had an identical genotype to X. The correction factor is the probability of Y having the same genotype as X. Note that the extra factor of 2 disappears if X is homozygote due to the inclusion of S_{ab} in the denominator.

Multi-locus weights are calculated using the inverse of the sampling variances of the locus given the genotype of the reference individual at that locus. The locus-specific sampling variance may be written in a general form:

$$\mathrm{Var}\left[R_{XYll}\right] = \frac{2p_a p_b}{0.5\left(1 + S_{ab}\right)\left(p_a + p_b\right) - 2p_a p_b} \qquad (2.6)$$

(Lynch and Ritland, 1999).

As described above, a single estimate for the relatedness is obtained by taking the arithmetic average of the estimates obtained when X is the reference and when Y is

the reference. RE gives estimates outside the range zero to one, and is undefined when X is homozygote at a bi-allelic locus where both alleles are equally frequent. In addition the estimator can give intuitively incorrect results when a single locus is examined. For example consider a locus with three alleles *i*, *j* and *k* with frequencies 0.2, 0.3 and 0.5 respectively. A pair with genotypes (*i*, *i*) and (*i*, *i*) has an estimated relatedness of 1, while a pair with genotypes (*i*, *i*) and (*i*, *k*) has an estimated relatedness of 1.1875.

2.2.1.4 The Queller and Goodnight estimator.

Queller and Goodnight (1989) described an estimator for relatedness that has been used primarily for assessing the average relatedness of groups although may be applied to individuals. It is also an asymmetric measure, with a proband and a reference individual. The estimator is derived from the regression of the within individual allele frequencies of the proband against the within individual allele frequencies of the reference. The locus-specific estimator may be written as:

$$R_{XYll} = \frac{0.5(S_{ac} + S_{ad} + S_{bc} + S_{bd}) - p_a - p_b}{1 + S_{ab} - p_a - p_b}, \qquad (2.7)$$

(Lynch and Ritland, 1999), where X is the reference and Y is the proband. Here the similarity measure is twice the average identity of the pair, and the reference is the value of the similarity that would be obtained if the proband had the same genotype as X. The correction factor is the probability of selecting either an allele of type *a* or an allele of type *b* from the proband and equals $p_a + p_b$.

Lynch and Ritland (1999) suggest calculating locus-specific weights using the inverse of the sampling variance for the whole locus. However, like RE, it is better to use the sampling variances of the locus conditional upon the genotype of the reference. Locus-specific sampling variances may then be calculated as:

$$\mathrm{Var}[R_{XYll}] = \frac{0.5(p_a + p_b)}{1 + S_{ab} - p_a - p_b} \qquad (2.8)$$

If the reference is heterozygote at a biallelic locus equation 2.7 is undefined.

2.2.2 A regression-based approach to variance component estimation.

For the sake of clarity, the regression estimator for relatedness will be referred to as RE and the regression approach to variance component estimation as the regression approach. The regression approach examines the data on a pair-wise level, regressing pair-wise phenotypic similarity against a marker-based estimate of the pair-wise relatedness (Ritland, 1996b; Lynch and Walsh, 1998).

The phenotypic similarity (Z_t) for a trait in pair t is defined as the product of the pair-wise phenotypic deviations, and may be viewed as the pair-wise phenotypic covariance:

$$Z_t = (y_t - \bar{y})(y_t' - \bar{y}), \tag{2.9}$$

where y_t and y_t' are the trait values for pair t, $\bar{y}$ is the phenotypic mean of the trait estimated from the sample. Under an additive model, Z_t may be expressed as the product of the additive genetic variance and the genetic relationship between pair t plus a residual error term specific to pair t:

$$Z_t = r_t \sigma_A^2 + e_t, \tag{2.10}$$

where e_t the residual error term and σ_A^2 the additive genetic variance of the trait. Regression theory therefore yields an estimator for the additive genetic variance:

$$\hat{\sigma}_A^2 = C_{Zr} / \hat{V}_r, \tag{2.11}$$

where C_{Zr} is the covariance of all possible pair-wise relationships and phenotypic similarities and $\hat{V}_r$ is the variance of pair-wise relationship. The environmental variance (σ_E^2) may then be estimated using the total phenotypic variance, estimated from the sample.

Replacing the actual value of the relationship with an estimate requires an additional step be added to the estimation procedure. This is because there is now noise associated with the measure of the relationship, thus straightforward calculation of the variance of the relationships results in an overestimate of that variance. Ritland (1996b), assuming that each locus provided an independent estimate of the relationship, outlined an ANOVA partitioning the variance of the relatedness into between and within locus components, with the intraclass covariance providing an estimate of the actual variance of the relationship. Loci do not provide equal amounts of information on the relationship, since they have different allelic distributions and so locus specific weights must be incorporated into the ANOVA. For ease of expression the weights are scaled so that $\sum_l w_l = 1$.

To estimate the intraclass covariance Ritland equated the expected value of squared relationship with the calculated relatedness:

$$R_t^2 = r_t^2 + \sum_l w_l^2 e_l^2 , \qquad (2.12)$$

and the expected weighted sum of squares of the locus-specific estimates with the calculated relatedness:

$$\sum_l w_l^2 R_{tl}^2 = r_t^2 \sum_l w_l^2 + \sum_l w_l^2 e_l^2 , \qquad (2.13)$$

where R_t and R_{tl} are the overall and locus-specific estimate of pair-wise relatedness between pair t and r_t is the actual value of the relationship. By solving equations 2.4 and 2.5 for r_t^2, dividing by the number of pairs and subtracting the square of mean relatedness, Ritland derived an estimator for the actual variance of the relationships:

$$\hat{V}_R = N^{-1}\sum_i \left[\frac{\left(\sum_l w_l R_{il}\right)^2 - \sum_l w_l^2 R_{il}^2}{1-\sum_l w_l^2} \right] - \left(\overline{R}\right)^2, \qquad (2.14)$$

where N is the number of pairs, $\hat{V}_R$ the estimate of the actual variance of relationship and $\overline{R}$ is the mean pair-wise relatedness, equal to:

$$\overline{R} = \frac{\sum_i \sum_l w_l R_{il}}{N}. \qquad (2.15)$$

Ritland (1996b) reports that simulation studies of marker data consisting of full-sib families showed that equation 2.6 recovered estimates of $\hat{V}_R$ that were within two to five percent of the true values.

For the purposes of this study, estimates of heritability are used as summary statistics for the additive and environmental variances. Heritability (h^2) is calculated as:

$$\hat{h}^2 = \hat{\sigma}_A^2 / \left(\hat{\sigma}_A^2 + \hat{\sigma}_E^2\right). \qquad (2.16)$$

2.2.3 Marker-based restricted maximum likelihood.

Restricted maximum likelihood (REML) techniques allow variance component estimation in populations with known pedigrees, with no demands placed on the structure of the pedigree (Lynch and Walsh, 1999). REML weights and incorporates information from different relationship classes and unbalanced data through the inclusion of a matrix describing the variance-covariance structure of the data. In the case of estimating additive genetic variance, the variance-covariance matrix is formed using the known additive genetic relationship between individuals.

The regression approach to variance component estimation loses information about higher order relationships (triplets etc). In addition, it weights information from families according to the number of pairs within which that family is represented, rather than on the information content of that family. An approach that allows relationship information derived from marker information to be used within a REML framework might help regain some of this lost information. Extensions to the REML machinery could be visualised where the likelihood of the population structure based on the observed marker data is incorporated into the REML model. Population structure could therefore also be maximised. However, in practice maximising over all possible population structures is not feasible due to the overwhelming number of possible structures. Chapter 4 describes a procedure that uses this type of approach, first maximising population structure and then estimating variance components, and so it will not be discussed further here. Foulley *et al.*, (1987, 1990) investigated an approach that allowed inclusion of uncertain paternities into a sire evaluation scheme. Their approach attached probabilities to a small number of candidate paternities before maximising the likelihood with respect to the variance components in a Bayesian framework. However, the approach is applicable to situations where there are only a small number of unknown individuals and has limited use in more general situations.

In this chapter pair-wise estimates of relatedness are used to form the basis of a variance-covariance matrix, before standard REML procedures are used to estimate the variance components. For convenience this approach has been termed marker-based REML (MB-REML).

2.3 The simulated populations.

Samples comprised of full sib-families were simulated to investigate the properties of the regression-based estimator, with respect to the level of marker information, the sample size (the number of full-sib families), the mean family size, the actual value of the heritability and the estimator of relatedness used.

Phenotypic data for full-sib data sets were generated using the infinitesimal model (Bulmer, 1980). An individual's phenotype was sampled as:

$$Y_{fj} = \left(\frac{a_{f1} + a_{f2}}{2} \right) + N\left(0, \frac{\sigma_a^2}{2} \right) + N\left(0, \sigma_e^2\right) \qquad (2.17)$$

where Y_{fj} is the phenotypic value of sib j in family f, σ_a^2 is the additive genetic variance, σ_e^2 is the residual or environmental variance, and a_{f1} and a_{f2} are the breeding values of the parents sampled from a N(0, σ_a^2) distribution. The phenotypic variance was set to 1; so $\sigma_a^2 = h^2$ and $\sigma_e^2 = 1 - h^2$. It was assumed that there was no common environmental correlation of sibs.

The simulations were run under different conditions: Marker information was varied, with populations simulated with 2, 5, 10 or 20 alleles at each of 5, 10 or 20 loci; full-sib family sizes were drawn from a truncated (*i.e.*, no null class) Poisson distribution (Po) with the mean set at 2, 5, 10 or 20; simulated heritabilities were set at 0, 0.25, 0.5, 0.75 or 1; and populations with 100, 150, 200, 300, 400 and 800 individuals in total were simulated. Allele frequencies were either assumed known or were estimated from the sample. Two distributions of alleles were considered: rectangular, where n_l alleles at locus l each had a frequency of $1/n_l$ and triangular, where allele i had a frequency of $2i/\left(n_l^2 + n_l\right)$, for $i = 1, 2 \ldots n_l$.

Variance components were calculated using each of the pair-wise measures of relatedness described in Section 2.2.1: The similarity index (SI) (Lynch, 1988; Li *et al.*, 1993), Queller and Goodnight's estimator (QG) (1989), the regression estimator (RE) (Lynch and Ritland, 1999) and the correlation estimator (CO) (Ritland, 1996a). For each estimator the locus specific weights used were those that were appropriate for unrelated pairs, since this was the most common class of relatedness in the simulations. For each pair, allele frequencies were recalculated excluding the pair under investigation. In addition variance components were calculated using the actual simulated relationships, but using the pair-wise regression framework; these estimates have been termed REAL. Each set of conditions was replicated 250 times.

Simulations were also run to investigate the MB-REML approach when the level of marker information and the sample size were varied. The relatedness measure used to form the relationship matrix was RE, which showed lowest sampling variances in the simulations described above. Fewer repetitions (100) and fewer sets of conditions were run due to the slow nature of inverting large matricies with few zero elements. Estimates of variance components obtained were compared against those calculated using RE in the regression framework.

In all cases estimates of heritability were calculated from the estimated variance components (Eqn. 2.16). Heritabilities were then compared against estimates calculated using REML with the simulated pedigree (i.e. using actual relationships). The REML package used was ASREML (Gilmour *et al.*, 1997). Four statistics were calculated for each set of simulations: the deviation of each marker-based estimate from the REML estimate, the sampling variance over simulations, the bias of the marker-based estimate from the simulated parameter, and the mean squared error (MSE) (sampling variance plus squared bias).

2.4 Results.

The figures presented in this section are all taken from the simulations in which triangular distributions of allele frequencies were simulated at each locus, and in which these allele frequencies were estimated from the sample. Differences observed when rectangular distributions were used are described in the text. In all cases there was very little difference found between estimating allele frequencies from the sample or using the correct (simulated) allele frequencies. REML in the figures refers to estimates made using the actual pedigree. Investigation of MB-REML is restricted to section 2.4.5.

2.4.1 Sample size.

The mean difference between the REAL estimates and the REML estimates is close to zero (Fig. 2.1a), indicating that little bias is introduced through pair-wise analysis when relationships are known. Mean difference between the marker-based estimates

and the REML estimates were in general larger than when known relationships are used (Fig 2.1a). The shape of the lines shown by the marker-based estimates as sample size increases is roughly the same, although with different magnitudes. QG-based estimates were generally the lowest, while SI-based estimates were generally the highest. Overall, CO-based estimates show the smallest difference from REML and REAL (Fig. 2.1a).

A more important measure of performance is the MSE of heritability estimates, which falls for all estimators as sample size increases (Fig. 2.1b). The MSE of the REAL estimates are slightly larger than the REML derived estimates. This difference, despite both approaches working with exact relationships, is due to the more efficient way in which REML weights family data, with REML weighting by information content as well as family size. The difference between the MSE of these approaches would be expected to be greater when the distribution of family sizes had a much larger variance than mean or when there are mixed relationship classes within the sample (see Chapter 5). The MSE of marker-based heritability estimates are all much greater than the MSE when known relationships are used. Out of the marker-based approaches, estimates derived using RE estimates of the relatedness show lowest MSE, although the MSEs of CO and QG based estimates are very similar. Heritability estimates derived using SI have distinctly larger MSE than the other three marker-based approaches. This is due to the relatively poorer way in which SI weights locus and allele-specific information. In all cases the MSE was dominated by the sampling variance, indicating that any bias shown by the estimates is comparatively unimportant.

The sampling variance of the REML-derived heritability estimates and pair-wise estimates using actual relationships both fall at a rate that is roughly inversely proportional to the sample size (Fig. 2.1b). The sampling variances of the marker-based estimates fall at a slightly slower rate (Fig 2.1b). This may be an artefact of the simulations or an indication of inefficiency introduced through the inclusion of unknown relationships; perhaps it is an effect of using locus specific weights appropriate for unrelated pairs in the calculation of all relatedness estimates.

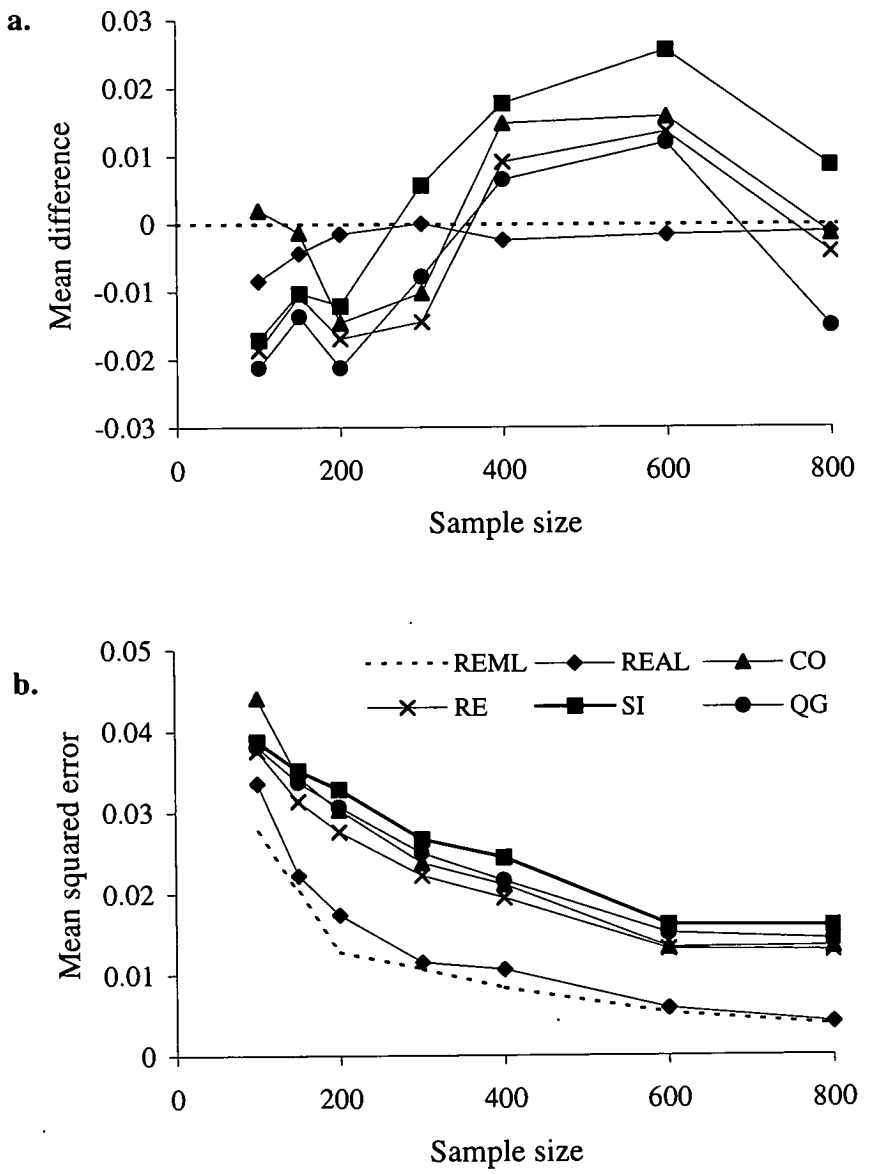

Figure 2.1: Results for the full-sib family simulations investigating the properties of heritability estimates obtained using the different estimators of relatedness, when the sample size was varied. Simulation conditions: Family size distribution ~ Po(10), heritability 0.5 and 10 loci each with 10 alleles arranged in a triangular distribution. (a) Mean difference of the heritability estimates from REML estimates (zero line). (b) Mean squared error of heritability estimates.

When rectangular allele distributions are used there is less difference between the results for the different marker-based estimators. Estimates made using the RE still show MSE closest to the REML derived estimates, although these are even less distinguishable from estimates made using CO.

2.4.2 Mean family size.

When family size is varied, the mean difference between the REAL estimates and the REML estimates is again close to zero (Fig. 2.2a). With small family sizes (mean 2) the heritability estimates based on SI are biased upwards, but follow the same trends as the other estimators at the other family sizes. Overall estimates based on RE are closest to the REAL estimates.

MSE falls with increased mean family size (Fig. 2.2b), with the MSE of marker-based estimates falling rapidly between a mean family size of 2 and 5. In populations of fixed total size there are many more full sib pairs when large families are simulated than when small families are simulated, the number of full-sib pairs being proportional to the square of family size. The lower MSE of estimates made using large family sizes reflects the larger number of full-sib pairs and highlights a requirement for there to be large numbers of related individuals in the sample. Ritland (1996b) described this restriction as the requirement for "*significant variation of actual relationship*", although more strictly this should be the requirement that there be large numbers of related pairs in the sample, since the variation of actual relationship falls with an increase in sample size. This fall is due to the number of unrelated pairs being proportional to the square of sample size, while the number of full-sib pairs is proportional to the number of families. For example in the full-sib structure investigated here, an increased sample size causes a reduction in the MSE (Fig. 2.1b) while variation of actual relationship also falls from about 0.005 to 0.001 between sample sizes 200 and 800 assuming that all families are of size 5.

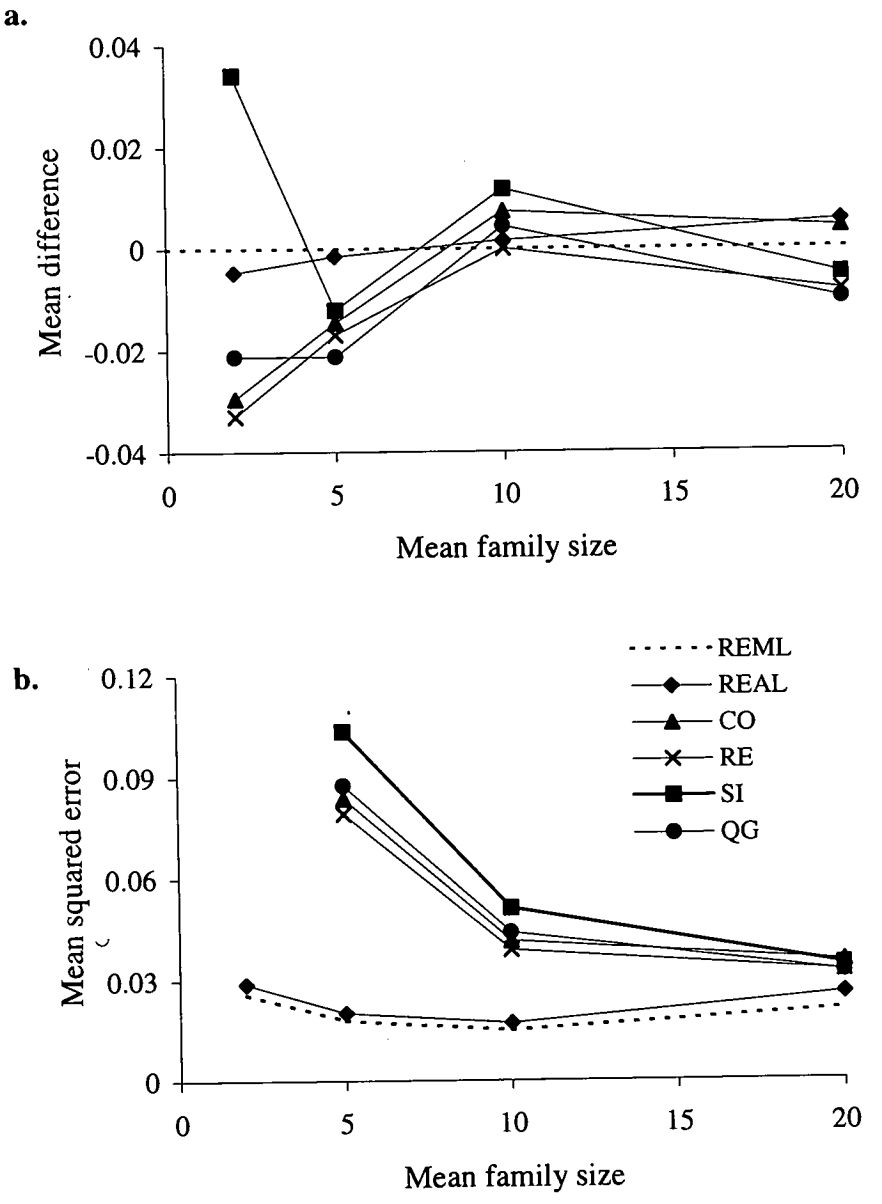

Figure 2.2: Results for the full-sib family simulations investigating the properties of heritability estimates obtained using the different estimators of relatedness, when mean family size was varied. Simulation conditions: Total sample size 200, heritability 0.5 and 10 loci each with 5 alleles arranged in a triangular distribution. (a) Mean difference of the heritability estimates from REML estimates. (b) Mean squared error of the heritability estimates (missing values for family size 2: CO = 0.26, RE = 0.25, SI = 0.5 and QG = 0.3).

When the total sample size is fixed there is a trade-off between the number of families and the number of individuals per family analysed. For example, in a balanced full-sib design the sampling variance is least when each family has a size of $2 / h^2$ (Falconer and Mackay, 1996). Thus, a minimum value is observed for the MSEs of the REML and REAL estimates (Fig 2.2b).

Again the results are closer together when rectangular rather than triangular allele frequency distributions are used.

2.4.3 Marker information.

When the amount of marker information is increased, the mean difference between the marker-based estimates and the REML estimates decreases (results not shown). Mean difference also falls with increasing numbers of alleles per locus, with a doubling of the number of alleles having roughly the same effect as doubling the number of loci. Again the estimates calculated using RE are closest to the estimates derived using REAL.

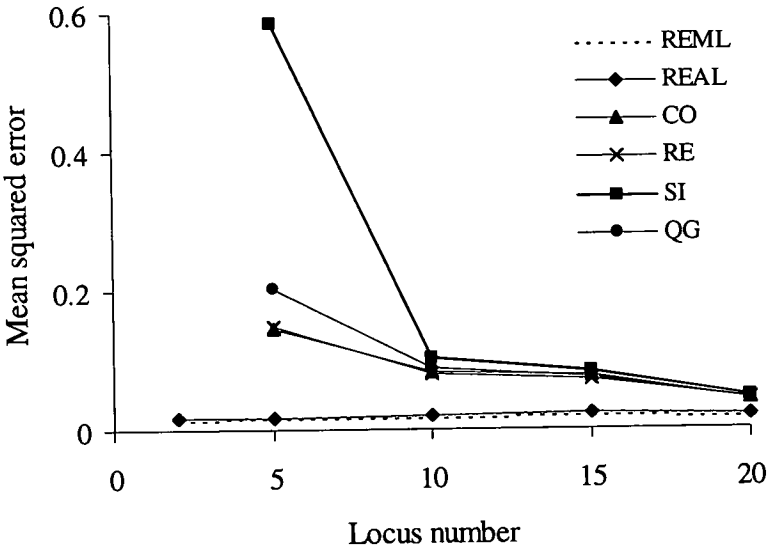

Figure 2.3: Mean squared errors of heritability estimates obtained using the different estimators of relatedness, when marker information was varied. Simulation conditions: Total sample size 200, mean family size 5, heritability 0.5 and each loci had 5 alleles with a triangular distribution. (missing values for two loci simulated: CO = 1.40, RE = 2.99, SI = 1.65, QG = 2.56).

MSE also falls rapidly with increasing locus number, with very large MSE at small numbers of loci (Fig. 2.3). At larger numbers of marker loci, marker-based estimates calculated using RE gave the lowest MSE, although any differences between estimators became marginal with 20 loci.

2.4.4 Simulated heritability.

When simulated heritability is low the marker-based estimates have similar bias to each other (within a range of 0.02) and are close to the REAL estimates. At larger values of simulated heritability, the marker-based estimates are less grouped, and are more distant from the REAL estimates. Again the RE estimates are the closest to the REAL estimates. Of note is a deviation between the mean bias of the REML based estimates and mean bias of pair-wise estimates made from known relationships when additive genetic variance is simulated as zero. This is due to the REML assigning all negative estimates of variance as zero, thereby fixing the additive genetic variance at zero, whereas the regression-based estimator allows negative estimates of variance components. The bias of the REML based estimate is therefore greater than zero ($\approx$ 0.04) with a simulated heritability of zero, while the bias of the regression-based estimator is approximately zero ($\approx$ 0.006).

MSE increases with increasing value of the simulated heritability (Fig. 2.4), with large MSE evident at high heritabilities. Again, SI showed the largest MSE and RE the smallest.

2.4.5 MB-REML estimates.

The relatedness estimator yielding heritability estimates with the lowest sampling variance was the regression estimator of Lynch and Ritland (1999), which was therefore used to estimate a relationship matrix for use in REML.

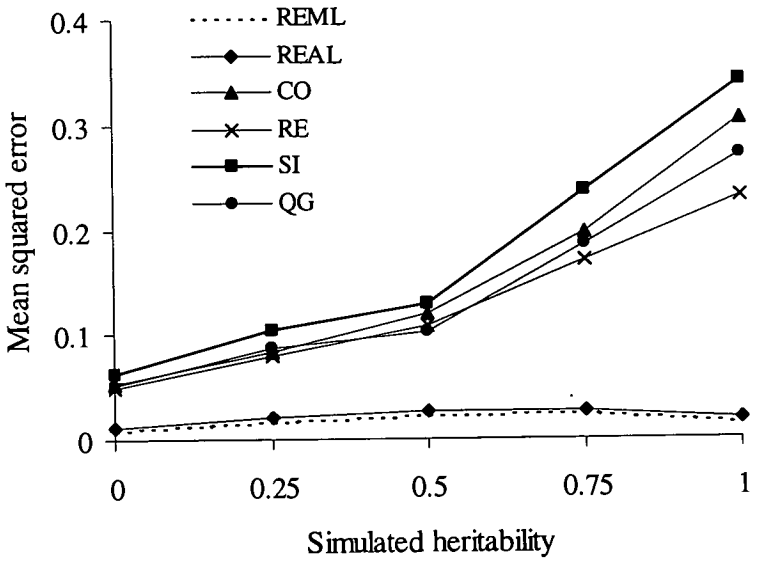

Figure 2.4: The mean squared errors of heritability estimates obtained using the different estimators of relatedness, when simulated heritability was varied. Simulation conditions: Total sample size 200, mean family size 5, and 10 marker loci, each with 5 alleles arranged in a triangular distribution.

2.4.5.1 Marker information.

Estimates of heritability made using MB-REML are biased downwards, although the magnitude of the bias decreases in a roughly linear manner as the amount of marker information increases (Fig 2.5a). In comparison, estimates determined using the pair-wise regression approach are much closer to the REML based estimates using the actual pedigree.

The MSE of heritability estimates made using MB-REML are higher than the MSE of both the regression-based and REML approaches (Fig 2.5b). When sampling variance is considered rather than MSE, the sampling variance of the MB-REML is the lowest, much lower even than the REML-based estimates, reflecting the substantial downwards bias.

When the allele frequencies at each locus follow a triangular distribution similar results to those presented in Figure 2.5 are obtained, but with a slightly increased downwards bias seen in the estimates made using MB-REML.

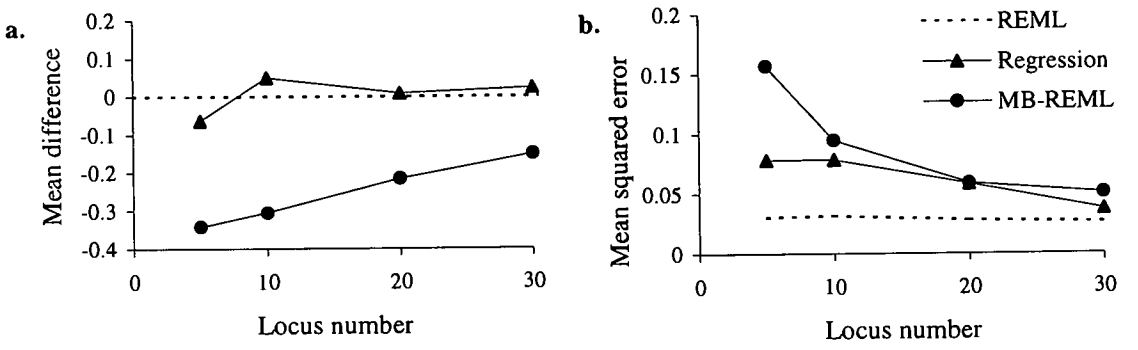

Figure 2.5: Results for the full-sib family simulations investigating marker-based REML when the number of loci was varied. Simulation conditions: Sample size 200, family size distribution ~ Po(5), heritability 0.5 and 10 equally frequent alleles per locus. (a) The mean difference of the heritability estimates from the REML estimates made from using actual pedigree. (b) The mean squared error of the heritability estimates.

2.4.5.2. Sample size.

Heritability estimates observed using MB-REML are again biased downwards, with the bias becoming larger as the sample size increases (Fig. 2.6a). This rather counter-intuitive result may be explained by examination of the relatedness estimates used in the relationship matrix. Apart from situations where allele frequencies are estimated from the sample, when there is a small decrease, the actual level of noise seen in each estimate of relatedness is independent of sample size. The overall noise in the relationship matrix therefore actually increases with increasing sample size, since c new estimates of relatedness are introduced to the relationship matrix for every extra individual added to the sample (where c is the current sample size). As a consequence the amount of variance subsequently attributed to the relationships within the sample decreases, resulting in underestimates of the heritability. Again the regression-based estimates show little bias in comparison.

MSE of MB-REML heritability estimates increases as sample size increases (Fig. 2.6b). Again this is due to the increase in the amount of noise introduced

through an increase in the number of relationships requiring estimation. The MSE of the regression-based estimates is much closer to the MSE of the REML estimates (Fig 2.6b) and as with Fig. 2.1b fall with almost the inverse of sample size.

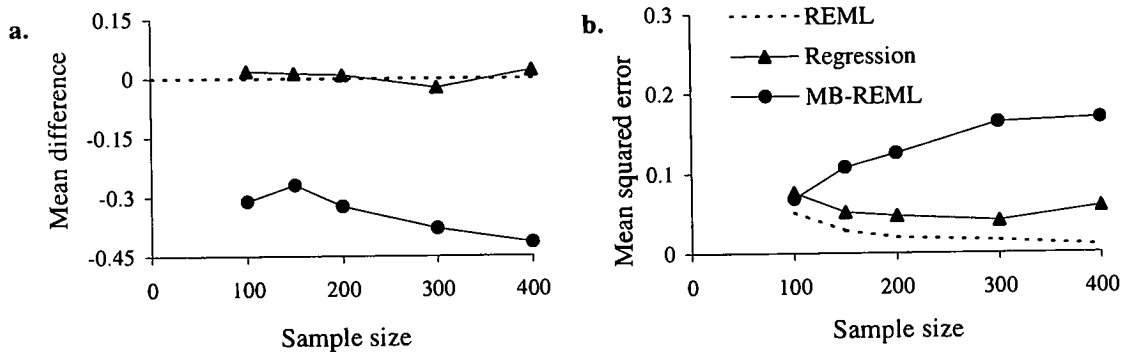

Figure 2.6: Results for the full-sib family simulations investigating marker-based REML when the sample size was varied. Simulation conditions: Family size distribution ~ Po(5), heritability 0.5 and 10 loci each with10 equally frequent alleles. (a) The mean difference of the heritability estimates from REML estimates made using actual pedigrees. (b) The mean squared error of the heritability estimates.

2.5 Discussion.

The regression approach provides a means to estimate heritabilities in natural populations. However, although it gives essentially unbiased estimates, the sampling variance and mean squared errors of estimates are large. Estimates of heritability would only be reliable in populations with large families, where sample sizes are large and where there is ample marker information. The ability of the regression-based procedure to accurately estimate heritabilities increases with increasing numbers of marker loci, increasing sample size and increasing family size. Estimates of heritability also improve at lower levels of actual heritability, but since the measures of relatedness are independent of the phenotype data, this is just a scaling effect. The sampling variance of the marker-based estimates falls when sample size increases, although at a slightly slower rate than the sampling variance of estimates made using known relationships, indicating some loss in efficiency in the technique

through the introduction of unknown relationships. This is perhaps an effect of using locus specific weights appropriate for unrelated pairs in the determination of relatedness for all pairs, regardless of the true relationship. Alternatively this might reflect inefficiency in the procedure for estimating the actual variance of relationship. In the weighted ANOVA procedure described in section 2.2.2 it is assumed that the within pair estimates of the relatedness have the same variance regardless of the true relationship, but this may not be correct. Previous studies have found that the level of the relationship does affect the sampling variance of the estimators of relatedness (Lynch and Ritland, 1999), hence different relationships are not estimated with equal accuracy. To account for the differences in the accuracy of the estimates of the different classes of relationship requires weighting pair-wise information according to the inverse of the sampling variance of the actual pair-wise relationship. This clearly requires *a priori* knowledge of the relationships within the sample, and is therefore impossible.

Of the four estimators of relatedness investigated here (the similarity index (Lynch, 1998; Li *et al.*, 1993), Queller and Goodnight's estimator (1989), the correlation estimator (Ritland, 1996a) and the regression estimator (Lynch and Ritland, 1999)), the regression estimator showed lowest mean bias and lowest mean squared error. There was, however, little actual difference in terms of performance between it and the correlation estimator.

Incorporation of relatedness measures directly into a relationship matrix for use with REML techniques was not successful. As found with previous studies (Van Vleck, 1970a, b) inaccuracies in the relationship matrix led to downwards bias in the heritability estimates made. Even with high levels of marker information, downwards bias still makes this approach less useful than the regression-based approach. In addition, simulations show that MSE of estimates actually increases with increasing sample size.

In this study allele frequencies were either taken as the simulated value, or were estimated from the sample. In all cases there was very little difference found between using calculated or actual frequencies. If allele frequencies are estimated from the sample under investigation they are subject to further random error, since there are relatives within the sample, which might bias subsequent relationship

estimates. To combat this problem, allele frequencies for each candidate pair were re-estimated excluding the information from that pair. This approach removed a small covariance between population and pair-wise allele frequencies and resulted in slightly improved estimates, although improvement was negligible with large samples.

The question of experimental design must also be addressed, with questions arising about whether it is more important to collect large sample sizes, or genotype more loci? This issue will be returned to in Chapter 7, and will be broadened to include other marker-based approaches for estimating variance components, and so will not be discussed here. The effects of using more relationship classes in the sample are also not investigated here but are returned to in a later chapter (Chapter 5).

A final consideration is how the simple model investigated here may be extended to incorporate other factors (assumed zero in these simulations). In natural populations many fixed effects (e.g. sex, or year of birth) may have considerable influence on quantitative traits and these additional effects must be included in the model to allow estimation of the variance components. For example a significant year effect on bill depth was noted in Darwin's Finches (*Geospiza*), caused by larger parents in certain years (Boag, 1983). Because the methods investigated here are based on pair-wise computations rather than on individuals it is harder to incorporate fixed effects into the model. One simple solution to this problem would be remove all fixed effects prior to estimation, using least squares techniques, thereby estimating values for each animal that are based only on the random effects. Marker-based analysis may then be used to estimate variance components from these calculated animal values. This approach would introduce bias by not accounting for the change in the degrees of freedom encountered in estimating the fixed effects. However in a large population such bias would be negligible.

In addition the model described here partitions the variance into additive genetic and environmental causes only. Ritland (1996b) outlined how the variation arising due to inbreeding, local environment and dominance may be included into the model. However these require that more parameters be incorporated into the model, and so result in larger MSE for estimates, undesirable given the already large MSE

observed in simulations compared to the MSE of known pedigree estimates. Indeed, simulations performed by Ritland indicated that under a model including dominance and additive genetic variance, the variance of subsequent narrow and broad sense heritabilities were often "*extreme*." Approaches that reduce the observed MSE in the simple case are required before more complex models are considered.

Chapter 3

A likelihood-based approach

3.1 Introduction

In Chapter two a regression-based technique for marker-based inference about quantitative genetic variation was described and the properties of the procedure were examined using simulated full-sib samples. The regression-procedure is not restricted by the population structure and requires no prior assumptions about the population. However it requires that there be large sample sizes, with large numbers of related pairs in the sample and a reasonably large amount of marker information before estimates of variance components become reliable. In some situations additional information may be available about the population structure and so other approaches to variance component estimation may become applicable. An example is a situation where the sample is assumed to contain only unrelated and full-sib pairs and the prior probability of each of these relationship classes is known (or derived from the average full-sib family size).

In this chapter the second of the marker-based approaches to variance component estimation, which is based upon likelihood techniques (Edwards, 1972), is described and investigated. The technique was first introduced by Mousseau *et al.* (1998), and requires that prior knowledge of the sort described above is available. In brief, the approach works by calculating the likelihood of each pair falling into a number of predetermined relationship classes according to the prior information and the likelihood of their pair-wise genotypes and phenotype given the relationship. Since the likelihood of the pair-wise phenotype is dependent upon the desired variance components, maximum likelihood estimates for them can be obtained through maximising the likelihood with respect to those parameters. A number of

different functions that describe the joint phenotypic distribution may defined; Mousseau *et al.* (1998) originally proposed using a function based on the sum of the normalised trait values. This requires that both the sample mean and sample variance be estimated prior to analysis. Other functions that require fewer parameters to be estimated prior to analysis can be determined, e.g. the pair-wise difference of phenotypes that have not been normalised. In this chapter investigation is made of a number of these functions using simulation and analytical techniques to determine their statistical properties. In addition the performance of the likelihood approach is compared to the performance of the regression approach using the Lynch and Ritland (1999) estimator for relatedness (RE of Section 2.2).

For both pair-wise approaches a concern is that they lose data from the higher order groups of relations, such as triplets, that are present in the sample. For example, if three individuals sampled from a single generation have genotypes a_ia_i, a_ja_j and a_ka_k (a_i, a_j and a_k being mutually exclusive alleles) they cannot be full sibs; but with pair-wise analysis such an exclusion is not possible. In the previous chapter the use of relatedness estimates directly in a relationship matrix suitable for REML analysis was investigated in an attempt to regain some of this lost information. However the relatedness measures used in the matrix were still based only on pairs. Moreover the noise associated with the relatedness measures resulted in heritability estimates that were very biased downwards. Another approach would be to extend the pair-wise procedures to examine triplets. With triplets, families would still be weighted by size, through the number of triplets within which they appear, but extra information from exclusions would be included. The regression-based technique is difficult to extend and requires that parameters describing the third moments of relatedness and additive genetic effects be introduced (Ritland, 1996b), the interpretation of which is difficult. The likelihood approach can, however, be readily extended to the triplet case, without the need to introduce extra parameters. Simulation of full-sib families is used in this chapter to investigated the use of triplet-wise analyses.

3.2 Statistical Methods.

3.2.1 The pair-wise likelihood technique.

The likelihood technique is designed for use on a sample of individuals which have been genotyped at a number of marker loci and have been measured for the phenotypic value (y) of the quantitative trait of interest. The likelihood-based procedure is applicable in situations where some prior knowledge of population structure is known (Mousseau *et al.*, 1998). In the case of a population comprised only of full-sib families this prior knowledge would be the probability that a pair of individuals randomly selected from the population are full-sibs.

Pair-wise genotype	Unrelated	Full-sib
$A_iA_i - A_iA_i$	p_i^4	$p_i^2(1+p_i)^2/4$
$A_iA_i - A_iA_j$	$4p_i^3p_j$	$p_i^2p_j(1+p_i)$
$A_iA_i - A_jA_j$	$2p_i^2p_j^2$	$p_i^2p_j^2/2$
$A_iA_j - A_iA_j$	$4p_i^2p_j^2$	$p_ip_j(1+p_i+p_j+2p_ip_j)/2$
$A_iA_i - A_jA_k$	$4p_i^2p_jp_k$	$p_i^2p_jp_k$
$A_iA_j - A_iA_k$	$8p_i^2p_jp_k$	$2p_ip_jp_k(1+2p_i)$
$A_iA_j - A_kA_l$	$8p_ip_jp_kp_l$	$2p_ip_jp_kp_l$

Table 3.1: The probabilities for the possible pair-wise genotypes observed in diploid individuals in unrelated and full-sib pairs. i, j, k and l denote mutually exclusive alleles and p_i denoting the frequency of allele i.

There are seven possible genotype patterns observable at a single locus with co-dominant alleles in a pair of diploid individuals. The probabilities of observing these patterns given the relationship between the pair may be calculated for a given relationship (Thompson, 1975) (Table 3.1). The probabilities are multiplied across

loci, assuming independence between loci (i.e. unlinked loci) to give the probability of the observed molecular data given a particular pair-wise relationship.

The phenotypic information also provides information on the relationship, because the distribution of some function of a pair's phenotypes is dependent on the level of the relationship between the pair. The three types of information: the prior information, the molecular information and the phenotypic information, are combined to give the joint likelihood of the observed data:

$$L = \prod_{b} \left(\sum_{r} a_r m_{b|r} z_{b|r} \right), \tag{3.1}$$

where L is the total likelihood for the population, b indexes a particular pair, r indexes a particular class of relationship (e.g. full-sib, half-sib, unrelated), a_r is the prior probability of a random pair sharing relationship r, $m_{b|r}$ is the likelihood of the molecular data of pair b given relationship r, and $z_{b|r}$ is the probability density of the phenotypic data for pair b given relationship r and the population parameters, such as the additive genetic variance, to be estimated.

A function that combines the phenotypic data from a pair is required to allow calculation of the likelihood of the observed phenotypes, given the relationship and the variance. Assuming that the trait under consideration is normally distributed, then the joint distribution of the observed phenotypes is a multivariate normal (MVN) distribution where the covariance term is dependent upon the relationship of the pair (Table 3.2). Linear transformation of the observations can be used to provide simpler distributions. Table 3.2 shows four such transformations, the associated linear function(s) and the associated unrelated and full-sib distributions.

	Transformation	Function(s)	Parameters	Distribution (unrelated)	Distribution (full-sib)
Original distribution	-	$\begin{bmatrix} y \\ y' \end{bmatrix}$	a, μ	$\text{MVN}\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_A^2+\sigma_E^2 & 0 \\ 0 & \sigma_A^2+\sigma_E^2 \end{bmatrix}\right)$	$\text{MVN}\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_A^2+\sigma_E^2 & \sigma_A^2/2 \\ \sigma_A^2/2 & \sigma_A^2+\sigma_E^2 \end{bmatrix}\right)$
NSUM	$[1/\hat{\sigma}_{TOT} \quad 1/\hat{\sigma}_{TOT}]$	$\dfrac{y_i + y_i' - 2\mu}{\sigma_{TOT}}$	a, μ, σ_{TOT}	N(0, 2)	$N(0, 2+h^2)$
SUM	[1 1]	$y_i + y_i' - 2\mu$	a, μ	$N(0, 2\sigma_A^2+2\sigma_E^2)$	$N(0, 3\sigma_A^2+2\sigma_E^2)$
DIFF	[1 -1]	$y_i - y_i'$	a	$N(0, 2\sigma_A^2+2\sigma_E^2)$	$N(0, \sigma_A^2+2\sigma_E^2)$
BOTH	$\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$	$\begin{bmatrix} y_i - y_i' \\ y_i + y_i' - 2\mu \end{bmatrix}$	a, μ	$\text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2\sigma_A^2+2\sigma_E^2 & 0 \\ 0 & 2\sigma_A^2+2\sigma_E^2 \end{bmatrix}\right)$	$\text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_A^2+2\sigma_E^2 & 0 \\ 0 & 3\sigma_A^2+2\sigma_E^2 \end{bmatrix}\right)$

Table 3.2: Summary table of the four phenotypic models investigated showing the linear functions, and their distributions, and the transformation matrix used to derive the function from the original multivariate normal distribution. Also shown are the parameters that are required before analysis using each model. a is the prior probability that a randomly picked pair are full-sibs, μ and σ_{TOT} are the population mean and standard deviation respectively; in practice these are estimated from the sample. σ_A^2 is the additive genetic variance, σ_E^2 the environmental variance and h^2 the heritability. (Note: $\sigma_{TOT}^2 = \sigma_A^2 + \sigma_E^2$).

The functions require that different numbers of population parameters be estimated prior to the likelihood calculation (Table 3.2). NSUM, which was the function originally proposed by Mousseau *et al.* (1998), is equivalent to the sum of the normalised trait values and requires the greatest number of parameters be estimated prior to calculation. SUM and DIFF use the sum and difference of the observed trait values respectively, and require that fewer parameters be estimated prior to calculation. BOTH is a combined form of the SUM and DIFF functions, since these functions are uncorrelated, and contains all the information that is in the original multivariate form. It is desirable to have fewer parameters requiring estimation prior to calculation as they lead to bias in subsequent estimates of the variance.

Maximization of likelihood equation 3.1 using standard iterative procedures (e.g. the Newton-Raphson algorithm (Edwards, 1972; Weir, 1996)) yields the maximum likelihood estimates for the variance components of the distributions associated with the linear function used (Table 3.2). NSUM estimates heritabilities directly, whereas SUM, DIFF and BOTH require use of equation 2.6. Slight bias is introduced into heritability estimates derived from variance component estimates, since the ratio of two expectations need not be the same as the expectation of a ratio. However bias is very small compared to the sampling variance of the estimator, and so error in the heritability estimate will be mainly due to the sampling variance. Simultaneous estimation of the mean during analysis yields the same estimate as straightforward calculation of the sample mean, thus maximisation of 3.1 was with respect to the variance components only.

3.2.2 Bias in pair-wise techniques with full pedigree information.

Variance components were estimated for a full-sib design using correct pedigree information (i.e. using the known relationships) using the pair-wise technique and restricted maximum likelihood (REML; Patterson and Thompson, 1971; Searle *et al.*, 1992; Lynch and Walsh, 1998). REML accommodates unbalanced population structures, optimally weighting unequally sized families through the use of a relationship matrix, and thereby making efficient use of the available information. REML estimates were used as reference values for the best available parameter

estimate for a particular population. Pair-wise parameter estimates were compared against the REML estimates. Both balanced and unbalanced family structures were examined.

3.2.2.1 The balanced case.

In the balanced case REML yields the ANOVA estimates for variance components and is unbiased. When correct pedigree information is used in the regression-based approach of Chapter 2, the covariance of phenotypic similarity with relationship and the actual variance of the relatedness are calculated without bias. Heritability estimates are identical to the restricted maximum likelihood-derived estimate as families are the same size and are therefore equally weighted in the calculation.

Bias in the pair-wise likelihood approaches can be assessed by taking the expectations of the square of the linear functions described in table 3.2 for each relationship category, given that the relationships are known. DIFF is shown to be unbiased, while SUM, NSUM and BOTH are biased because they include the sample estimates of the mean and standard deviation rather than the population values. The derivation of the bias of SUM is summarised in Appendix 1. The proof of the unbiased nature of DIFF and derivation of the bias of NSUM follow the same format as the proof for DIFF, and so are not shown.

For SUM the bias affects only the estimate of σ_E^2, which has an expected value less than the REML-derived estimate of σ_E^2 by:

$$\frac{2}{nf}\left(\hat{\sigma}_E^2 + \frac{(1+n)}{2}\hat{\sigma}_A^2\right), \qquad (3.2)$$

where n is the family size, f is the number of families and $\hat{\sigma}_A^2$ and $\hat{\sigma}_E^2$ are the REML estimates of the additive and environmental variance. Estimates of heritability are therefore upwardly biased (Eqn 2.15). This bias is removed if the actual population mean is known and used in place of the sample mean (Appendix 1).

The expectation of the estimate of heritability from NSUM deviates from the REML- derived estimate of heritability by:

$$\frac{4\left(\frac{1}{nf}-1\right)\hat{\sigma}_E^2+2\left(\frac{nf}{2}+\frac{1}{nf}-\frac{n}{2}-\frac{1}{f}-2\right)\hat{\sigma}_A^2}{(nf-1)\hat{\sigma}_E^2+(2nf-n-1)\frac{\hat{\sigma}_A^2}{2}}-\frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2+\hat{\sigma}_E^2}. \qquad (3.3)$$

Again the bias is a result of using estimates of population parameters in the likelihood calculation and disappears if the true population mean and variance are used. With NSUM and SUM the bias decreases with larger sample sizes.

It is not possible to determine the expected bias of BOTH estimates analytically. This is because three equations describing V_w and V_b can be derived from the model using expectations (as in Appendix 1) and the known relationships, while there are only two unknowns. Closed solutions to the estimates of V_w and V_b cannot therefore be formed and maximum likelihood techniques must be adopted. Simulated studies show that bias is generally downwards and usually between a third to a half of the bias seen when using NSUM.

3.2.2.2 The unbalanced case.

REML estimates of variance components weight families according to the phenotypic correlation and the family size (Lynch and Walsh,1999; Patterson and Thompson, 1971; Searle *et al.*, 1992). Because the weights given to each family depend upon the estimates of the parameters, REML techniques yield slightly biased estimates of variance components in the unbalanced case.

In both regression and likelihood marker-based procedures, pairs are given equal weighting, and as a result families are weighted only by the number of pairs within which they are represented, and not by the phenotypic correlation. Variance component estimates therefore differ from the REML derived estimates and have higher sampling error.

3.2.2.3 Incomplete pedigree information.

When the exact nature of the relationships is unknown, marker-based estimates of relationships must be used. The use of inferred relationships in these estimators may

cause bias, introduced through estimating relationships. Assessment of this sort of bias is most easily achieved through simulation.

3.2.3 The triplet-wise likelihood technique.

Extensions to the likelihood approach allow triplets to be investigated rather than pairs. Triplets allow exclusions due to incompatible genotypes, and may therefore contain more information than simple pair-wise analysis. Here, simulated full-sib data sets are used to investigate the properties of the triplet-wise analysis, and so a version of the technique applicable to a full-sib design will be presented.

To include triplets Equation 3.1 must be modified to:

$$L = \prod_{t} \left(\sum_{\mathrm{R}} a_{\mathrm{R}} m_{t|\mathrm{R}} z_{y_t^1, y_t^2, y_t^3|\mathrm{R}} \right), \tag{3.4}$$

where t indexes all the possible triplet combinations, R is a set of candidate relationships for the triplet and y_t^1, y_t^2 and y_t^3 are the phenotypes of the three members of t. In a full-sib family structure there are five possible combinations for the relationships between the individuals, 1, 2 and 3: All three may be unrelated; 1 and 2 may be full-sibs and 3 unrelated; 1 and 3 may be full-sibs and 2 unrelated; 2 and 3 may be full-sibs and 1 unrelated; all three may be full sibs.

With triplets there are 23 different unordered genotypic patterns at a single locus. A general approach to calculation of the likelihood of observing the 5 sets of relationships is therefore appropriate. The likelihood of a putative full-sib family at a single locus may be expressed as:

$$L_{\text{family}} = \sum_{w=1}^{m} \sum_{x=1}^{m} \sum_{y=1}^{m} \sum_{z=1}^{m} \left[p_{wx}^1, p_{yz}^2 \prod_{c=1}^{n} L(g_c) \right], \tag{3.5}$$

where m is the number of alleles at the locus; w, x, y and z index alleles; n is the number of individuals in the family; c indexes an individual from the putative family,

p^1_{wx} is the ordered genotype frequency of parent one, p^2_{yz} is the ordered genotype frequency of parent two and $L(g_c)$ is the likelihood of observing the genotype of c at the locus given the parental genotypes. For example if p^1 and p^2 share the genotype (1, 2) then $L(g_c) = 0.25$, when the offspring genotype, g, is (1, 1) or (2, 2); $L(g_c) = 0.5$ when g is (1, 2); and $L(g_c) = 0$ otherwise. Equation 3.5 is multiplied across loci to give the total likelihood for a single family and across families to give the likelihood of a set of relationships. Equation 3.5 is not the most efficient approach to calculating the triplet-wise likelihoods for the five sets of relationships, but it is the most general, being easily extended to calculate the likelihood of half-sibs or nested full-sib/half-sib designs (Appendix 2). In addition, it is applicable to the calculation of the likelihood of any size of full-sib family. For a more efficient approach to the case of just full-sibs see Painter (1997). The order of the individuals within the triplet (or group of any size) becomes unimportant since the likelihoods of the relationship sets are compared only against each other.

Only three phenotypic distributions need be considered, however, since the order of the individuals does not affect the joint likelihood of their phenotypes. The three distributions describe the phenotypes when the three are unrelated, when two are full-sib and one is unrelated and when all three are full-sib. These may be expressed as a multi-variate normal (MVN) where the covariance terms only depend on the relationships within the triplet, with, for example, a full-sib pair plus unrelated having a distribution:

$$\begin{bmatrix} y^1 \\ y^2 \\ y^3 \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \mu \\ \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma^2_A + \sigma^2_E & \sigma^2_A/2 & 0 \\ \sigma^2_A/2 & \sigma^2_A + \sigma^2_E & 0 \\ 0 & 0 & \sigma^2_A + \sigma^2_E \end{bmatrix} \right),$$

where y^1 and y^2 are the full-sib pair. Again transformation allows simpler distributions to be determined (Table 3.3). TDIFF is the triplet equivalent of DIFF, examining the contrasts between the phenotypes and removing the need to estimate the sample mean. TSUM is the triplet equivalent of SUM.

Structure	TSUM	TDIFF
Transformation	[1 1 1]	$\begin{bmatrix} 1 & -1 & 0 \\ 1 & 1 & -2 \end{bmatrix}$
All unrelated	$\left[y^1 + y^2 + y^3 - 3\bar{y}\right] \sim \mathrm{N}\left(0, \left[3\sigma_A^2 + 3\sigma_E^2\right]\right)$	$\begin{bmatrix} y^1 - y^2 \\ y^1 + y^2 - 2y^3 \end{bmatrix} \sim \mathrm{MVN}\left(0, \begin{bmatrix} 2\sigma_A^2 + 2\sigma_E^2 & 0 \\ 0 & 6\sigma_A^2 + 6\sigma_E^2 \end{bmatrix}\right)$
$y^1 + y^2$ full-sib, y^3 unrelated *.	$\left[y^1 + y^2 + y^3 - 3\bar{y}\right] \sim \mathrm{N}\left(0, \left[4\sigma_A^2 + 3\sigma_E^2\right]\right)$	$\begin{bmatrix} y^1 - y^2 \\ y^1 + y^2 - 2y^3 \end{bmatrix} \sim \mathrm{MVN}\left(0, \begin{bmatrix} \sigma_A^2 + 2\sigma_E^2 & 0 \\ 0 & 7\sigma_A^2 + 6\sigma_E^2 \end{bmatrix}\right)$
All full-sib	$\left[y^1 + y^2 + y^3 - 3\bar{y}\right] \sim \mathrm{N}\left(0, \left[6\sigma_A^2 + 3\sigma_E^2\right]\right)$	$\begin{bmatrix} y^1 - y^2 \\ y^1 + y^2 - 2y^3 \end{bmatrix} \sim \mathrm{MVN}\left(0, \begin{bmatrix} \sigma_A^2 + 2\sigma_E^2 & 0 \\ 0 & 5\sigma_A^2 + 6\sigma_E^2 \end{bmatrix}\right)$

* Since the order of the individuals does not affect the joint likelihood of their phenotypes, phenotypes may be swapped around so that y^1 and y^2 are always the putative full-sib pair when calculating the likelihood of the triplet being a full-sib pair and an unrelated.

Table 3.3: Summary of phenotypic models used in TSUM and TDIFF.

TDIFF AND TSUM can be combined in a similar way to DIFF and SUM, since they are uncorrelated, giving the model TBOTH. Maximising 3.4 with respect to the desired parameters allows estimation of V_w and V_b.

The likelihood approach could be extended to a tetrad-wise analysis in a similar manner, however computation time becomes a significant problem. In addition, it is questionable whether tetrad analysis would add much new information because in a full-sib design there is no tetrad that leads to an exclusion where all of the four triplets within that tetrad are possible (non-excluded). This fact is easily proved using an exhaustive search of all genotype combinations using Equation 3.5.

3.3 The simulated populations.

Phenotypic values for full-sib data sets were generated in the manner described in 2.3. The simulations were run under different conditions: Marker information was varied, with populations simulated with 5, 10 or 20 alleles at each of 2, 4, 6, 10, 15, 20 or 30 loci; population structure was altered, with full-sib families either having a constant size of either 2, 3, 5, 10 or 15 or being drawn from a truncated (*i.e.*, no null class) Poisson distribution (Po) with the mean set at 2, 5 or 10; simulated heritabilities were set at 0, 0.25, 0.5, 0.75 or 1; and populations with 50, 100, 150, 200, 300, 400, 600, 800 and 1000 individuals in total were simulated. In all cases allele frequencies were calculated from the sample, and were simulated as having a rectangular (uniform) distribution. Heritabilities were used as a summary statistic and were calculated using each of the phenotypic functions described above: NSUM (Mousseau *et al.*, 1998), SUM, DIFF and BOTH. In addition heritabilities were calculated using the regression-based approach and the RE estimator of relatedness (Section 2.2; Lynch and Ritland, 1999). Each set of conditions was replicated 500 times.

Finally simulations were used to investigate the triplet-wise likelihood approach. Locus number was simulated as 5, 10 or 20, with 5, 10, or 20 equally frequent alleles at each. Heritabilities were set to 0.5, and 15 families were simulated with size 10. Sample size was kept constant in the simulations due to the slow nature of triplet-wise analysis, computation time being dependent upon the cube of

population size. For example a population of size 150, with 10 alleles at 10 loci, takes fifteen minutes to converge, and a population of size 400 takes five hours even when using a more efficient algorithm for calculating likelihoods than Equation 3.5. Each set of conditions was replicated 100 times. Heritabilities were estimated using TSUM, TDIFF and TSUM and compared to estimates made using DIFF, since DIFF required that the same number of population parameters be estimated prior to the analysis.

As in Chapter 2, estimates of heritability were compared to REML derived estimates of heritability calculated using the actual relationships.

3.4 Results.

3.4.1 Sample size.

Bias decreases as sample size increases (not shown). Sampling variance and MSE decrease with increasing sample size (Figure 3.1). The MSE for the REML-derived estimates falls in proportion to the inverse of the sample size, as do the MSEs of the likelihood-based estimators. The MSE of the regression-based estimator, however, falls at a slower rate, reflecting a less efficient use of the data, and confirming the observation of the previous chapter. Increasing the sample size while maintaining the same family size results in a linear increase in the number of pairs that are full-sibs, but a quadratic increase in the number that are unrelated, so the variance of relatedness in the population decreases. As population structure is assumed known prior to estimation using the maximum likelihood techniques, this fall in the true variance of relatedness is not so important. However, the compensatory effects of including the prior information on population structure is not complete, and bias and MSE were still affected by the variance of relatedness. Mean squared error is dominated by the sampling variance, indicating that the major source of error is through sampling error rather than bias.

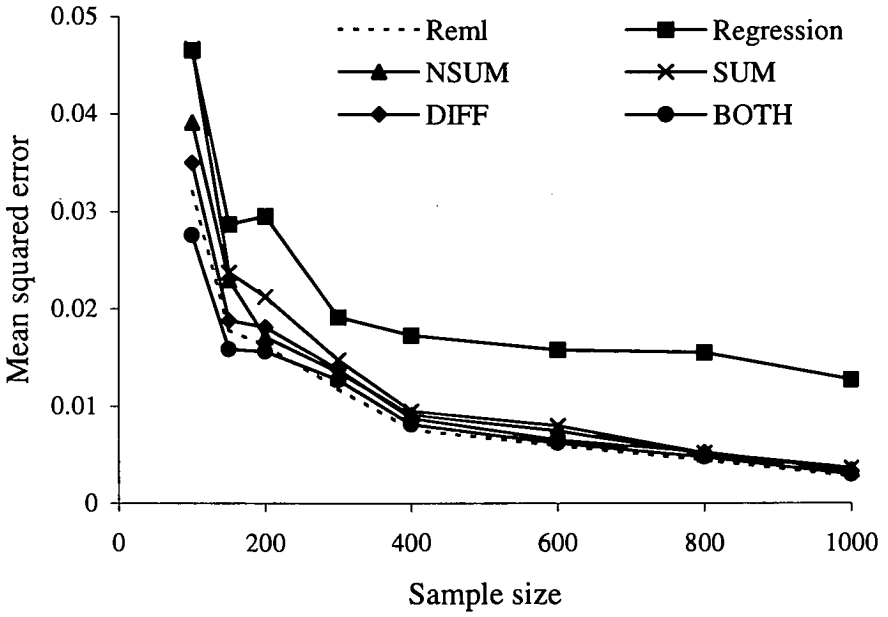

Figure 3.1: The change in mean squared error of the heritability estimates obtained using the likelihood approach and different phenotypic functions, when sample size was varied. Simulation conditions: Each family was of size 10, simulated heritability 0.5 and 10 loci each had 10 equally frequent alleles.

3.4.2 Marker information.

In general, the estimates approach the REML derived estimates as marker information increases. This trend is observable for both an increase in the number of marker loci (Figure 3.2a) and for an increase in alleles per locus (not shown). The regression-based estimator yields estimates that are very close to the REML derived estimates across the range of locus numbers. As marker information increases, the likelihood-based estimates approach the deviations predicted by theory (Equations 3.2 and 3.3). BOTH, which could not be analysed analytically, even in the balanced case, is biased downwards across the range of simulated marker numbers.

The relative importance of the phenotypic information in the likelihood techniques has a larger effect at lower numbers of marker loci (5 to 15). With few marker loci the posterior probabilities of the relationships becomes more dependent on the phenotypic information, whereas with larger numbers of loci the dependency is not so strong. With lower marker numbers (5 - 15 loci) the phenotypic information

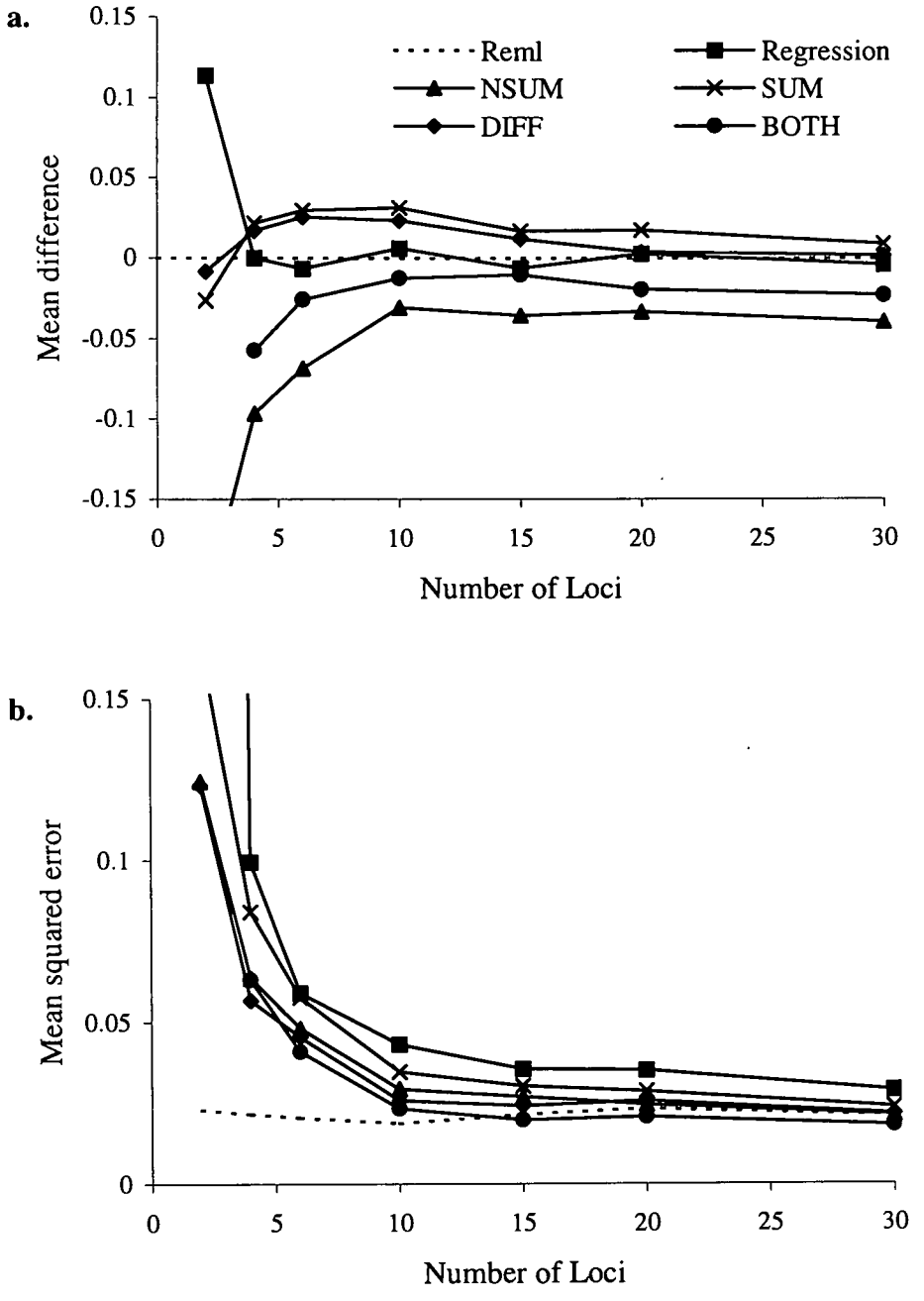

Figure 3.2: Results for the full-sib family simulations investigating the properties of heritability estimates obtained using the likelihood approach and different phenotypic functions, when the number of marker loci was varied. Simulation conditions: 20 families of size 10, heritability 0.5 and each locus had 5 equally frequent alleles. (a) Mean difference of the heritability estimates from REML estimates. (b) Mean squared error of the estimates.

causes an upward bias in the heritability estimates, because a phenotypically similar pair have a higher probability of being classed as full-sibs, including pairs which are not actually full-sibs. This trend is most notable in the DIFF line for the set of simulation conditions of Figure 3.2a, although it is seen with other phenotype functions under different simulation conditions.

With a very small numbers of markers (<5) and small numbers of alleles per locus (<10) the estimates of the relationships are extremely noisy, resulting in a decrease in the proportion of the variance assigned to additive genetic effects. Hence there is a downturn in the likelihood estimates at very low marker information (Figure 3.2a). Similar graphs are obtained when the plots of mean deviation against marker number are made under different family structures, heritabilities and allele numbers.

As marker information increases, the sampling variance of the heritability estimates decreases, approaching the sampling variance for REML estimates (Figure 3.2b). As might be expected, the regression approach shows the largest sampling variances since the additional information about population structure (the prior) is only used in the likelihood estimators.

The sampling variance of SUM is larger than that of DIFF. DIFF approaches the sampling variance of the REML-derived estimates because both yield unbiased estimates of the same two population parameters (σ_A^2 and σ_E^2) in the balanced case. NSUM estimates a single parameter only, the heritability, and so yields smaller sampling variance for that parameter. This causes the MSE of the NSUM estimates to fall below the MSE of the REML-derived estimates at higher levels of marker information when relationships are estimated with high accuracy. With increased numbers of alleles per locus relationships are also more accurately estimated, resulting in increased accuracy of variance component estimation (results not shown). Since REML gives unbiased estimates of variance components in balanced situations, average difference is very close to average bias. The square of the average bias is small compared to the sampling variance, so the mean squared error is dominated by the sampling variance. Thus deviations in estimates from the parameter value are more likely to be through sampling rather than bias in the technique.

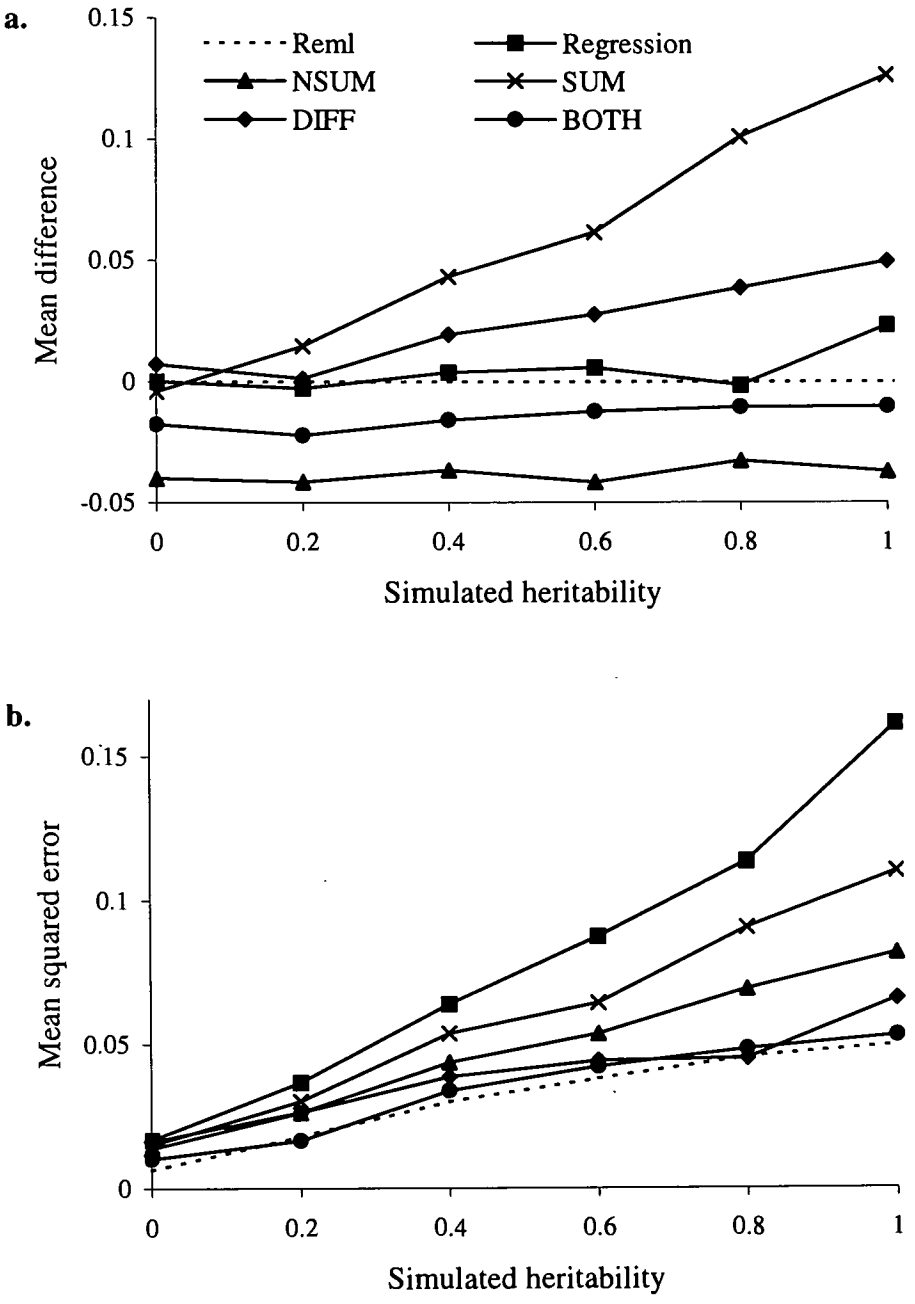

Figure 3.3: Results for the full-sib family simulations investigating the properties of heritability estimates obtained using the likelihood approach and different phenotypic functions, when simulated heritability was varied. Simulation conditions: 20 families of size 10 and 10 loci each had 10 equally frequent alleles. (a) The change in mean difference of heritability estimates from the REML estimates. (b) The change in mean squared error of the heritability estimates.

3.4.3 Simulated heritability.

NSUM and BOTH yield smaller estimates of heritability than the REML derived estimates (Figure 3.3a), BOTH being less biased than NSUM. The magnitude of the mean differences of NSUM and BOTH from the REML estimates decreases as simulated heritability increases. DIFF estimates and the regression-based estimates of heritability do not differ on average from the REML-derived estimates. SUM is unbiased at zero heritability but becomes increasingly biased upwards with increasing simulated heritability. Comparison of the observed biases in NSUM and SUM, with the biases predicted using equations 3.4 and 3.3 respectively shows that bias is more positive than is predicted for known relationships. This is confirmed by DIFF, which shows increasing bias with increasing simulated heritability, despite being unbiased when exact relationships are used. This extra bias is due to the inclusion of phenotypic information in the likelihood calculation. The relative weight placed on the phenotypic information increases as the simulated heritability increases because the difference between the phenotypic distributions for unrelated and full-pairs becomes larger. As explained previously, a pair that is phenotypically similar has a higher chance of being classed as full-sib and so inclusion of phenotypic information biases subsequent heritability estimates upwards.

The MSE of the estimates of heritability increases with simulated heritability for all the techniques (Figure 3.3b), with the exception of DIFF, where MSE falls relative to the MSE of the REML-derived estimates as simulated heritability increases. Because bias is small relative to the sampling variance, the mean squared error is again dominated by the sampling variance.

3.4.4 Population structure.

Different population structures, such as alterations in the sample size, give rise to different variances in the relationships. Mean difference is little affected by change in the simulated variance of relationship (Figure 3.4a), except in situations where the variance of relationship is low, e.g. with 75 families of size two.

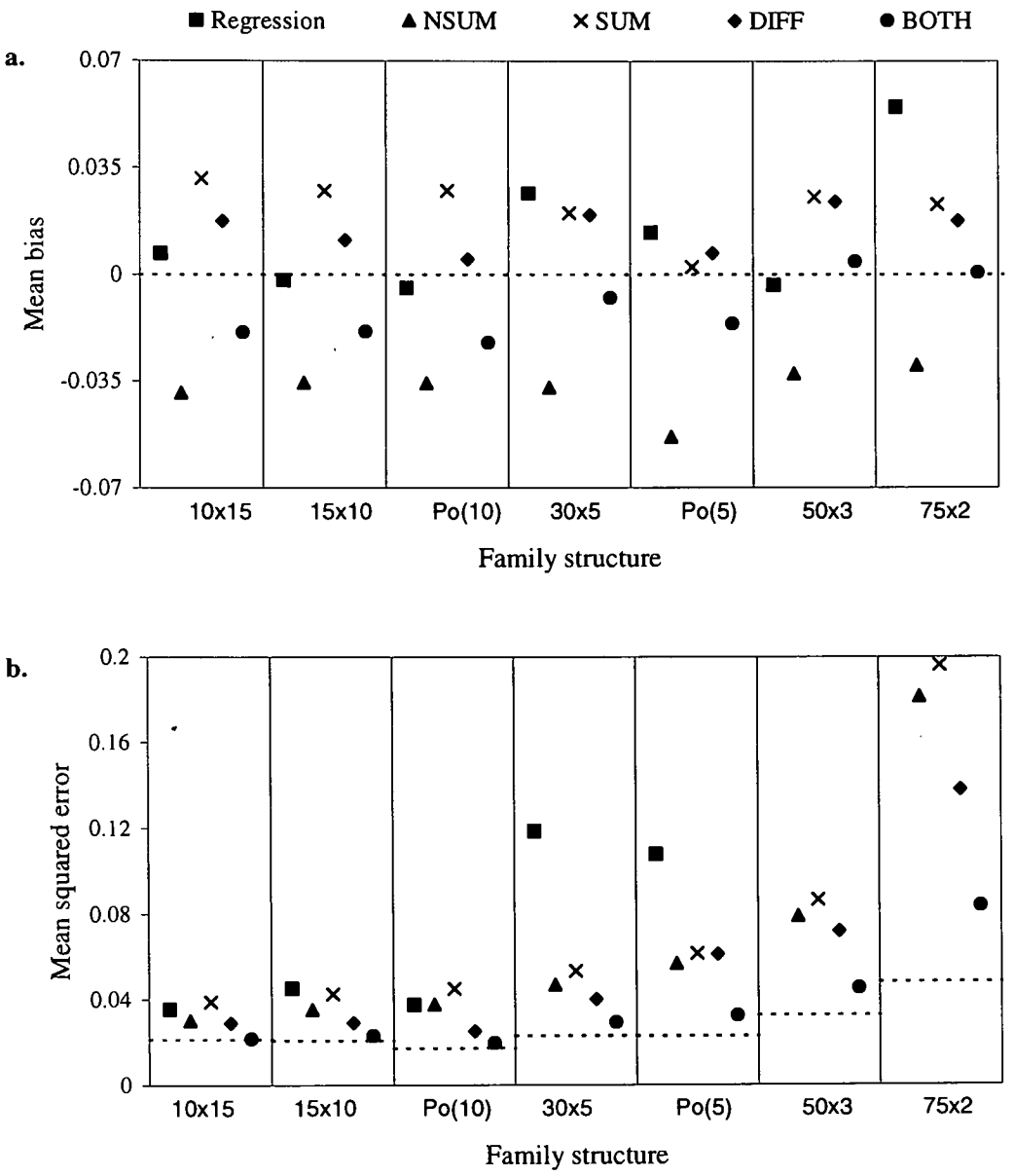

Figure 3.4: Results for the full-sib family simulations investigating the properties of heritability estimates obtained when population structure was varied. Simulation conditions: Total sample size was 150, heritability 0.5 and 10 loci each had 10 equally frequent alleles. (a) The change in mean difference of the heritability estimates from the REML estimates. (b) The change in mean squared error of the heritability estimates. (10x15 = 10 families of size 15; Po(10) = families followed Poisson distribution with mean 10). MSE was 0.28 and 1.3 for the regression-based technique for structures 50x3 and 75x2 respectively.

The higher the variance of relationship in the simulated population the lower the MSE and the smaller the bias of the estimate (Fig 3.4b). With the family structure 75 by two, giving the smallest variance of relationship, the regression-based estimator performed considerably worse than the likelihood-based estimators, having a sampling variance of 0.4. Overall, estimates in populations with random family size showed slightly larger mean bias and MSE than those of populations with balanced family of the same mean size (Figs 3.4a,b), reflecting the poorer ability of the pair-wise techniques to weight the data from each family.

3.4.5 Triplet-wise analysis.

At lower levels of marker information triplet-wise analysis using TDIFF yields estimates of heritability that have large downwards bias, and large MSE compared to estimates determined using the computationally simpler pair-wise approach (Table 3.3). When the amount of marker information is increased the bias of TDIFF decreases and approaches the bias seen in the REML-derived estimates. The MSE also approaches the MSE of the REML-derived estimates (Table 3.3). When exact relationships are used in place of inferred relationship information and a balanced population is simulated, TDIFF gives unbiased estimates of the heritability (not shown). This is presumably because no prior estimates of the phenotype distribution parameters are required in the calculation. TBOTH follows the same pattern as TDIFF, although the magnitude of the bias is smaller (Table 3.3). In addition, since an estimate of the sample mean is used in the calculation, estimates of heritability made when exact relationships are used are biased downwards. TSUM shows upwards bias, which decreases as marker information increases, and has MSE that approaches that of the REML-derived analysis (Table 3.3). When the known relationships are used with TSUM, estimates of heritability are biased upwards, again due to an estimate of the sample mean being included in the calculation. In contrast, the pair-wise analysis shows the smallest mean differences and the smallest MSE out of the marker-based approaches at low levels of marker information. At higher levels

Marker		REML	DIFF		TDIFF		TSUM		TBOTH	
No. Loci	Allele	MSE	ΔD	MSE	ΔD	MSE	ΔD	MSE	ΔD	MSE
5	5	0.032	0.041	0.052	-0.452	0.317	0.083	0.118	-0.254	0.231
5	10	0.039	0.013	0.053	-0.430	0.203	0.102	0.064	-0.151	0.053
5	20	0.044	0.005	0.042	-0.146	0.074	0.057	0.033	-0.056	0.041
10	5	0.034	0.016	0.040	-0.421	0.245	0.104	0.085	-0.160	0.038
10	10	0.039	0.010	0.041	-0.229	0.082	0.051	0.058	-0.065	0.039
10	20	0.031	0.003	0.031	-0.014	0.031	0.038	0.031	-0.036	0.031
20	5	0.039	0.023	0.042	-0.327	0.220	0.063	0.061	-0.085	0.031
20	10	0.035	0.006	0.037	-0.075	0.043	0.039	0.048	-0.042	0.036
20	20	0.036	-0.001	0.037	-0.001	0.035	0.035	0.031	-0.036	0.036

Table 3.3: Summary of the mean differences (ΔD) from the REML estimates and the mean squared errors of the heritabilities estimated using the triplet-wise likelihood approach and varying amounts of marker information. Simulation conditions: 20 full-sib families of size 10, heritability 0.5 and equally frequent alleles.

of marker information (e.g. 20 loci, with 10 alleles at each) the mean differences and MSE become more comparable.

Contrary to the prediction that it would improve estimation through the inclusion of higher order groups of relations, triplet-wise analysis gives estimates that are less reliable. It is unclear why this is the case. It might be because in triplet-analysis the marker data are used to split the phenotypic data into five relationship classes while in pair-wise analysis the same amount of marker data are being used to split the phenotypic data into two relationship classes. Alternatively, it may be due to increased weight being placed on the phenotypic function since three sets of phenotypic data are used instead of two. Overall, given the comparative length of time it takes to run a triplet analysis rather than a pair-wise analysis and due to the observed bias and large MSE of the results with low amounts of marker information, triplet-based analysis is not recommended.

3.4 Discussion.

In all cases the average deviation of the pair-wise marker based estimators from the REML estimate is very close to the average bias, because the REML estimates are unbiased (with balanced families) or nearly so. In most cases the sampling variance was much larger than the square of the mean bias, indicating that errors in estimation are likely to be mainly through sampling rather than through bias.

The regression-based method shows least average bias in its estimates over the range of the tests; however, this must be set against the higher variance of estimates. DIFF, the likelihood-based technique that uses the phenotypic differences within pairs to calculate variance components, showed least bias out of the likelihood-based procedures, presumably because fewer assumptions about population parameters are required prior to calculation. At higher levels of marker information DIFF showed less bias than the regression-based technique (Figure 3.1a). Simulations run where both the sum and the difference of phenotype were combined into one estimator indicate that there is some further information to be gained over the difference or over the sum, alone. This is because the sum and the difference are uncorrelated for both full-sib and unrelated pairs. Results indicated

that BOTH estimates were more biased than DIFF (because more population parameters require estimation prior to analyses), but yielded lower sample variances at larger sample sizes.

Triplet-wise analysis was biased at low levels of marker information, and estimates were very slow to converge. It is unclear why estimates are so biased with low levels of marker information, while at higher levels of marker information are comparable to the pair-wise estimates. Possible explanations are that the errors associated with assigning triplets into the five relationship categories are larger than the errors associated with assigning pairs into two relationship classes given the same level of marker information. Inaccuracies in the relationship information would lead to downwards bias. This is because estimates of the additive genetic variance are derived from the differences between the phenotypic distributions of related and unrelated individuals (Table 3.3). Regarding either related individuals as unrelated or unrelated individuals as related decreases the differences between the distributions and therefore leads to an underestimate of the additive genetic variance. Overall, there was no observed advantage of using triplet analysis over pair-wise analysis at higher levels of marker information. Due to the inefficient use of computer time, the observed bias, and the comparative complexity of the triplet-wise analysis compared with pair-wise analysis, triplet based analysis is not recommended and will not be considered further.

All the methods share a number of basic properties. For accurate results they require that adequate amounts of marker information be used in the estimation of relationships. They also require that a sufficiently large variance of relationship is present in the population under investigation. With extremely low variance in relationship (0.014 in the example of 75 full-sib families of size two), the regression-based estimator was considerably worse than the likelihood-based estimators. This is because there are fewer full-sib pairs in the population and therefore less information available to make inferences about the distribution of the full-sib phenotypic information. The likelihood procedures also performed worse under these conditions for similar reasons; however, because these techniques require the use of prior information on the population structure, this lack of information has less effect on the estimates. In this study, small population sizes were simulated in order to reflect the

small sample sizes often available from natural populations. The properties of all the estimators are improved with respect to the level of bias and the sampling properties of the estimates when larger sample sizes are used. Mean squared errors fell in proportion to the inverse of sample size for the likelihood-based estimators and at a slower rate for the regression-based estimator, indicating the importance of the variance of the relationship and that the regression approach is a less efficient method. This inefficiency, as mentioned in Chapter 2, may be due to using locus-specific weights appropriate for unrelated pairs in the determination of relatedness for all pairs regardless of the true relationship. Alternatively it might reflect inefficiency in the weighted ANOVA procedure for estimating the actual variance of relationship.

The likelihood techniques that require prior estimates of population parameters other than the probability that a randomly selected pair are full-sibs show bias caused by the inclusion of these sample estimates. This bias is removed if the actual parameter values are included in place of estimates from the sample. Additionally, with all of the pair-wise likelihood functions there is upwards bias arising from the inclusion of phenotypic information in the likelihood calculation because a pair containing individuals that are phenotypically similar has a higher probability of being classed as full-sibs. If the phenotypic information is separated from the marker-based information the problem reduces to a regression, with phenotypic difference (for example) plotted against the relationship. Each pair-wise phenotypic difference is regarded twice, once for the unrelated category and once for full-sib category, but weighted by the marker and prior-based likelihood that it falls into each category. The slope of the regression line and the y-intercept can then be used to estimate the desired variance components, with the slope being dependant on the additive genetic variance and the y-intercept on the additive and environmental variances. However, this approach is more biased than the likelihood approach, since all pairs have a finite probability of being in each category. Inclusion of either full-sibs as unrelated individuals or unrelated individuals as full-sibs causes the slope of the regression line to decrease and hence introduces downwards bias to estimates of the additive genetic variance.

The two types of method are designed for use under slightly different circumstances. The regression-based technique, which requires assumptions about the population mean and variance, may be used when little information is available on family sizes or relationship structure. The likelihood-based techniques can be used only when such information is available. The cost of the increased generality of the regression-based procedure is a large increase in the mean squared error of estimates.

It is evident from these results that these techniques may only be used in natural populations with sufficient marker information and suitable population structure. For example in a natural population comprised of small families, a much larger sample of individuals than 150 from the population would be required, with a larger amount of marker data, before variance components can be estimated with useful accuracy.

Natural populations contain more than two classes of relatives and techniques must be applicable to such heterogeneous populations. The regression-based estimator uses a method of moments to estimate a measure of relatedness and therefore does not require extension to deal with combinations of relationships, provided there is a large number of relationships within the sample. The likelihood-based procedure requires the calculation of the likelihood that a pair falls into each of the relationship classes under consideration, given the marker and the phenotypic information. As a result, likelihood-based approaches are readily extendable to include other categories of relationship. However, for both types of technique the ability to distinguish between more distant relationships classes using marker information falls rapidly with the increase in the distance of the relationship (Thompson, 1975; Blouin *et al.*, 1996). This would result in poorer estimates of variance components in populations with low variance in relatedness. Chapter 5 includes investigation of half-sib samples, and mixtures of half-sibs and full-sibs.

An additional consideration is how known relationships may be incorporated into each model, e.g. mother-offspring pairs might be known. In the regression-based technique, known and unknown relationships may be used together in the estimator provided that some means of scaling the estimated relationships is adopted so that they are in line with known relationships. This may be accomplished by equating

relationship estimates of known pairs against the known relationship. However known relationships also alter the likelihood of other relationships within the sample. Estimates of relatedness should therefore be made between groups of known relationship, rather than between single individuals within those groups. As with the inclusion of higher-order relationships, comparison between groups requires that coefficients describing the higher-order moments of the relationship and the additive genetic effects be estimated, thereby making calculation complex and the interpretation of results difficult. This problem is discussed further in Chapter 5.

Accommodating known relationships into the likelihood techniques is achieved more simply, by setting the likelihood for the known relationship class for a pair to one and the likelihoods for the other relationship classes to zero. A further benefit of the likelihood techniques is that it is simple to update the likelihood of a particular relationship by knowledge of another relationship. For example, if mother-offspring pairs are known, the origin of one of the alleles at each locus within an individual is accounted for and so estimates of relationships through the father may be based upon the remaining allele. In practice, errors in genotyping must be accounted for if this technique is to be adopted in a practice (Marshall *et al.*, 1998).

Chapter 4
A Markov chain Monte Carlo approach

4.1 Introduction

Molecular-based tools for inferring genetic relationships may be grouped into two categories: method-of-moments estimators which are used to estimate relatedness, as a continuous measure, based on shared alleles at marker loci (Lynch, 1988; Queller and Goodnight, 1989; Ritland, 1996a; Lynch and Ritland, 1999); and likelihood techniques used to determine the likelihood of a pair falling into particular relationship classes, e.g. full-sibs or non-sibs, given the observed marker information (Thompson, 1975; Mousseau *et al.*, 1998).

Similarly, two methods that allow the estimation of quantitative genetic parameters associated with a trait without reference to the exact pedigree have been described (Chapters 2 and 3; Ritland, 1996b; Lynch and Walsh, 1998; Mousseau *et al.*, 1998). These use molecular data to infer pair-wise relationships between individuals, since this is the least complex level at which relationships may be estimated. Ritland (1996b) proposed a regression approach to parameter estimation, where measures of pair-wise phenotypic similarity are regressed against pair-wise relatedness (Ritland, 1996b; Lynch and Walsh, 1998). Alternatively, if prior information is available on population structure, likelihood-based procedures may be adopted, in which pairs are placed into a predetermined population structure according to the probability of observing their genotype and phenotype (Chapter 3; Mousseau *et al.*, 1998).

Pair-wise techniques lose valuable information in the form of higher order relationships. For example, if three individuals sampled from a single generation

have genotypes a_ia_i, a_ja_j and a_ka_k (a_i, a_j and a_k are mutually exclusive alleles) they can not be full sibs; but with pair-wise analysis such exclusion is not possible. Additionally, with pair-wise techniques the weight placed on information from a single family depends on the number of pairs of individuals that can be chosen from that family. It is therefore dependent only upon family size, and not information content. Consequently, pair-wise methods do not yield the most efficient estimates for parameters, and are prone to larger standard errors than efficient methods of estimation such as restricted maximum likelihood. Only in the case of balanced populations containing two classes of relationship are families weighted equally, and then give estimates identical to ANOVA derived estimates when exact pedigree information is known (Chapter 3).

A secondary problem is obtaining estimates of the allele frequencies at the marker loci. In previous studies allele frequencies have been assumed known, or have been estimated from the sample. If allele frequencies are estimated from the sample under investigation they are subject to further random error, since there are relatives within the sample, which might bias subsequent estimates of pair-wise relationships. To combat this problem, Queller and Goodnight (1989) proposed recalculating the allele frequencies for each pair under investigation excluding the information from that pair. This removes a small covariance between population and individual allele frequencies and results in slightly improved estimates, although change is negligible with large numbers. Ritland (1996b) adopted the same approach.

A final problem with pair-wise methods is how they may be extended to include other factors such as sex or year in the model. Since they operate on a pair-wise level other factors must also be investigated on a pair-wise level and as a result the optimum estimate may not be achieved.

In this chapter a simple two step procedure for estimating variance components is described. Firstly families of sibs are reconstructed using a Markov Chain Monte Carlo (MCMC) procedure, and secondly the reconstructed sib-ships are used to estimate variance components. The MCMC procedure is based upon likelihood techniques and, to clarify the nomenclature used, will be referred to as the MCMC approach, while the pair-wise procedure of Chapter 3 will be referred to as the likelihood approach.

The MCMC procedure reconstructs sib-ships within a single generation allowing improved parameter estimation through more efficient weighting of families and use of more than pair-wise pedigree information. Conceptually the sib-ship reconstruction procedure shares features with Bayesian approaches using MCMC procedures in phylogeny reconstruction (Kuhner *et al.*, 1995; Larget and Simon, 1999; Yang and Rannala, 1997) where, given the sequence data, the most plausible phylogenetic trees are generated from a large number of potential trees without the need to investigate every possible tree. Similarly, in sib-ship reconstruction plausible sib-ships are generated from the sample using the marker data without the need to investigate every possible combination of sib-ships. However, in sib-ship reconstruction the aim is to reconstruct a number of groups with specific relationships rather than determine likely distances between each member (or taxon) in the sample. This approach of reconstructing specific groups is equivalent to fixing the possible branch lengths to either one length (representing full-sib) or to double that length (representing unrelated) in phylogeny reconstruction. Moreover, no attempt is made to update the assumed prior distributions of the parameters used in pedigree reconstruction, since these are not the parameters of interest. In this light the techniques used in pedigree reconstruction are not Bayesian in nature.

Reconstructed pedigrees are subsequently used to form a relationship matrix suitable for use in an animal model run with restricted maximum likelihood (REML) (Lynch and Walsh, 1998; Patterson and Thompson, 1971; Searle *et al.*, 1992), specifically using the ASREML program (Gilmour *et al.*, 1997). This approach allows traditional efficient methods for parameter estimation to be used and hence simplifies the inclusion of additional factors or the use of multivariate analysis if data have been collected from several traits. In addition, methods are outlined that allow the estimation of population allele frequencies that account in part for relationships within the sample. Simulation of full-sib families is used to investigate the properties of this approach, and to compare estimates of heritability with estimates made using the other marker-based approaches.

4.2 Statistical Methods.

4.2.1 Inferring sib-ships.

4.2.1.1 Markov Chains.

Markov Chain Monte Carlo simulations facilitate the determination of solutions to problems that can not readily be solved by theoretical calculations (Norris, 1997). A Markov chain is a random walk through the parameter space of a system, where each step of the walk depends only upon the current state of the chain. If the likelihood for the set of parameters at the current point of the chain is calculated and compared against the likelihood at the next point then the random walk may be 'guided' to points of high likelihood within the parameter space. This provides a way to estimate parameter values with a high likelihood (though not necessarily the highest) without having to search the entire parameter space. These techniques are therefore of particular use in solving complex likelihood problems, especially when the parameter space is large. A necessary feature of these procedures, however, is the requirement that the chain be able to make moves that decrease the likelihood of the parameters (Gilks *et al.*, 1996). This feature decreases the chance that the chain becomes stranded upon a false likelihood peak. Thus, when a decrease in the likelihood of the parameters is encountered the chain only moves to the new state with a probability dependent upon the size of the decrease in the likelihood.

In this study a modified from of the *Metropolis-Hastings* algorithm (Metropolis *et al.*, 1953; Gilks *et al.*, 1996) is used to reconstruct the sibships. With Metropolis-Hastings algorithm the probability of making a step that increases the likelihood is always one (i.e. the step is always made) but the probability of making a step that decreases the likelihood is dependent upon the magnitude of the change in the likelihood. Thus the chance of the chain becoming 'stranded' on a false likelihood peak is decreased. Here, the probability of moving between two points of the parameter space depends on the change in the likelihood between the two points, regardless of whether the likelihood increases or decreases. If there is a large

increase in the likelihood the chain will almost certainly move, if there is a large decrease the chain will almost certainly not move. At intermediate values the chain may move more freely through the parameter space (with steps that increase and decrease the likelihood being possible), although the probability of making a step that increases the likelihood is always greater than the probability of making a step that decreases the likelihood. In this way the chances of the chain becoming stuck on a false likelihood peak are decreased.

In this study, molecular data are first used to reconstruct sib-ships assuming individuals are either sibs or unrelated using an MCMC approach and then the reconstructed sib-ships are used to estimate variance components for a quantitative trait. Errors in pedigree reconstruction are of two types: denoted Type I, where genuinely unrelated individuals are classed as related, and Type II, where genuinely related pairs are classed as unrelated. It is shown that type I errors lead to large downwards bias in parameter estimation, while type II errors lead only to trivial downwards bias. It is not necessary to find the point with highest likelihood, but merely a point of high likelihood since, firstly, sib-ship reconstruction leads to few errors of type I, and secondly the true sib-ship may not have the highest likelihood given the marker information.

4.2.1.2 The population.

Suppose a sample of N individuals has been taken from a single generation of a population. Each individual has been scored for genotype at ℓ physically unlinked marker loci, and there is some information on which to base assumptions about relationship structure. In the case described here it is assumed that the sample contains only full sibs and unrelated individuals and that the distribution of full sib family sizes is known. Other relevant information might be about known relationships, such as between offspring and dam in a half-sib structure. The likelihood of the relationship structure, allele frequencies and genotypes of the individual animals may be calculated from the sample. This is a function of the observed marker information and any previous knowledge of the allele frequencies and relationship structure. The likelihood may be expressed as:

$$L_{\text{population}} = L(a, g, s|m, d), \tag{4.1}$$

where a represents the marker allele frequencies within the population, g denotes the N genotypes within the sample, s denotes the sample space of possible family structures (i.e. the sib family membership), m denotes the observed marker information and d represents the previous knowledge, such as the distribution of family size, about relationship structures.

Maximising (4.1) over all possible family structures is prohibitive, since an extremely large number is possible, even in small samples. For example with just ten individuals, restricted to being either full sib or unrelated, there are 115,975 possible family structures. Markov Chains or other optimisation techniques are therefore required. The number of possible family structures in a full-sib design may be calculated from Stirling numbers of the second kind (see Abramowitz and Stegun, 1965). The numbers are of the form S_N^F, and total the number of ways of partitioning N individuals into F families. For example for 5 individuals the total number of possible family structures would be equal to $\sum_{f=1}^{5} S_5^f$.

4.2.1.3 With known allele frequencies.

The likelihood of individual families: If allele frequencies are assumed to be known then individual family likelihoods become independent and equation 1 may be expressed as:

$$L_{\text{population}} = \prod_f L(g_f, s_f|m_f, d, a), \tag{4.2}$$

where f indexes family.

In the model the likelihood of any single family f of size n_f is equal to the probability of observing the genotypes given that all the members of f are full sibs,

multiplied by the likelihood of observing the structure (here the size) of family f given the prior information:

$$L_{\text{family}} = L_{\text{genotypes}} \cdot L_{\text{structure}} \tag{4.3}$$

The probability of the observed genotypes within a putative full sib family is calculated as:

$$L_{\text{genotypes}} = \prod_{\ell} \left[\sum_{w=1}^{n_\ell} \sum_{x=1}^{n_\ell} \sum_{y=1}^{n_\ell} \sum_{z=1}^{n_\ell} \left[p^1_{wx} p^2_{yz} \prod_{c=1}^{n_f} L(g_{c\ell}) \right] \right], \tag{4.4}$$

where ℓ denotes independent marker loci; n_ℓ the number of alleles at locus ℓ; w, x, y and z index alleles; c indexes an individual from the putative family, p^1_{wx} is the ordered genotype frequency of parent one, p^2_{yz} is the ordered genotype frequency of parent two and $L(g_{c\ell})$ is the probability of observing the genotype of individual c at locus ℓ given the parental genotypes. For example if p^1 and p^2 share the genotype (1, 2) then $L(g_{c\ell}) = ¼$ when the offspring genotype, g, is (1, 1) or (2, 2), $L(g_{c\ell}) = ½$ when g is (1, 2), and $L(g_{c\ell}) = 0$ otherwise.

In practical computing it is much more efficient, reducing running time to its square root, to take the first offspring and assign one of its alleles to one parent and the other allele to the other parent, and then sum over the remaining alleles, (see Appendix 2).

In the simulations the likelihood of the family structure depends only upon the family size (since category information is included in the way the Markov Chain mixes the population). Either a non-informative distribution for full-sib family size, where each family size is equally likely, or a truncated Poisson (Po) distribution (no zero class) describing the probability of each family size was used. The independence of families allows fast Monte Carlo algorithms to be written, since at

each step in the chain only the likelihoods of individual families rather than the likelihood of the whole population need to be considered.

The 'hill climbing' algorithm for full sib family reconstruction:

a) Start with each member of the sample assigned to a different family. This starting point avoids the problem of generating populations with likelihoods of zero, which is almost certain with randomly selected families.

b) Calculate the likelihood for each family and store.

c) Select a random individual, x, from a randomly chosen family f_1. This individual is to be moved at random to a new location (new family) within the sample.

d) Select a random destination family, f_2, for individual x. The new location is chosen in a way that allows the individual to stay in the same place, or to be placed in a new family on its own. In the simulated study described below all the possible destinations had an equal probability of being chosen. Other schemes are possible; e.g. the probability of choosing a new 'blank' family to move x to could be assigned a greater probability of any individual family or the probability a family is chosen may be made dependant upon its size.

e) Calculate $L_{old} = L(f_1) \times L(f_2)$. Use the stored likelihoods to calculate the likelihood of observing families f_1 and f_2 prior to moving x. This equals the product of the likelihoods of each family on its own, since families are independent.

f) Move x from f_1 to f_2.

g) Calculate new likelihoods for f_1 and f_2 after the move of x; these are termed $L(f_1)_{new}$ and $L(f_2)_{new}$.

h) Calculate the new likelihood of observing both families, $L_{new} = L(f_1)_{new} \times L(f_2)_{new}$.

i) Calculate $r = L_{new} / (L_{new} + L_{old})$.

j) Draw z from a uniform distribution between 0 and 1.

k) Compare z with r. If $z < r$ move x back to f_1. If $z \geq r$ store $L(f_1)_{new}$ and $L(f_2)_{new}$. This step means that the probability of accepting a change, i.e. keeping x in f_2 depends on the change of the likelihood. If $L_{new} >> L_{old}$ the change is almost certainly accepted, but if $L_{old} << L_{new}$ then the change is almost certainly rejected. In

addition this allows backsteps, a decrease in the likelihood, to occur, thereby reducing the chance that the chain will become stranded on a false maximum. As explained above, these rules for accepting and rejecting a change are a modified form of the rules governing a Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Gilks *et al.*, 1996), although here, the probability of making a step that increases the likelihood is also dependent upon the magnitude of the change in the likelihood.

l). Return to (c). Continue to move (mix) individuals between families, until stopping criteria are reached; these are discussed below.

Stopping the chain: A number of criteria may be used to stop the chain:

1). A fixed number of iterates has been run. This method must be repeated a number of times for the sample, and the resulting full sib families compared for similarity. The population with the greatest likelihood may then be selected, or some composite structure determined (although this requires additional checking for exclusions).

2). The likelihood for the whole population (the product of the stored likelihoods) remains constant or nearly so for a fixed number of iterates.

3). The average family size approaches the expected family size, and then remains constant or nearly so for a fixed number of iterates.

In practice there is little difference between using criteria 2 and 3 to stop the chain. In populations of size 200, with five alleles at each of ten loci and a family size distribution that is Po(5), the population likelihood and mean family size level out together, with the values stabilised by 300,000 cycles (often by 220,000). With the same level of marker information a population of size 800 stabilises after about 900,000 cycles.

Half-sib reconstruction: The algorithm is easily modified to accommodate the reconstruction of half-sib families. For half-sib families the probability of observing the genotypes of a putative half-sib family, over all the possible genotypes of the shared parent, is computed for each locus and then multiplied across loci. The likelihood of each offspring depends on the likelihood of receiving one allele from the common parent and the other from an allele pool with the same allele frequencies as the population. Parental genotype information may be incorporated into both half-

and full-sib algorithms by constraining the parental genotypes over which the offspring genotype likelihoods must be summed. The likelihood equation for a half-sib family is included in Chapter 5.

4.2.1.4 With unknown allele frequencies.

Calculating parental allele frequencies from samples containing relatives: Population allele frequencies are usually unknown and must also be calculated from the sample. In sib-ship reconstruction the likelihood of observing a particular sib-ship depends on the allele frequency in the parental generation, since these are the alleles that are sampled to form the offspring generation. Allele frequencies may be estimated by using a weighted least squares approach (Dillon and Goldstein, 1984), with correlations of the allele counts between relatives accounted for by inclusion of the relationship matrix. The derived estimator is dependent only upon the relationship matrix and the allele counts:

$$\hat{a}_i = \left(\mathbf{1}^T \mathbf{R}^{-1} \mathbf{a}\right)\left(\mathbf{1}^T \mathbf{R}^{-1} \mathbf{1}\right)^{-1}, \tag{4.5}$$

where $\hat{a}_i$ is the mean allele count, **R** is the relationship matrix, **a** is the vector containing the allele counts for each individual and **1** is a vector of ones. Allele frequency is then estimated as $\hat{a}_i/2$.

An updated algorithm: The previous algorithm can be modified using the allele frequency estimator to simultaneously update allele frequencies. The process is begun by calculating the allele frequencies as though all members of the population are unrelated, and then periodically updating the estimates as groups of full sibs are generated (e.g. every 5000 iterates). Recalculation every step is unnecessary: firstly, there may be no change made in population structure, and secondly a single change does not affect allele frequency estimates significantly. Updating allele frequencies reduces the population frequency of alleles shared by grouped individuals and also reduces the probability that a group reconstructed as full sibs will be broken down

again, even if the reconstruction is wrong. It is therefore recommended that allele updating starts after a number of cycles has already been run (say, 100,000).

4.2.2 Measuring the accuracy of the reconstructed family.

A statistic that enables measurement of the accuracy of each reconstructed family is useful for the purposes of comparison. Simulating populations with known relationships using the same parameters (or estimates of the parameters) for the distribution of family size as those of the study population allows percentage confidence levels for a given size of family to be estimated. Two confidence levels may be determined: the probability that full-sib family members in the family reconstructed are genuine full sibs, and the probability that the family is complete (i.e. is not the result of a larger family being split – a possible problem with this approach to sib-ship reconstruction).

To assess the properties of the estimators in the simulated study, where the real family structure is known, an additional statistic that scores the reconstructed pedigree for accuracy was defined:

$$accuracy = \left(S_{\mathrm{fs|fs}} - S_{\mathrm{fs|ur}}\right) / Tot_{\mathrm{fs}} \qquad (4.6)$$

where $S_{\mathrm{fs|fs}}$ is the total number of correctly reconstructed full sib pairs, $S_{\mathrm{fs|ur}}$ is the total number of incorrectly reconstructed full sib pairs and Tot_{fs} is the total number of full sib pairs in the true pedigree. This statistic equals zero when all members of the population are in different families, and one when the population structure is reconstructed exactly. Since the statistic actively penalises accuracy when unrelated individuals are reconstructed as full sibs (type I errors), it may become negative in poorly reconstructed populations.

4.3 The simulated populations.

Simulation was used to compare the properties of heritability estimates made using the reconstructed pedigree approach with those of the pair-wise approaches. Phenotypic data for full sib data sets were generated as described in section 2.3.

The simulations were run under different conditions: Marker information was varied, with populations simulated with 2, 3, 5, 8 and 10 equally frequent alleles at each of 10 loci; full sib family sizes were drawn from a truncated (i.e. no null class) Poisson distribution with parameter 2, 5 and 10; and populations with 100, 200, 400 and 800 individuals in total were simulated. Each set of conditions was run 250 times on independently generated random populations. Heritability was set to 0.5.

To test the robustness of the algorithm to reconstruct families, populations were simulated from a Po(5) distribution of family size, but different assumptions were used about this distribution during reconstruction, namely uninformative (where every family size is equally likely), Po(5) and Po(10).

Simulations were run on the populations with allele frequencies updated after every 2000, 5000, 10000, 20000 cycles, or not at all. The accuracy of reconstruction statistic was also calculated and compared between each level of allele update.

In each set of simulations, MCMC iterations were continued for 1,400,000 cycles, a greater number than required for the levelling off of both mean family size and population likelihood.

Reconstructed sib-ships were used under an animal model to estimate the additive genetic and residual variances for the simulated trait, employing a standard package, ASREML (Gilmour *et al.*, 1997). Heritability estimates were taken as the summary statistic. Heritabilities were also estimated by the pair-wise approaches (Chapters 2 and 3; Ritland, 1996b; Lynch and Walsh, 1998; Mousseau *et al.*, 1998). There are a number of forms of the likelihood technique, and in this study the procedure based on the difference in phenotype was used (Chapter 3). Results were compared in terms of the mean deviation of heritability estimates from the "best" achievable estimates (those estimated by REML from the true pedigree and the same quantitative data), which reflects bias, and mean squared errors (MSE), a composite statistic of bias and sampling variance over simulations.

4.4 Results.

4.4.1 Sample size.

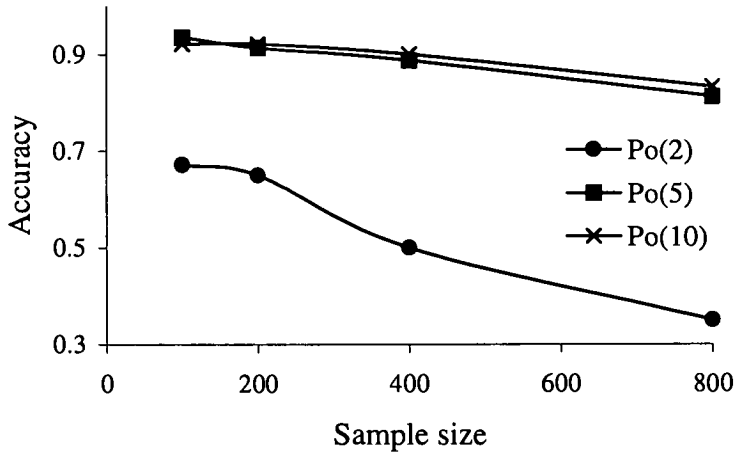

Figure 4.1: The change in accuracy of family reconstruction with changing sample size, for the three simulated distributions of family size [Po(2), Po(5) and Po(10)]. Simulation conditions: 200 individuals, 10 marker loci with five alleles each and heritability 0.5.

Figure 4.1 shows the change in the accuracy statistic as sample size increases for three distributions of family size. Accuracy decreases approximately linearly with an increase in the sample size. This is due to the increased chance that unrelated individuals have similar genotypes through random sampling, and may be compensated for by increasing the marker information. The accuracy for the Po(2) distribution of family size was much less than the accuracy of the Po(5) or Po(10) graph, reflecting much poorer reconstruction of pedigrees. This is discussed below.

Consider first the results for the Po(5) distribution. Figure 4.2a-ii shows the mean deviation of heritability estimates obtained using marker-based approaches from those using the known pedigrees (the zero line). Estimates using the reconstructed populations deviate less from the true pedigree estimates than pair-wise estimates, and show trivial negative bias. The size of the negative bias increases in a roughly linear manner as sample size increases (and hence also increases linearly with the accuracy statistic). This is probably due to the splitting of large families into

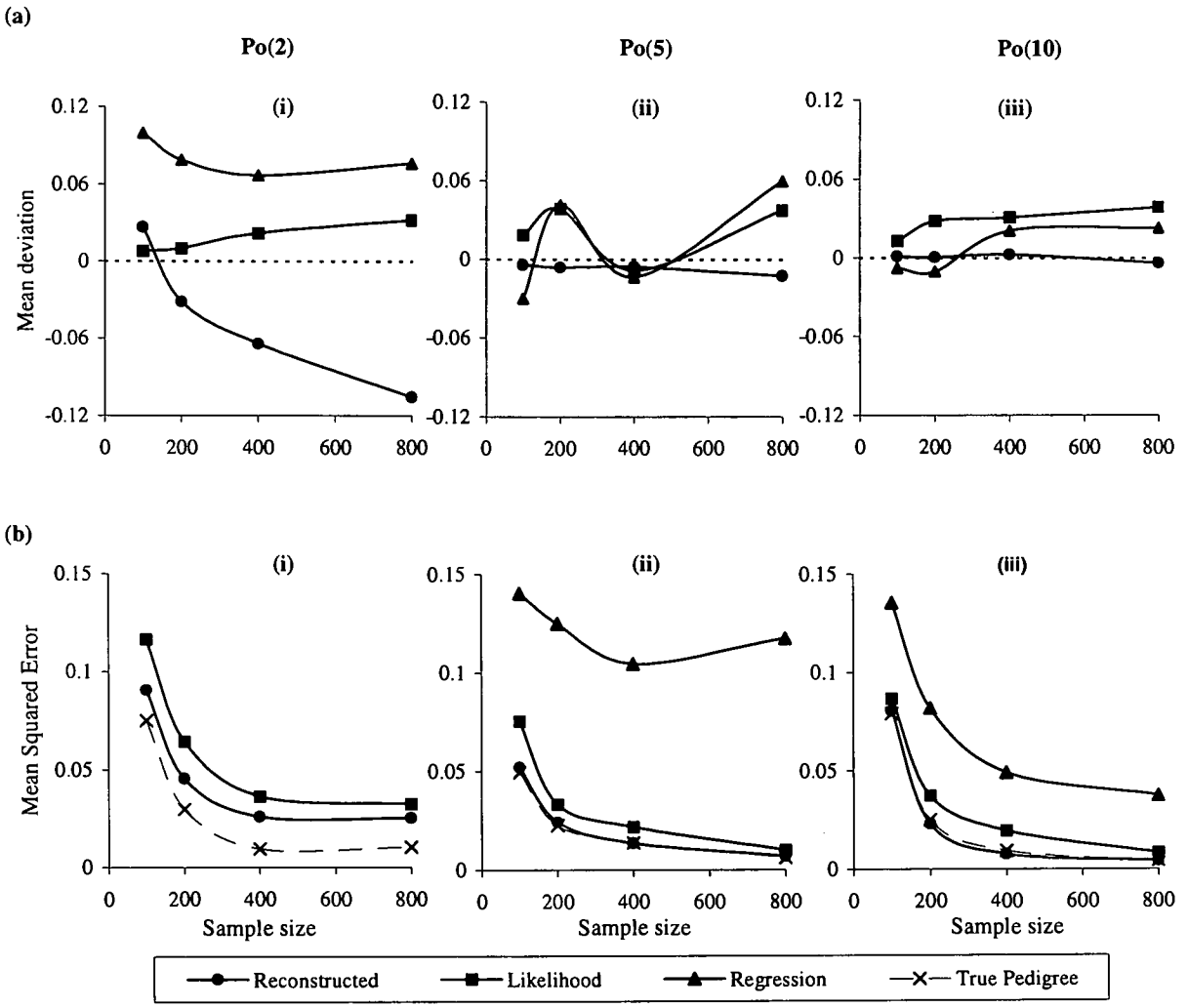

Figure 4.2: Results for full-sib family simulations with 10 marker loci with five alleles each, heritability 0.5, and varied numbers of individuals in the sample for the three simulated distributions of family size. (a) The change in mean deviation of marker-based heritability estimates from estimates obtained using actual pedigrees (zero line) with changing sample size for the three simulated distributions of family size. (b) The change in mean squared error of heritability estimates with changing sample size. Columns i, ii, and iii refer to family size distributions Po(2), Po(5) and Po(10). Values for the MSE of regression-based estimates of Figure 4.2b-i are off the scale (see text).

two or more smaller ones during the sib-ship reconstruction procedure, which reduces estimates of the variance between families and thereby heritability estimates (Falconer and Mackay, 1996). The pair-wise techniques share the same trends across the sample sizes, a result that possibly reflects the similar manner in which they weight family information, using size rather than information content.

A more important measure of performance of the techniques is summarised in Figure 4.2b-ii, which displays the change in MSE across the range of sample sizes. In all cases MSE is dominated by the sampling variance, rather than the bias, indicating that any bias is trivial compared with the level of precision of the techniques. Confirming previous results (Figure 3.1b), the regression procedure has much larger MSE than the pair-wise likelihood approach and has a slower decline in value than other techniques as sample size increases, indicating a less efficient technique. The pair-wise likelihood procedure has MSEs about 50% greater than those of the reconstructed pedigree, which are virtually indistinguishable from those of the true pedigree (the small difference being explained by the downwards deviation seen in Figure 4.2a-ii). MSE is approximately inversely proportional to the sample size for all the techniques except for the regression procedure.

4.4.2 Family size.

Simulations run using different distributions for family size showed similar trends to those obtained for families simulated with a Po(5) distribution, with those for the Po(10) distribution being virtually identical. Pair-wise techniques showed more consistent mean deviations in heritability estimates across the range of sample sizes with small mean family size (Po(2)) (Figure 4.2a-i) than with larger family size distributions (Figs. 4.2a-ii and 4.2a-iii). This is because information for variance component estimation from a population in which families are small comes mainly from pairs of individuals, rather than larger groups. The downwards bias in estimates obtained using reconstructed pedigrees with Po(2) family sizes is due to an increase in the number of type I errors, which at sample size 800 make up about a quarter of the number of pairs assigned as full-sibs. A greater amount of marker information would be required to increase the accuracy of reconstruction and reduce this bias in

estimates. Figure 4.2b-i shows that the MSE of reconstructed pedigree estimates is smaller than that of the likelihood-based estimates, with one-third of the MSE being explained by the bias at sample size 800. Sample variances for the reconstructed pedigree estimates are two-thirds those for the likelihood procedure.

With a Po(10) family size distribution, estimates using reconstructed pedigrees show almost no deviation from those using actual pedigrees (Figs. 4.2a-iii and 4.2b-iii). This indicates that few type I errors are made during pedigree reconstruction. Exclusions due to incompatible genotypes become more frequent with larger family sizes. Therefore at smaller family sizes there is a lower chance of incorrect families being detected than at larger family sizes, leading to greater numbers of type I errors, reducing accuracy (Figure 4.1), and increasing bias (Figure 4.2a-i).

Previous results (Chapters 2 and 3; Ritland 1996b) indicate that pair-wise procedures depend on there being sufficient variance of relatedness in order to be effective (i.e. they require that there be adequate numbers of groups of relatives within the sample). The simulation results supported this, with the regression-based procedure having extremely large MSE when actual variance of relatedness is low (with Po(2)) and smaller MSE with larger actual variance of relatedness (with Po(10)). The MSE values for the regression-based procedure are not plotted on Figure 4.2b-i since they are off the scale (0.53, 0.34, 0.38 and 0.29 for sample sizes 100, 200, 400 and 800, respectively). Less dramatic improvements in MSE are noted in the likelihood-based procedure where prior information on population structure compensates in part for less actual variation in relatedness.

4.4.3 Marker data.

Figure 4.3a shows the change in the accuracy as the amount of marker information is varied, simulated by changing the number of alleles at each locus. Accuracy improves at a diminishing rate with increasing allele number, with little difference in accuracy between six and ten alleles per locus. At the minimum number of alleles per locus (two) mean accuracy is about –0.2, reflecting a large number of type I errors (Equation 4.6) and resulting in large downward bias in heritability estimates.

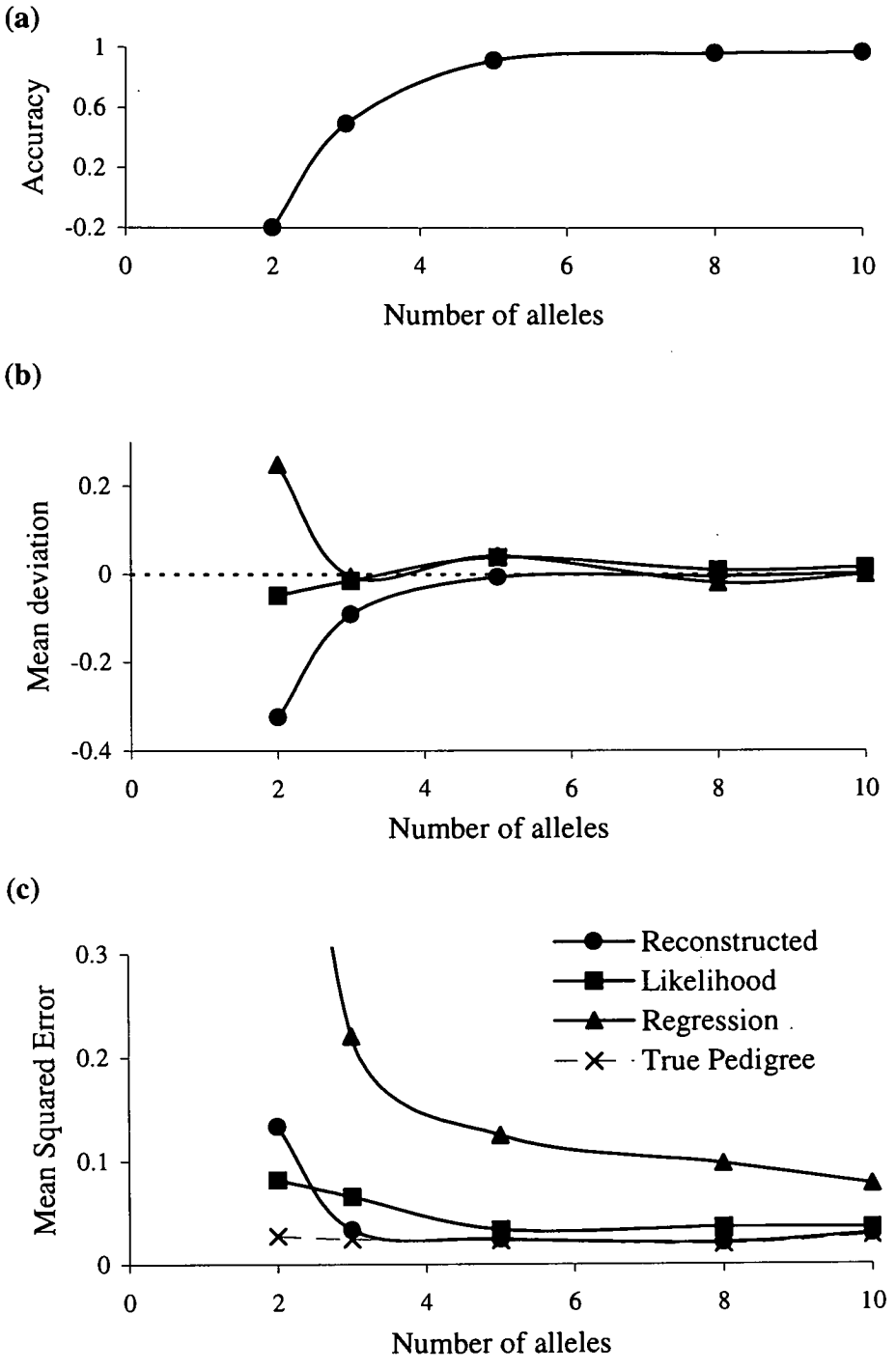

Figure 4.3: Results for full-sib family simulations with 200 individuals, 10 marker loci, heritability 0.5, actual family size distribution Po(5), assumed family size distribution for sib-ship reconstruction Po(5) and varied numbers of alleles per locus. (a) The change in accuracy of reconstructed families with changing numbers of alleles. (b) The change in mean deviation of marker-based heritability estimates from estimates obtained using actual pedigree (zero line). (c) Change in MSE of heritability estimates.

Figure 4.3b illustrates this point with the largest mean deviation of estimates occurring with low allele numbers. With the exception of low marker information (less than five alleles per locus), estimates made using reconstructed pedigrees are closer to true pedigree estimates than using either pair-wise technique. At low marker information the likelihood procedure shows least mean deviation from the true pedigree estimates.

Figure 4.3c shows the change in MSE with allele number. Again the regression procedure shows the largest MSE and sampling variances of parameter estimate. Deviations in the MSE of estimates using reconstructed pedigrees from those using the true pedigree were almost entirely explained by the bias (indicated by mean deviation). Since mean deviation for the likelihood procedure is also small (Figure 4.3b), its MSE is higher than that for the true pedigree due to sampling variance, and hence estimates made using the likelihood procedure have lower precision.

4.4.4 Assumed distribution of family sizes.

Table 4.1 summarises the change in accuracy, mean deviation and MSE when different assumptions are made about the family size distribution. Accuracy is lowest when an uninformative distribution for family size (i.e. every family size is equally likely) is assumed. Despite this, the mean difference between heritability estimates determined using pedigrees reconstructed with uninformative family size distributions and correct pedigrees is very small. Moreover, there is little increase in the MSE of these estimates, indicating only a little loss in precision. Using the correct distribution of family sizes, in this case Po(5), then accuracy and estimates are improved slightly, with MSE being almost identical to that of the true pedigree.

Accuracy is improved further if a Po(10) distribution is assumed, even though the true distribution is Po(5). This is because comparatively larger weights are placed on larger family sizes, thereby reducing the problem of large families being split into smaller families. However, if marker information is low, so that the probability of full-sib triplet exclusion due to incompatible genotypes is low, then increasing the weights of larger families can result in large numbers of incorrectly grouped

individuals. As mentioned previously, this causes larger bias in estimates of heritability than related pairs being classed as unrelated.

Distribution	Accuracy (Var)	Mean deviation	MSE
True pedigree	1 (0)	-	0.0260
Uninformative	0.848 (0.003)	-0.0111	0.0296
Po (5)	0.911 (0.002)	-0.0056	0.0266
Po (10)	0.943 (0.001)	-0.0229	0.0265

TABLE 4.1: Simulation results when different family size distributions are assumed during pedigree reconstruction (the same populations were reconstructed in each case). Simulated populations contained 200 individuals with the true family size distribution being Po(5). 10 loci with 5 equally frequent alleles were simulated. Heritability was set at 0.5. Mean deviation is the average deviation of the estimated parameter from the REML-derived estimate using correct pedigree information.

4.4.5 Updating allele frequencies.

Method	Accuracy (Var)	Mean deviation	MSE
True pedigree	1 (0)	-	0.0492
Not recalculated	0.518 (0.015)	-0.0608	0.0735
20,000 cycles	0.544 (0.014)	-0.0677	0.0717
10,000 cycles	0.545 (0.012)	-0.0651	0.0724
5000 cycles	0.552 (0.014)	-0.0503	0.0665
2000 cycles	0.545 (0.016)	-0.0570	0.0669

TABLE 4.2: Simulation results when parental allele frequencies were estimated after a different numbers of cycles. Simulation conditions: 100 individuals, 5 marker loci with 5 alleles each, heritability 0.5, actual family size distribution Po(5) and assumed distribution for sib-ship reconstruction Po(5).

Table 4.2 summarises the simulations investigating the recalculation of parental allele frequencies. Results show that there is some improvement in accuracy and in

parameter estimates as the number of reestimations of allele frequencies is increased. It would be expected that such allele reestimation of allele frequencies would have a greater effect in small populations where the variance in family size is large, since under these conditions the weights placed on allele counts from each family would be most incorrect. In such cases allele frequencies in the offspring generation might poorly represent allele frequencies in the parent generation. In larger populations, especially those with small family sizes, allele frequencies are more constant between generations (Falconer and Mackay, 1996).

4.4.6 Confidence levels

Simulations of 250 populations of size 200 were used to estimate the percentages of families of sizes three and four (numbers chosen as examples) that were reconstructed correctly. In each case two quantities were determined: the percentage of reconstructed families comprising only true full-sibs, and the percentage that were actually of that size, rather than a subset of a larger family. A Po(5) distribution of family size was assumed in the simulations and marker information was set at ten loci with five alleles each.

Family size	Number	Percentage true full-sibs	Percentage correct size
3	1819	98	56
4	1717	99	78

TABLE 4.3: Percentage confidence levels, determined by simulation, for the accuracy of families of size 3 and 4. Simulation conditions: 200 individuals, 10 marker loci with 5 alleles each, actual family size distribution Po(5) and assumed distribution for sib-ship reconstruction Po(5).

More families of size three than four were reconstructed, although families of size four were expected to be more frequent (Table 4.3). This is because the procedure tends to split larger families, which is reflected in the lower confidence that the families reconstructed as size three are actually of size three. Simulations

also show that reconstructed families of size four are more likely to be a genuine collection of full-sibs because of the relatively greater chance that an incorrect group of size four is excluded through incompatible marker information. Figure 4.4 shows the distribution of the actual sizes of families that were split to give reconstructed families of sizes three and four. Of particular note is the drop in the second point of each curve relative to the rest of the curve, which is due to the low likelihood placed on a family of size one under a Poisson distribution of family sizes. For example, a family of size four is unlikely to be split into a family of three and another of one, due to the low probability of observing a family of size one; while a family of size five may be more easily split into families of three and two.

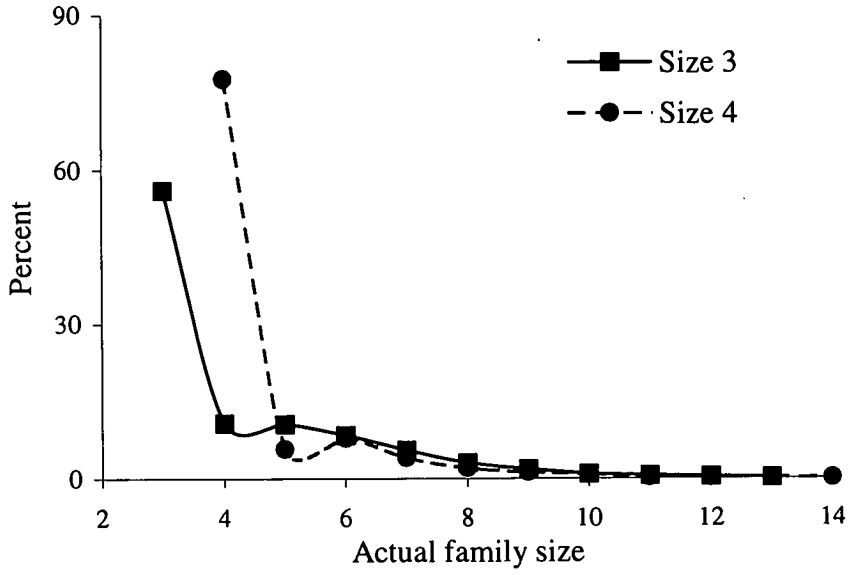

Figure 4.5: The distribution of the actual size of families reconstructed as being of sizes 3 and 4

4.5 Discussion.

Monte Carlo Markov Chain procedures to reconstruct sib-ships from a single generation of a population provide an improved means of estimating variance components compared to earlier techniques. Reconstructing the pedigree in this manner recovers in part some of the family specific weights lost in pair-wise

techniques, resulting in more efficient use of the information and lower mean squared errors in parameter estimates. Moreover since pedigrees are then assumed known, traditional procedures for partitioning the variance can be used, facilitating the incorporation of additional effects into the model, or the use of multivariate analysis on data collected from several traits. The sib-ship reconstruction process is independent of the quantitative data, and so actual values for the genetic parameters should not affect the technique's accuracy in estimating those parameters. For this reason, simulations examining the effects of the actual level of heritability were not run. A final attractive feature of these procedures is that population allele frequencies may be estimated simultaneously.

Since the Markov chain depends on the calculation of likelihoods, it is relatively straightforward to incorporate additional information, for example maternal genotype information, year of birth, or, in the case of plants, separation by distance provided a suitable dispersion parameter is known. This ease of modification allows the incorporation of possible genotyping errors into the algorithm, providing the probability of incorrectly typing a genotype (which may be done overall, or on a locus or allele specific basis) can be estimated prior to running the algorithm. Equation 4.4 could be modified to still sum over all parent allele combinations, but now allowing each of these alleles to change with some probability based on the probability of a mistyped locus. As this would slow the algorithm considerably, some assumptions restricting the number of transitions allowed may be required. Since mistyped alleles are more likely to cause families to be split rather than incorrect families to be formed, however, the present algorithm can cope with low levels of mistyping without modification, as only small bias in variance component estimates is introduced by this type of error. The effects of genotyping errors on the MCMC approach and other marker-based approaches are investigated in chapter 5.

There is interaction between sample size and the amount of marker information required to accurately reconstruct the families. With large sample sizes, the probability of obtaining type I errors increases, and more marker information is required to counteract this effect. Further investigation is required to determine the

extent of the interaction and to investigate the balance between the collection of individual data and the amount of marker data genotyped.

When compared to previous techniques this new approach performs admirably well, in many cases having lower mean deviation from the best available estimator, calculated from the known pedigree, and lower mean bias from the true parameter. In addition it yields mean squared errors that are often almost indistinguishable from those of the known pedigree; and as the MSE in most cases is dominated by sampling variance, any biases in parameter estimates become trivial.

There are a number of areas where caution must be taken when using relationships based on marker information to infer parameters. For example, in populations that are not in linkage equilibrium the information from each locus is not independent. Instead the likelihood of the marker data in any putative full sib family must be calculated from the probability of observing parental genotypes across all loci simultaneously rather than individually.

A second area for caution is in using reconstructed sib-ships to determine other parameters such as the average size of families, or the distribution of family sizes, which might be used in studies of reproductive success or other life history traits (Stearns, 1992). Reconstructed sib-ships have a tendency to under-estimate mean family size, and do not give an accurate description of its distribution. For example the reconstructed pedigrees examined to determine the confidence levels for families of size three and size four showed that more families of size three were reconstructed than of size four (Table 4.3), even though the latter were expected to be more common. Family sizes are underestimated due to the low probability of breaking down a correctly grouped set of individuals in order to join it to another group. For example, a family of size six may be reconstructed initially as two families of size three, but due to the low probability of moving through smaller family sizes be unable to combine into the correct single family. To combat this problem a step might be added to the algorithm that, with some probability, periodically attempts to combine two entire families (thereby attempting a direct jump across a valley of the likelihood surface), and/or break up an entire family (although this might prevent mean family size or population likelihood from stabilising). Results indicated that the use of an incorrect distribution of family size

that increases the expected frequency of larger families (the reconstructions operating under a Po(10) distribution) had improved accuracy over reconstruction using the correct prior (in this case a Po(5) distribution). However, such an approach to estimate mean family sizes and distributions is not advisable since it may cause family size to be overestimated, especially in populations with low marker information where exclusions based on incompatible genotypes are rare. In populations with ample marker information, exclusions often prevent large families being formed incorrectly. Simulation to estimate the expected bias in family size results in a circular problem, since the distribution of family sizes required to simulate the families is unknown. However, it may be possible to use simulation using the same sample size, the same level of marker information, the estimated family size distributions and the variance component estimates to estimate the size of bias shown in the variance components. This would require the assumption that any bias in subsequent variance component estimation approximately equals the bias in the original estimates, an assumption that may hold only if that original bias is small since variance components are bounded below by zero.

The choice of distribution of family size must also be considered cautiously for, as previously mentioned, assigning unrelated individuals to the same family can cause large downwards bias in estimates of between family variance and of genetic parameters derived from them. It is best therefore to choose a distribution that results in an underestimate of mean family size. Results indicate that when using an uninformative distribution of family sizes, the mean size of families in the reconstructed pedigree is consistently underestimated if the true distribution of family size is Poisson. This is because an uninformative distribution does not weight the creation of large families enough to break up two families of roughly the same size to recombine them as one larger family, even if they are actually one large family. The same problem occurs even when the correct distribution for family size is used, although to a lesser extent.

There are ways that the algorithm itself might be improved, leading to more likely population structures. These include the possibility of combining whole families, or subdividing an entire family (perhaps that with the lowest likelihood), mentioned above. An alternative approach to the population mixing, that could speed

up the algorithm but would not lead to better solutions, would be to treat the individuals systematically, moving first individual 1 to a random family, then individual 2 etc., rather than selecting and moving individuals completely at random. In addition alteration of the optimum acceptance/rejection rule for each change in the MCMC iteration could be considered.

In natural populations there are more than two classes of relationship, and in addition full-sib and half-sib groups are unlikely to be completely independent, perhaps being full-cousins. There are (at least) two approaches to the problem of dealing with multiple relationships: one approach is to assume that, since most of the information on heritability would come from close relatives, only these classes need to be considered (e.g. assume only full-sib and unrelated individuals are present, and ignore half-sibs, cousins etc.). The robustness of these techniques to deviations from the assumption of two classes of relationship is a complex problem and is examined in Chapter 5. Another approach is to attempt to include other classes of relationship into the model. Of particular interest are extensions to these techniques that allow nested maternal full-sib families within paternal half-sib families to be reconstructed. This is achieved through modification of equation 4.4 to multiply across the likelihood of the maternal half-sib families given a particular paternal genotype (see Chapter 5). Mixing would then move individuals between different mothers as well as fathers. However, the number of potential family structures would be extremely large and possibly intractable, even when using Markov Chain approaches. In addition, calculation of the likelihoods for individual families would be slow. Moreover, the ability to distinguish between relationships falls quickly with increased distance of relationship, and extremely large amounts of marker information, or some known relationships to build upon (e.g. known mothers), would be required to reconstruct accurate pedigrees (Thompson 1975). Furthermore, incorrectly assigned relationships would bias estimates of variance components to an unknown extent. For example, assigning groups of full-sibs as half-sibs would bias heritabilities upwards since a larger similarity in phenotype is attributed to smaller familial relationship.

Short cut methods and assumptions would need to be applied to make more complex situations tractable. For example if information is available on two or three

non-overlapping generations of a population, sib-ships could be reconstructed for each generation, constraining the sum of possible parental genotypes using the probability (if known) that a parent is contained within the samples collected from previous years. Generations could then be linked using the likelihood of the observed marker data and the probability that one or both parents are from the previous generation.

Chapter 5

Estimating variance components in more complex situations.

5.1 Introduction.

The studies presented in the preceding three chapters used simulated full-sib data sets to investigate the properties of three marker-based approaches to genetic variance component estimation. The studies evaluated their performance relative to each other with respect to, for example, the amount of marker information and the size of the sample. There are a number of situations, however, that deviate from the idealised conditions that were simulated and the behaviour of the estimation procedures must be assessed in these situations. For example, the marker information may be incomplete or contain errors, there may not be only full-sib groups within the sample and there may be other effects influencing trait variation in addition to additive genetic and uncorrelated environmental factors. The effects of these deviations from the idealised samples studied previously are investigated in this chapter.

In practice marker information is subject to genotyping errors. For example, if the locus contains a non-amplifying (null) allele, individuals heterozygous for that allele will be genotyped as homozygous for the remaining allele (e.g. Pemberton *et al.*, 1995). All types of mistyping will increase the level of noise in the inferred relationship information, and will thereby result in poorer estimates of the desired parameters. In the case of the MCMC approach, mistyping will more likely result in an exclusion of correct sib-ships (the type I errors of Chapter 4) than the creation of incorrect sib-ships (type II error). This type of error leads to only small bias being introduced to parameter estimates and only increases mean squared error slightly, provided that the level of mistyping is not too great (Chapter 4). A locus that has not

been typed would also result in more noise in the inferred relationship information. In the regression approach missing marker data must be taken into account in the calculation of the locus specific weights (Ritland, 1996b), with weights being recalculated so that they sum to one. In the likelihood approach a locus with a missing genotype is dropped completely from the calculation of relationship likelihoods. In the MCMC approach genotype information from a locus will still provide relationship data within a putative sib-ship even if one individual is not genotyped for that locus. Thus only that locus for that single individual should be excluded from the analysis. It is therefore expected that the MCMC will compensate for missing locus information better than the likelihood approach. For the purposes of clarification, both MCMC and likelihood approaches are based upon likelihood techniques, but throughout this study the likelihood approach refers exclusively to the pair-wise method of Chapter 3.

A second potential problem is the existence of additional classes of relationship within the sample. A number of predictions may be made about the behaviour of the approaches when the population contains additional classes of relationship and when the sample structure deviates from the assumed structure. When the sample contains additional classes of relationship, pair-wise estimates will be unbiased but will have larger sampling variance than REML based estimates, even when relationships are known exactly. In REML analysis the value of the relationship is taken into account by the inclusion of the relationship matrix and so different relationship types are weighted accordingly (Lynch and Walsh, 1998). The pair-wise procedures do not weight the information according to the relationship (chapter 3), and so it is unlikely that they will yield the REML based estimates even in balanced designs when there are greater than two classes of relationship.

Another prediction is that when the assumed population structure in the likelihood technique (i.e. the priors) is inaccurate then bias will be introduced into parameter estimates. For example, if it is assumed that there are only half-sibs and unrelated individuals within a sample that actually contains full-sibs as well, there might be upward bias found in heritability estimates because of full-sibs being classed as half-sib. Incorrect prior probabilities attached to the different relationship classes would have similar effects on variance component estimates, even when the

correct relationship classes are included in the analysis. The exact effect of incorrect priors will be specific to the nature of the inaccuracies of the priors. The MCMC approach operates in a different manner to the pair-wise approaches, and assigns specific relationships rather than keeping relationship information non-specific. Incorrectly assigned relationships lead to large bias in estimated variance components for the same reasons as outlined above (Chapter 4).

In this chapter, hierarchical populations, with full-sib groups within half-sib groups, are simulated to investigate the inclusion of additional relationships in the sample. In the regression approach relationship information is inferred through the calculation of pair-wise relatedness, which is a measure of the genetic distance between a pair rather than the exact relationship (Chapter 2; Ritland, 1996b). As a result the approach needs no modification to incorporate half-sibs into the analysis. The likelihood procedure is readily modified, and requires that the likelihood for each genotype pattern observable in a pair be determined for each type of relationship included in the analysis. In addition it requires that the probability of a pair falling into each category be known prior to analysis. The MCMC procedure must be modified to mix individuals over half-sib as well as full-sib families, and the likelihood equation for full-sib families (Equation 4.4) must be modified to sum over half-sib families.

The likelihood and MCMC approaches operate using definite relationship classes; the likelihood approach calculates the likelihood of each pair falling into each class and the MCMC actually assigns pairs to particular classes. As a result the inclusion of half-sibs into the sample allows the environmental covariance of full-sibs (due to mother and common environment) to be fitted into the phenotypic model. The likelihood approach is modified to include the covariance of full-sibs through re-parameterisation of the phenotype function. The MCMC-based approach need not be altered, since sib-ship reconstruction is independent of phenotype, but the additional effect is added at the stage of the REML analysis.

In natural populations some relationships might be known. For example mother-offspring pairs may be known through behavioural observation. Known relationships may affect the relationship data directly, through knowledge of the exact relationship between a pair, or indirectly, where the likelihood of a pair-wise

relationship is modified through knowledge of relationships between the individuals of a pair and others. For example, if the mother's genotype is known, the source of one of the alleles in the offspring can often be determined and this extra information may be included in the analysis. The likelihood and MCMC approaches are extendable to include information from known relationships, through alteration of the likelihood calculations used in each. Known relationships may be included directly into the regression-based approach, simply by setting all the locus-specific relationship estimates for that pair to the exact value of the relationship. However, it is more difficult to incorporate the indirect information gained about other relationships involving the members of the pair of known relationship. A known pair must be regarded as a single unit, and estimates of relatedness calculated between that pair and other individuals. This approach requires the inclusion of additional coefficients describing higher moments of the relatedness and the additive genetic effects, and so is difficult to implement.

Finally, the approaches may be modified to include information from other types of marker loci, for example, loci with dominant alleles or mitochondrial loci. Inclusion of data on a mitochondrial locus into both the likelihood and MCMC analyses immediately allows some relationship classes to be excluded. For example a pair with different alleles at a mitochondrial locus cannot be a mother-offspring pair, full-sibs or maternal half-sibs. However, mitochondrial loci tend to be much less polymorphic than microsatellite loci, and so their relative benefit might be low compared to the inclusion of a more polymorphic autosomal locus. Mitochondria are maternally inherited, and as a result do not provide suitable information for the estimation of additive pair-wise relationship, but can only exclude definite classes of relationship (see above). The regression approach does not reference exact relationship classes, and so information from a mitochondrial locus is not useful. Loci with dominant alleles may also be used to determine relationship information. However, they provide less information on the relationship than loci with co-dominant alleles and the same allele frequency distribution (Thompson, 1975; Milligan and McMurry, 1993), because the exact genotype at the locus cannot always be determined. Milligan and McMurry (1993) investigated the use of dominant versus codominant markers in the estimation of male mating success.

Results indicated that twice the amount of bias was observed in estimates of mating probabilities when dominant markers were used rather than the same number of codominant markers, and that the variance of estimates was about 50 percent larger. The regression approach, as well as the likelihood and MCMC approaches may be modified to include information from a locus with dominant alleles.

The objectives of the study presented in this chapter are to: 1). Investigate the effects of incorrect or missing genotype information on the variance parameters estimated using each of the approaches. 2). Investigate the inclusion of additional relationships into the samples through the simulation of hierarchical samples and derive modified forms of likelihood equation 4.4 to allow the MCMC approach to be applied to half-sib and hierarchical samples. 3). Investigate the performance of the likelihood and MCMC approaches when incorrect assumptions are made about the sample structure. 4). Extend the likelihood and MCMC approaches to include maternal genotype information. 5). Present methodology that allows the inclusion of information from mitochondrial and loci exhibiting dominance into the approaches, and assess their information content relative to loci with co-dominant alleles.

5.2 Statistical methods.

The statistical methods presented in this chapter are divided into three parts: the modifications to 1) the regression approach, 2) the likelihood approach and 3) the MCMC approach.

5.2.1 The regression approach.

5.2.1.1 A mitochondrial locus.

The probability that the alleles at a mitochondrial locus, in two individuals are identical is dependent upon whether they share a female ancestor through the maternal line (i.e. no male individuals break the direct line of descent). If the pair do share such an ancestor the probability of identity is one. Otherwise, the probability of sharing identical haplotypes is dependent upon the allele frequencies at the locus.

Identity is therefore independent of the additive genetic relationship, and so information from a mitochondrial locus may only be used to exclude certain relationships, e.g. full-sibs, and not to estimate relatedness. Information from a mitochondrial locus cannot therefore be included into regression approach.

5.2.1.2 A locus with a dominant allele.

The regression-based approach may also be extended to incorporate information from a locus that exhibits dominance amongst its alleles. Again this is through modification of the estimator of relatedness used. For the purposes of this study dominant loci with only two alleles, one allele (*B*) being completely dominant over the other (*b*), are investigated, although more complex patterns of dominance may be used. With dominant alleles some of the possible genotype patterns cannot be distinguished from each other. In these situations estimates of the relatedness must be calculated for all of the possible genotypes that can give rise to the observed pattern in the pair, and an average relatedness based on the probability of obtaining the underlying genotypes calculated. For example, using the Lynch and Ritland (1999) estimator, the estimate of the relatedness between an individual (X) homozygous for the recessive allele and an individual (Y) that displays the dominant phenotype (denoted *B* -) is:

$$R_{XY} = [R_{XY} \mid X = bb, Y = BB]P(BB \mid B-) + [R_{XY} \mid X = bb, Y = Bb]P(Bb \mid Bb) \quad (5.1)$$

where the recessive individual (X) is the reference individual. $P(BB \mid B-)$ and $P(Bb \mid B-)$ are the probability that the *proband* (Y, the non-reference individual) is homozygous and heterozygous respectively for the dominant allele given that a dominant phenotype was observed. $[R_{XY} \mid X = bb, Y = BB]$ is the estimate of the relationship given X has genotype *bb* and Y has genotype *Bb*. In the biallelic example used in this study the sampling variance is calculated assuming that the pair are unrelated and depends on whether the reference individual displays the recessive or dominant phenotype. In the recessive case the sampling variance is:

$$\mathrm{Var}[R_{XY} \mid X = bb] = \frac{p_b^2}{1 - p_b^2}, \tag{5.2}$$

where P_b is the frequency of the recessive allele and in the dominant case:

$$\mathrm{Var}[R_{XY} \mid X = B-] = \frac{1 - 11p_b + 43p_b^2 - 65p_b^3 + 8p_b^4 + 56p_b^5 - 16p_b^6 - 16p_b^7}{(1 + p_b)^3 (1 - 2p_b)^4}. \tag{5.3}$$

As with the standard relatedness estimator of Lynch and Ritland (1999), the above estimator gives undefined results when the dominant and recessive allele are equally frequent. With loci that show more complex patterns of dominance, calculation of the sampling variance may be achieved numerically by exhaustive summation over the possible genotypes for the pair assuming that they are unrelated.

5.2.2 The likelihood-based approach.

5.2.2.1 Including half sibs.

Inclusion of half-sibs into the likelihood approach (Equation 3.1) is straightforward, and requires that the prior probability of a pair being half-sib be known as well as the prior probabilities of the other relationship classes. In addition, inclusion requires the derivation of the likelihoods for the seven possible genotypes observable at a single locus in a diploid pair and derivation of the distribution of the function describing pair-wise phenotype. Table 5.1 summarises the genotype likelihoods for unrelated, half-sib and full-sib pairs. The overall likelihood of a pair is calculated as the product of the individual locus likelihoods.

Inclusion of half-sibs allows the environmental covariance of full-sibs (σ_C^2), due to mother and common environment to be included into the phenotypic model, which is achieved through the re-parameterisation of the phenotypic function. In this study the phenotypic function based on the difference of the individuals phenotypes

is used, since this is computationally simple and shows low bias and low mean squared errors (Chapter 3). Table 5.2 summarises the distributions of the phenotypic difference under an additive model for unrelated, half-sib and full-sib pairs when is σ_C^2 is included and excluded.

Genotype	Unrelated	Half-sib	Full-sib	Factor
$A_iA_i - A_iA_i$	$4p_i^2$	$2p_i(1+p_i)$	$(1+p_i)^2$	$0.25p_i^2$
$A_iA_i - A_iA_j$	$4p_i$	$1+2p_i$	$1+p_i$	$p_i^2 p_j$
$A_iA_i - A_jA_j$	4	2	1	$0.5p_i^2 p_j^2$
$A_iA_j - A_iA_j$	$8p_ip_j$	$p_i+p_j+4p_ip_j$	$1+p_i+p_j+2p_ip_j$	$0.5p_i p_j$
$A_iA_i - A_jA_k$	4	2	1	$p_i^2 p_j p_k$
$A_iA_j - A_iA_k$	$4p_i$	$1+4p_i$	$1+2p_i$	$2p_i p_j p_k$
$A_iA_j - A_kA_l$	4	2	1	$2p_i p_j p_k p_l$

Table 5.1: The probabilities for the possible pair-wise genotypes observed in diploid individuals in unrelated, half-sib and full-sib pairs. *i*, *j*, *k* and *l* index mutually exclusive alleles and p_i denotes the frequency of allele *i*. For ease of expression, factors common to each pair-wise genotype have been listed separately.

	Model	
Relationship	Additive	Environmental covariance of full-sibs
Unrelated	$N(0,\ 2\sigma_A^2 + 2\sigma_E^2)$	$N(0,\ 2\sigma_A^2 + 2\sigma_C^2 + 2\sigma_E^2)$
Paternal half-sib	$N(0,\ 1.5\sigma_A^2 + 2\sigma_E^2)$	$N(0,\ 1.5\sigma_A^2 + 2\sigma_C^2 + 2\sigma_E^2)$
Full-sib	$N(0,\ \sigma_A^2 + 2\sigma_E^2)$	$N(0,\ \sigma_A^2 + 2\sigma_E^2)$

Table 5.2: The distributions of phenotypic difference under additive models, including and excluding the covariance of full-sibs (σ_C^2) for unrelated, half-sib and full-sib pairs.

5.2.2.2 With known maternal information.

When maternal identity and genotype information are known, the source of one of the alleles in the offspring is known. If the offspring is heterozygous for the same allele as the mother the maternal allele can not be determined. Table 5.3 shows two

simple equations that may be used to calculate the likelihood of any genotype pattern for a pair with known maternal genotypes. When the maternal allele cannot be determined for one individual of the pair, the appropriate equations from Table 5.3 must be weighted by a half and summed over the two possible maternal alleles.

Known maternal genotypes		
Genotype pattern	Unrelated [1]	Half-sib [1]
A_i* , A_i*	p_i^2	$\frac{p_i}{2}(1+p_i)$
A_i* , A_i*	$2p_i p_j$	$\frac{p_i p_j}{2}$
Mitochondrial locus		
Haplotype pattern	Unrelated [2]	Full-sib [2]
A_i , A_i	p_i^2	p_i
A_i , A_j	$2p_i p_j$	0
Diallelic locus with dominance		
Genotype pattern	Unrelated	Full-sib
B- , *B-*	$p_B^2\left(2-p_B^2\right)^2$	$\frac{p_B^2}{4}\left(4+5p_B-6p_B^2+p_B^3\right)$
B- , *bb*	$2p_B\left(1-p_B\right)^2\left(2-p_B^2\right)$	$\frac{p_B}{2}\left(1-p_B\right)^2\left(4-p_B\right)$
bb , *bb*	$\left(1-p_B^2\right)^4$	$\frac{1}{4}\left(1-p_B\right)^2\left(2-p_B\right)^2$

1 These categories assume the mothers of the individuals are unrelated. If the individuals share the same mother the unrelated and half-sib would become would be half-sib and full-sib respectively.

2 These categories assume no paternal relationship.

Table 5.3: The likelihoods for the pair-wise genotype patterns when maternal information is known and for a mitochondrial locus and a diallelic locus with a dominant allele for unrelated and full-sib pairs. *i* and *j* index mutually exclusive alleles. * indicates a second allele that is known to be inherited from the mother.

When the maternal allele cannot be determined for both individuals the equations are weighted by a quarter and summed. The equations of Table 3.1 and Table 5.1 may be derived in a similar manner to this, with the equations of Table 5.3 being weighted using the probability of an allele from the population rather than from a single individual. Using the same approach, the likelihood equations for situations where only one of the individuals has a known mother can be derived. Thus situations where only some of the maternal genotypes are known are easily accommodated.

In addition, the equations of Table 5.3 are used to determine the likelihood of a pair being sibs or unrelated. Whether the pair are full-sibs or half-sibs depends upon the identity of the mother. In a hierarchical population where full-sib families are nested within half-sib families, a pair sharing the same mother are automatically full-sib, and only half-sib groups need be distinguished.

5.2.2.3 Including other types of marker loci.

Inclusion of information from additional types of marker loci such as dominant or mitochondrial loci is straightforward for the likelihood-based procedure, again requiring the derivation of the likelihoods for all observable genotype patterns. In this study dominant loci are assumed to have two alleles, one that is always detected when present and one that is detected only when it is present in the homozygous state. Other more complex patterns of dominance may be easily modelled. Table 5.4 summarises the likelihoods of the pair-wise genotype patterns for a mitochondrial and a dominant marker for unrelated and full-sib pairs. The likelihood of a pair is again calculated as the product of the likelihoods for the individual loci.

5.2.3 The MCMC-based approach.

5.2.3.1 Paternal half-sib populations.

The basic algorithm for samples containing only paternal half-sib families is identical to the algorithm outlined in the previous chapter, with individuals being mixed over half-sib families using the same conditions for accepting or rejecting a change.

Unlike the investigation of the previous chapter, in this study individuals were considered sequentially rather than randomly. It was found that this improved the time taken for the sib-ships to be reconstructed. It is assumed that each individual in the paternal half-sib families has a different mother.

Calculation of the probability of the genotypes observed in a putative half-sib family requires that equation 4.4 be modified to sum over the genotypes of one parent only. The resulting equation is:

$$L_{\text{genotypes}} = \prod_{\ell} \left[\sum_{w=1}^{b_\ell} \sum_{x=1}^{b_\ell} p_{wx} \prod_{c=1}^{n_f} [L(g_{c\ell})] \right], \tag{5.4}$$

where ℓ denotes independent marker loci, b_ℓ the number of alleles at locus ℓ, w and x index the paternal alleles, c indexes an individual from a putative half-sib family of size n_f, p_{wx} is the ordered genotype frequency of the common parent and $L(g_{c\ell})$ is the probability of observing the genotype of individual c at locus ℓ given that one of its alleles is from the father and one is from the mother. For example, if the father has genotype (1, 2) then $L(g_{c\ell}) = \frac{1}{2} p_1 + \frac{1}{2} p_2$ when the offspring genotype is also (1, 2), and when the parent has genotype (1, 1) and the offspring has genotype (1, 2) then $L(g_{c\ell}) = p_2$. Allele frequencies p_1 and p_2 are the probability of selecting an allele of type 1 and type 2 from a random individual from the population.

With all approaches to inference of relationship the ability to resolve half-sibs is lower than the ability to resolve full-sibs for the same level of marker information (Thompson, 1975; Blouin *et al.*, 1996; Ritland, 1996a). Consequently, a larger amount of mixing (i.e. a greater number of iterations) is required before a stable or reasonably stable set of sib-ships is constructed.

Calculations using 5.4 are much faster than calculations using 4.4, since summation is only over one parent instead of two. Even so, 5.4 may be speeded up in a similar manner to 4.4 by fixing one of the paternal alleles using the genotype of one of the individuals in the putative half-sib family. This modified form of 5.4 is presented in Appendix 2.

5.2.3.2 Hierarchical populations.

The algorithm for the reconstruction of full-sib families within half-sib families is also similar to the algorithm outlined in the previous chapter. The major difference is in the mixing of individuals. At each step of the chain the candidate individual was either moved to a (randomly selected) half-sib family or remained in the same half-sib family, each option having a probability of one half. Then the individual was either moved to an existing full-sib family within the chosen half-sib family or formed a new full-sib family. An individual had an equal chance of moving to each of the existing full-sib families as it did of forming a new one.

For this study an additional step was added to the mixing routine. As mentioned previously, full-sib families have greater resolving power than half-sib families and so full-sib families tended to be generated in preference to half-sib families. This meant that half-sib families were often split into their component full-sib families. Since mixing occurred previously only at an individual level, it was difficult for two full-sib families to be joined, as this required the crossing of a "trough" in the likelihood surface. Therefore, periodically (e.g. every 200 iterates) an entire half-sib family was joined to a randomly selected half-sib family (full sib families within those families remained separate) in an attempt to step directly over these likelihood "troughs". The same conditions for accepting and rejecting a change as for single individual mixing were used for this type of mixing.

The probability equation for the observed genotypes within this type of sib-ship is expressed as:

$$L_{\text{genotypes}} = \prod_{\ell}\left[\sum_{w=1}^{b_\ell}\sum_{x=1}^{b_\ell} p_{wx}\cdot\left[\prod_{m=1}^{n_f}\sum_{y_m=1}^{b_\ell}\sum_{z_m=1}^{b_\ell} p_{y_m z_m}\prod_{c=1}^{n_m} L(g_{c\ell})\right]\right], \qquad (5.5)$$

using the notation of 5.4, with n_f now indicating the number of full sib-ships within the half sib-ship and with $L(g_{c\ell})$ now being the probability of observing the genotype of individual c at locus ℓ given that one of its alleles is from the father (whos genotype is wx) and one is from the mother (genotype yz). In addition m

indexes the full-sib family, y_m and z_m index the maternal alleles of full-sib family m, $p_{y_m z_m}$ is the ordered genotype frequency of the mother and n_m indicates the size of full-sib family m. For ease of expression, the square brackets in 5.5 divide the expression into paternal and maternal sections. Equation 5.5 reduces to 4.4 when n_f= 1 and to 5.4 when $n_m = 1$.

Calculations using 5.5 are much slower than calculations using 4.4, since summation is only over many parents. However calculation may be speeded up by fixing one of the paternal alleles using the genotype of one of the individuals in his half-sib family, and by fixing one of the maternal alleles using the genotype of one of the individuals in her full-sib family. This modified approach to 5.5 is presented in Appendix 2.

5.2.3.3 With maternal information.

When maternal identity and genotype information are known, the information should be included in the analysis to improve parameter estimation. Equation 5.4 may be modified to calculate the likelihoods of the sib-ships when maternal information is known, and simply requires the redefinition of $L(g_{c\ell})$. Unless the mother and offspring are both heterozygous for the same alleles, the paternal allele within the offspring can be determined and $L(g_{c\ell})$ would simply be the probability of the offspring receiving that allele from the father given the paternal genotype currently defined by the sum. In cases where the mother and offspring are heterozygote for the same allele, each offspring allele in turn is assumed to come from the mother, and the two probabilities are weighted by a half and summed.

The likelihood of the paternal sib-ship may also calculated in situations where only some of the maternal genotypes are known. In these situations the likelihoods of the individuals with known mothers (given the paternal genotype) are multiplied by the likelihoods of the putative maternal sib-ships (given the paternal genotype) and summed across all possible paternal genotypes.

5.2.3.4 Including other types of marker loci.

Information from a mitochondrial marker locus may be easily included into the MCMC approach, and does not require the calculation of any additional likelihoods. All members of a maternal sib-ship must contain the same haplotype and so putative maternal sib-ships where this is not the case are automatically rejected. No information concerning paternal sib-ships can be gained from a mitochondrial locus.

Information from a locus exhibiting dominance in its alleles is also easily included and requires that $L(g_{c\ell})$ of equations 4.4, 5.4 and 5.5 be redefined as the likelihood of observing the genotype (dominant or recessive) of locus ℓ in individual c summed given parental genotypes. Allele frequencies for a diallelic locus with one dominant allele are calculated as the square root of the frequency of the homozygous recessive genotype.

5.3 The simulated populations.

For all the samples, phenotypic data were simulated as described in section 2.3. In simulations including σ_C^2, an additional normal deviate was added to each full-sib family. As in previous chapters, heritability estimates were used as a summary statistic for the additive genetic and the environmental variance components. In all cases heritability estimates were compared against REML estimates made using the simulated relationships. In samples without σ_C^2 both additive genetic and environmental variance were simulated as having values of 0.5.

5.3.1 Genotyping errors.

Simulated full-sib data sets were used to evaluate the performance of each of the approaches in the presence of genotyping errors and un-typed loci. Populations were simulated with 200 individuals in full sib families following a Po(5) distribution, with each individual having 10 loci with 5 equally frequent alleles at each. Populations where 0, 1, 2, 4, 8, 16 and 32 percent of all marker loci were randomly

chosen and dropped from the analysis, and populations where 0, 1, 2, 4, 8, 16 and 32 percent of all loci were mistyped were simulated. Mistyping was simulated by selecting a random locus and assuming that one of the alleles at that locus failed to amplify. A heterozygous locus would therefore be typed as homozygous for the remaining allele. Each set of conditions was replicated 250 times.

Variance components were estimated by : i). The regression-based technique using the Lynch and Ritland (1999) estimator for pair-wise relatedness. ii). The likelihood technique using the phenotypic function based on the pair-wise difference (Chapter 3) using correct prior probabilities. iii). The MCMC approach reconstructing full-sib families using a uniform prior for the family size distribution (Chapter 4). In all cases allele frequencies were estimated from the sample, and in the case of the MCMC approach they were not updated during reconstruction.

5.3.2 Mitochondrial and dominant marker loci.

Simulations were run to compare the relative benefit of a single mitochondrial locus with two equally frequent alleles with the benefit of an autosomal locus. Full-sib family populations were simulated with 5 autosomal loci each with 5 co-dominant alleles, a single additional autosomal locus with either 2, 3, 4 or 5 equally frequent alleles and a single mitochondrial locus with 2 equally frequent alleles. Variance components were estimated twice for each simulated sample and for MCMC and likelihood approaches: once with the mitochondrial locus and the 5 constant autosomal loci, and once with the variable autosomal locus and the 5 constant autosomal loci.

Similar simulations were run to determine the comparative benefit of diallelic loci where one allele is dominant. In this set the number of loci exhibiting dominance was varied, being set as 1, 2, 3 or 4 and comparison was made with a single autosomal locus with two co-dominant alleles. The allele frequencies of the dominant loci were set so that the genotype frequency of the homozygous recessive was 0.5 (i.e. f(bb) = 0.5 and f(b) = 0.707). An identical allele frequency distribution was simulated at the autosomal locus. Again comparison was made in a background

of 5 autosomal loci with 5 co-dominant alleles each. Each set of conditions was replicated 120 times.

5.3.3 Simulations with half-sibs.

A number of different structures were simulated to investigate the inclusion of half-sibs into the data-set, with marker information and half-sib and full-sib family structures being varied (Table 5.4). Only data from the offspring generation were included in the variance component analysis. Each set of conditions was replicated 100 times.

Set	Sires	Dams / sire	Offspring / Dam	Loci
A	20	2	5	5
B	20	2	5	10
C	20	2	5	20
D	20	5	2	5
E	20	5	2	10
F	20	5	2	20
G	10	4	5	10
H	10	5	4	10
I	40	1	5	10
J	40	5	1	5
K	40	5	1	10
L	40	5	1	20

Table 5.4: Summary of the half-sib / full-sib population structures simulated. Five equally frequent co-dominant alleles were simulated at each locus. A to L are names given to each structure. Sire – number of sires. Dams / sire – number of dams within sire. Offspring / Dam – number of offspring within Dam. Note that I contains full-sib families only, and J, K &L contain half-sib families only.

Variance components were estimated for all the simulated samples using the three approaches to variance component estimation, but using different assumptions about population structure or available data. For the regression-based approach samples were analysed using both estimated relationships and known relationships. Known relationships were also used to estimate variance components using the likelihood framework. Estimated relationship data were then used in the likelihood procedure under four models containing different assumptions about the sample: i). The sample contained full-sibs and unrelated pairs only. ii). The sample contained half-sibs and unrelated pairs only. iii). Full-sib, half-sib and unrelated pairs were all present. iv). The maternal genotype information of individuals in the sample was known. The MCMC approach was run using the same four models, with the additional assumption that for the hierarchical samples full sib-ships were nested within half sib-ships.

For the likelihood approach, accurate prior probabilities were used. Thus analysis assuming both half-sibs and full-sibs were present gave the same variance component estimates as analysis assuming only one class of sib was present, when there was actually only that particular class of sib present. For the MCMC approach an uninformative distribution was placed on sib-ship sizes, so that every family size was equally likely, and allele frequencies were not recalculated during reconstruction.

The simulated population structures containing both half and full-sibs (Sets A to H, Table 5.4) were repeated an additional 50 times, but under a phenotypic model that included a σ_C^2. For these populations the additive genetic variance and σ_C^2 were simulated as 0.25, and the (residual) environmental variance simulated as 0.5. These populations were analysed using the likelihood and MCMC approaches and under the assumption that both half sib-ship and full sib-ship families were present in the sample.

The results of Chapter 3 showed that bias was introduced to variance component estimates made using the likelihood approach because phenotype was included in the calculation. A set of simulations was therefore run where the covariance of full-sibs was varied, while the other variance components remained constant. Hierarchical population structures of size 320 were generated, with each

structure comprising 20 paternal half-sib families containing four full-sibs families of size four. Additive genetic and environmental variance remained constant at 0.25 and 1 respectively and the covariance of full-sibs was simulated as 0, 0.2, 0.4, 0.6, 0.8 and 1. Marker data comprised 10 loci, each with 5 equally frequent alleles. Each set of conditions was replicated 250 times. Estimates of heritability and the covariance of full-sibs obtained using the likelihood approach were compared against estimates obtained using REML techniques and the known pedigree.

5.4 Results.

5.4.1 Genotyping errors

When using the likelihood and MCMC approaches, the bias of the heritability estimates becomes more negative as the percentage of mistyped loci increases (Fig 5.1a). The regression based approach gives some indication of an initial upwards bias as the percentage of mistyped loci increases, before bias starts to fall at about the same rate as the likelihood and MCMC approaches. The largest bias observed was about -0.06, which was for the MCMC approach when 32% of the marker loci were mistyped. Downward bias is observed in the techniques because random errors in the marker data decrease the chance of detecting information about related individuals. Related individuals are more likely to be classed as unrelated thereby causing heritabilities to be underestimated. MSE increases only slightly as the percentage of mistyped loci increases (Fig 5.1b).

Similar trends were observed when the percentage of loci with missing genotype information increases (Fig 5.1c and 5.1d), with the exception of an increasing upwards bias in the regression approach. Of note is the more rapid increase in the MSE of the regression approach than for the other two approaches, indicating a lower tolerance of the approach to missing marker data.

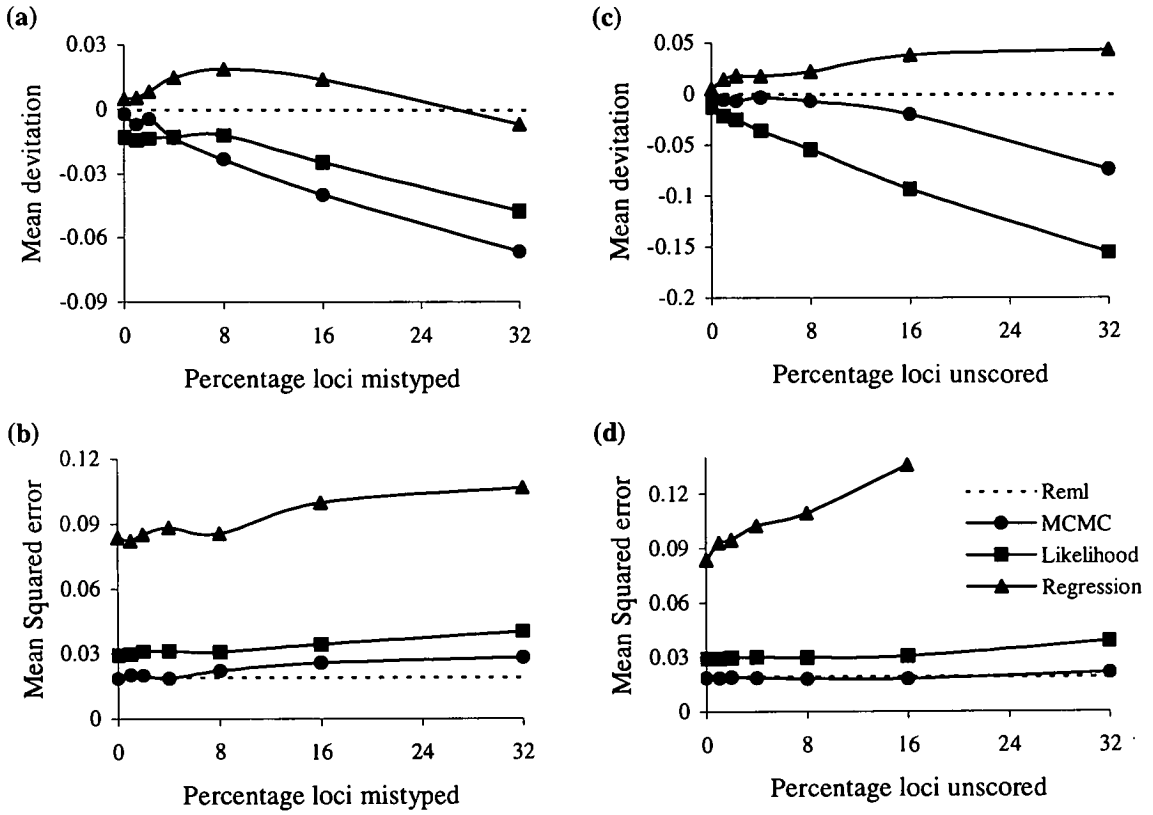

Figure 5.1: **a**) Change in mean deviation of heritability estimates from REML estimates and **b**) Mean squared error as the percentage of mistyped locus increases. **c**) Change in mean deviation and **d**) Mean squared error as the percentage of missing loci increases. For simulation conditions see text.

5.4.2 Mitochondrial and dominant marker loci.

For both the likelihood and MCMC approaches the presence of a single mitochondrial locus with two equally frequent alleles yields heritability estimates with the same MSE as an autosomal locus with between 3 and 4 equally frequent alleles (Fig 5.2). With three or less alleles at the autosomal locus analysis including a mitochondrial locus has lower MSE, and with four or more alleles at the autosomal locus analysis including a mitochondrial locus has larger MSE. Since the regression approach does not specify any relationship classes, the presence of a mitochondrial locus does not contribute any information to an analysis.

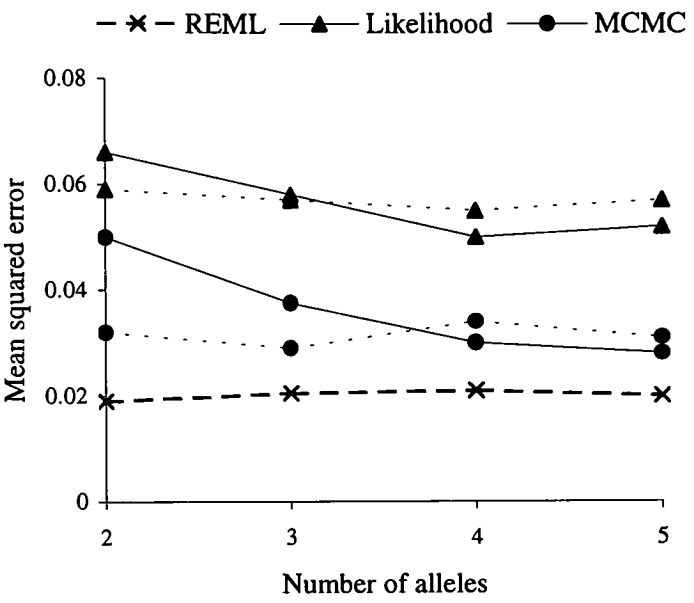

Figure 5.2: The MSE of heritability estimates obtained using the marker-based approaches when, in addition to 5 autosomal loci, either a single mitochondrial locus with two equally frequent alleles is included in the analysis (dotted lines) or an autosomal locus with varying numbers of alleles (solid lines). Allele number indicates the number of alleles simulated at the variable autosomal locus. Sample size was 200, and comprised full-sib families with a Po(5) distribution.

Similar investigation of dominant loci indicates that for the likelihood and MCMC approaches it takes about two dominant loci with alleles set so that the frequency of the homozygous recessive is 0.5 to provide the same level of information as a single autosomal locus with the same allele frequency distribution (Fig. 5.3). The use of a dominant locus in the regression approach improves estimation properties, but within the range of loci simulated never improves upon the presence of a co-dominant autosomal locus. The slow decrease in size of the MSE when the number of dominant loci simulated increases indicates that the relative benefit of a dominant locus with 2 alleles is low. Since the amount of marker information was low, and only 120 simulations were run for each set of conditions, fluctuations were observed in the simulations at constant marker information (the solid lines of Fig. 5.3).

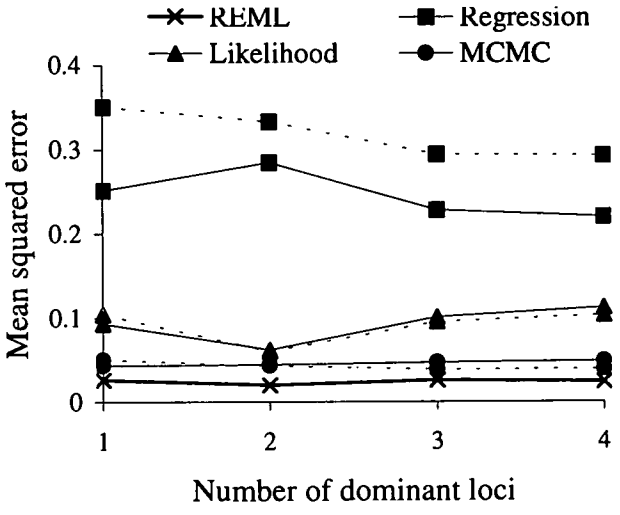

Figure 5.3: The MSE of heritability estimates obtained using the marker-based approaches when, in addition to 5 autosomal co-dominant loci, either a single autosomal locus is included in the analysis (solid lines) or a variable number of dominant loci are included (dotted lines). Sample size was 200, and comprised full-sib families with a Po(5) distribution.

5.4.3 Populations containing half-sibs.

For convenience, analyses of these results are divided into sections., with the results for samples including a simulated σ_C^2 being presented in section 5.4.4.

5.4.3.1 Half-sib families only.

Since the sample is balanced and contains only one class of relative (half-sib), using known relationships in either regression or likelihood techniques yields heritability estimates identical to those derived from REML-based estimates using known relationships. Samples containing half-sibs yield much less accurate estimates of variance components than populations containing full-sib families of the same size and the same level of marker information (Set I versus K, Table 5.5). The regression technique contains the least prior information about the population structure, and gives heritability estimates with low bias but large MSE when inferred relationship

Set	REML	Regression		Likelihood		MCMC *
		Known	Inferred	Known	Inferred *	
I	0.007	0.007	-0.053	0.007	0.039	0.006
40 \ 1 \ 5 \ 10	(0.022)	(0.022)	(0.145)	(0.022)	(0.030)	(0.023)
J	-0.038	-0.038	0.025	-0.038	-0.338	-0.430
40 \ 5 \ 1 \ 5	(0.066)	(0.066)	(1.243)	(0.066)	(1.454)	(0.199)
K	-0.005	-0.005	-0.081	-0.005	0.073	-0.192
40 \ 5 \ 1 \ 10	(0.063)	(0.063)	(0.852)	(0.063)	(0.442)	(0.087)
L	0.041	0.041	-0.036	0.041	0.053	-0.017
40 \ 5 \ 1 \ 20	(0.064)	(0.064)	(0.465)	(0.064)	(0.247)	(0.079)

* Set I analysed using full-sib form of the approach and sets K–L analysed using half-sib form.

Table 5.5: Summary of bias (upper line) and MSE (bracketed - lower line) of the heritability estimates obtained from different sample structures. Summary of structure of sets I to L (see Table 5.4) are in the column headed 'Set' in form: sire \ dams within sire \ progeny within dam \ locus number. Known – Actual relationships used in the analysis. Inferred – marker-based relationship information used in analysis.

information is used. Even with larger amounts of marker data (Set L, Table 5.5), estimates still have very large MSE. This unreliability reflects the lower accuracy with which information on more distant relationships can be inferred from marker data. The likelihood technique compensates in part for the reduced ability to resolve relationship information from more distant relationships by incorporating prior information into the analysis. However with low marker information (5 loci) heritability estimates are biased downwards and MSE is large (Set J, Table 5.5). With higher levels of marker information (10 and 20 loci) the magnitude of the bias decreases, and MSE is about half that of the regression technique (Sets K and L, Table 5.5). When plotted graphically, and including a greater range in the number of loci (not shown), the trend of the curve describing bias of likelihood estimates for half-sib populations is the same as that for full-sib populations (Fig. 3.2b) but shifted to the right, reflecting the larger amount of marker information required in half-sib analysis. Additive genetic variance is estimated as four times the between family variance in half-sib designs, and twice the between family variance in full-sib designs. Therefore inaccuracies in variance component estimates caused by inaccurately inferred relationship data, which are already more noisy than data inferred about full-sib relationships due to the low level of the relationship, are

further magnified. The MCMC approach uses less prior information than the likelihood approach, requiring only the assumption that the population is comprised of half-sibs. At lower levels of marker information (5 and 10 loci) the MCMC approach is biased downwards, but as marker information increases (20 loci) estimates approach the REML derived estimates made using the simulated relationships.

5.4.3.2 Hierarchical sample structures.

With known relationships: When known relationships are used in the pair-wise regression and likelihood frameworks the resulting estimates of the heritability show little mean bias and have larger MSE than the REML derived estimates. This reflects the poorer ability of the pair-wise approaches to weight data from different relationships. MSEs of the regression estimates are up to 50% larger than the REML derived estimates, while the MSEs of the likelihood estimates are closer to the those of the REML estimate (Table 5.6). This suggests that the likelihood method weights the data from the different relationships classes in a more efficient manner than the regression method. When all the relationships are known the MCMC approach yield is simply a straightforward REML analysis.

With inferred relationship information: The structure of the sample affects the accuracy of parameter estimation. In samples with comparatively large numbers of full-sibs, bias and MSE tend to be much smaller, regardless of which approach to variance component estimation is used, and taking into account the different MSE of the REML derived estimates (for example see E versus B in Table 5.6). Information about the additive genetic variance therefore mainly comes from the full-sib relationships. As in the analyses of situations described in previous chapters, the likelihood approach gives heritability estimates that tend to have larger bias than estimates made using the regression approach. However the MSE of the likelihood estimates are much closer to the MSE of the REML estimates. With low levels of marker information, the MCMC approach gives estimates of heritability that are

biased downwards. However, with larger amounts of marker information, estimates show comparatively little bias and have the smallest MSE.

Set	REML	Regression		Likelihood		MCMC
		Known	Inferred	Known	Inferred	
A	-0.026	-0.021	-0.006	-0.016	0.031	-0.241
20 \ 2 \ 5 \ 5	(0.023)	(0.031)	(0.126)	(0.027)	(0.077)	(0.086)
B	-0.010	0.004	-0.007	-0.009	0.016	0.005
20 \ 2 \ 5 \ 10	(0.030)	(0.041)	(0.075)	(0.036)	(0.045)	(0.036)
C	-0.011	-0.007	0.003	-0.003	-0.002	0.016
20 \ 2 \ 5 \ 20	(0.029)	(0.042)	(0.081)	(0.036)	(0.036)	(0.036)
D	-0.014	0.013	-0.018	-0.016	0.129	-0.346
20 \ 5 \ 2 \ 5	(0.034)	(0.053)	(0.488)	(0.043)	(0.132)	(0.147)
E	0.006	0.014	0.017	0.012	0.083	-0.122
20 \ 5 \ 2 \ 10	(0.043)	(0.060)	(0.258)	(0.052)	(0.154)	(0.052)
F	-0.066	-0.069	-0.060	-0.063	0.024	-0.084
20 \ 5 \ 2 \ 20	(0.032)	(0.042)	(0.122)	(0.037)	(0.058)	(0.036)
G	-0.001	0.044	0.051	0.025	0.061	-0.018
10 \ 4 \ 5 \ 10	(0.023)	(0.051)	(0.088)	(0.037)	(0.049)	(0.025)
H	0.001	-0.010	-0.062	-0.030	0.001	-0.007
10 \ 5 \ 4 \ 10	(0.041)	(0.060)	(0.069)	(0.051)	(0.071)	(0.048)

Table 5.6: Summary of bias (upper line) and MSE (bracketed - lower line) of the heritability estimates obtained from different sample structures. Summary of structure of sets A to H (Table 5.4) are in the column headed 'Set' in form: sire \ dams within sire \ progeny within dam \ locus number. Known – Actual relationships used in the analysis. Inferred – marker-based relationship information used in analysis.

5.4.3.3 With known maternal information.

With known maternal information, the source of one of the alleles in each individual is known (except in the case where mother and offspring are heterozygote for the same alleles) and so parameter estimation is improved, since there is less noise in the information inferred about the paternal sib-ships. In the likelihood approach, there is indication of a slight downward bias with known maternal information (Set C, Table 5.7). However MSE is generally smaller than in the analysis performed without maternal information and when the correct assumption is made about sample structure. Improvement in parameter estimation is most notable in populations that

contain only half-sibs (Set J, Table 5.7). This is due to the comparatively poorer ability to resolve information on half-sib relationships than full-sib relationships with lower amounts of marker data. When maternal information was included in the MCMC approach, parameter estimates were virtually identical to the REML based results, but with slight downwards bias and only a small increase in MSE. Again the most improved estimates noted were those for the half-sib-populations with 5 loci, each with 5 alleles when there was about a 50% decrease in MSE (Set J, Table 5.7).

		Likelihood		MCMC	
Set	REML	Inferred *	Mother	Inferred *	Mother
A	-0.026	0.031	-0.069	-0.241	-0.043
20 \ 2 \ 5 \ 5	(0.023)	(0.077)	(0.038)	(0.086)	(0.028)
B	-0.010	0.016	-0.058	0.005	-0.015
20 \ 2 \ 5 \ 10	(0.030)	(0.045)	(0.046)	(0.036)	(0.032)
C	-0.011	-0.002	-0.074	0.016	-0.012
20 \ 2 \ 5 \ 20	(0.029)	(0.036)	(0.049)	(0.036)	(0.030)
D	-0.014	0.129	-0.066	-0.346	-0.023
20 \ 5 \ 2 \ 5	(0.034)	(0.132)	(0.048)	(0.147)	(0.038)
E	0.006	0.083	-0.085	-0.122	-0.022
20 \ 5 \ 2 \ 10	(0.043)	(0.154)	(0.052)	(0.052)	(0.045)
F	-0.066	0.024	-0.064	-0.084	-0.063
20 \ 5 \ 2 \ 20	(0.032)	(0.058)	(0.055)	(0.036)	(0.033)
G	-0.001	0.061	-0.060	-0.018	-0.005
10 \ 4 \ 5 \ 10	(0.023)	(0.049)	(0.041)	(0.025)	(0.026)
H	0.001	0.001	-0.016	-0.007	0.004
10 \ 5 \ 4 \ 10	(0.041)	(0.071)	(0.056)	(0048)	(0.043)
I	0.007	0.039	0.007	0.006	0.008
40 \ 1 \ 5 \ 10	(0.022)	(0.030)	(0.022)	(0.023)	(0.022)
J	-0.038	-0.338	-0.143	-0.430	-0.043
40 \ 5 \ 1 \ 5	(0.066)	(1.454)	(0.254)	(0.199)	(0.081)
K	-0.005	0.073	0.073	-0.192	-0.010
40 \ 5 \ 1 \ 10	(0.063)	(0.442)	(0.139)	(0.087)	(0.070)
L	0.041	0.053	0.042	-0.017	0.042
40 \ 5 \ 1 \ 20	(0.064)	(0.247)	(0.075)	(0.079)	(0.066)

* Set I analysed using full-sib form of the approach and sets K–L analysed using half-sib form. Rest analysed using hierarchical forms.

Table 5.7: Summary of bias (upper line) and MSE (bracketed - lower line) of the heritability estimates obtained from different sample structures when maternal data was included in the analysis. Inferred – only marker-based relationship information used in analysis. Mother – Maternal and marker data used in analysis.

5.4.3.4 With incorrect assumptions.

The MCMC and likelihood approaches require that prior assumptions are made, or that prior knowledge is available about the population structure before analysis. Exact prior information was used in the simulations to describe population structure in the likelihood approach. Likelihood analysis therefore gave undefined parameter estimates when the assumption of only full-sibs was used on a sample containing only half-sib families and when the assumption of only half-sibs was used in samples containing only full-sib families (Table 5.8).

In samples that are genuinely hierarchical in structure, heritability estimates tend to have large upwards bias if either the assumption of only half-sib or only full-sib is applied (Table 5.8). In the full-sib case this is because half-sibs have a higher likelihood of being classed as unrelated than as full-sib (due to the prior probabilities). Estimates of the variance of the phenotypic difference of unrelated pairs are therefore biased downwards, since the phenotypic difference is distributed $N(0,\ 1.5\sigma_A^2 + 2\sigma_E^2)$ for half-sibs and $N(0,\ 2\sigma_A^2 + 2\sigma_E^2)$ for unrelated (Table 5.2). The information on the value of σ_A^2 is obtained from the difference between the distribution of phenotypic difference for full-sib pairs and for the unrelated pairs. Thus estimates of σ_A^2 are biased upwards and so are estimates of heritability. In the half-sib case the upward bias is more easily explained and is because of the high likelihood that full-sib pairs are classed as half-sib. A larger degree of phenotypic similarity is therefore assigned to a smaller relationship, therefore biasing heritability estimates upwards. Bias increases with the inaccuracy of the prior assumption. For example, compare the bias of sets, A, B and C against D, E and F for both the half-sib and the full-sib assumption (Table 5.8). In sets A to C there are larger number of full-sibs and bias is smaller when a full-sib only assumption is made than in sets D to F where there are fewer full-sibs. The opposite is true when a half-sib only assumption is made, with sets D to F having smaller bias than sets A to C. MSE is also much larger when incorrect assumptions are made. This can be attributed in a number of cases to the large bias.

Set	REML	Likelihood			MCMC		
		Half-sib	Full-sib	Both	Half-sib	Full-sib	Both
A	-0.026	0.206	0.500	**0.031**	-0.379	-0.179	**-0.241**
20\2\5\5	(0.023)	(0.537)	(0.285)	**(0.077)**	(0.163)	(0.052)	**(0.086)**
B	-0.010	0.472	0.348	**0.016**	-0.129	-0.020	**0.005**
20\2\5\10	(0.030)	(0.462)	(0.163)	**(0.045)**	(0.085)	(0.029)	**(0.036)**
C	-0.011	0.371	0.163	**-0.002**	0.181	-0.007	**0.016**
20\2\5\20	(0.029)	(0.260)	(0.063)	**(0.036)**	(0115)	(0.028)	**(0.036)**
D	-0.014	-0.207	0.719	**0.129**	-0.425	-0.407	**-0.346**
20\5\2\5	(0.034)	(0.230)	(0.585)	**(0.132)**	(0.197)	(0.176)	**(0.147)**
E	0.006	0.134	0.784	**0.083**	-0.268	-0.247	**-0.122**
20\5\2\10	(0.043)	(0.168)	(0.680)	**(0.154)**	(0.111)	(0.090)	**(0.052)**
F	-0.066	0.125	0.624	**0.024**	-0.028	-0.136	**-0.084**
20\5\2\20	(0.032)	(0.129)	(0.674)	**(0.058)**	(0.070)	(0.053)	**(0.036)**
G	-0.001	0.378	0.461	**0.061**	0.122	-0.015	**-0.018**
10\4\5\10	(0.023)	0.271)	(0.232)	**(0.049)**	(0.088)	(0.023)	**(0.025)**
H	0.001	0.184	0.443	**0.001**	-0.022	-0.041	**-0.007**
10\5\4\10	(0.041)	(0.186)	(0.274)	**(0.071)**	(0.070)	(0.042)	**(0048)**
I	0.007	***	**0.039**	0.039	-0.241	**0.006**	0.023
40\1\5\10	(0.022)		**(0.030)**	(0.030)	(0.108)	**(0.023)**	(0.030)
J	-0.038	**-0.338**	***	-0.338	**-0.430**	-0.453	-0.457
40\5\1\5	(0.066)	**(1.454)**		(1.454)	**(0.199)**	(0.225)	(0.384)
K	-0.005	**0.073**	***	0.073	**-0.192**	-0.431	-0.285
40\5\1\10	(0.063)	**(0.442)**		(0.442)	**(0.087)**	(0.195)	(0.170)
L	0.041	**0.053**	***	0.053	**-0.017**	-0.425	0.045
40\5\1\20	(0.064)	**(0.247)**		(0.247)	**(0.079)**	(0.199)	(0.146)

*** Undefined result.

Table 5.8: Summary of the bias (upper line) and MSE (bracketed – lower line) of heritability estimates for the different sample structures when inferred relationship data is used. Bold figures indicate marker-based analysis using the correct prior assumptions about the population structure.

With the MCMC approach, bias tends to be downwards, and is due to the poorer ability to resolve sib-ships when there is a mixture of relationships. Visual examination of the reconstructed sib-ships indicates that a large number of incorrect sib-ships are reconstructed at low levels of marker information, which as explained previously results in downwards bias. Under the assumption of only full-sib families, however, heritability estimates improve when there is a comparatively large number of full-sibs in the sample and when marker information is increased to 20 loci (e.g. Set C, Table 5.8). This is because the majority of the information about heritability

comes from full-sib groups, which are accurately reconstructed at higher levels of marker information. Under an assumption of half-sibs only, sib-ship reconstruction groups full-sibs more readily than half-sibs, and assigns them as half-sib groups. Therefore as marker information increases the downward bias from the incorrect sib-ships is cancelled due to the assignment of full-sibs as half-sibs, and in some cases the bias becomes positive (Sets C and G, Table 5.8). When a hierarchical structure is assumed for samples that contain only full-sibs, subsequent parameter estimation is good, and is comparable to estimates made using the correct assumption. With a hierarchical assumption and half-sib samples, there is a slight increase in downwards bias at lower levels of marker information, and a slight upward bias at higher amounts. In all cases of the MCMC approach, the size of the MSE tends to reflect the size of the bias, and when bias is small MSE is very close to the MSE of the REML estimates.

5.4.4 Simulations including an environmental covariance of full-sibs.

5.4.4.1 Population structure and marker data.

Known relationships were initially used in the likelihood approach to investigate the effects of using pair-wise analysis for the estimation of σ_C^2 (due to maternal effects and common environment). Estimates made of the heritability and σ_C^2 showed similar bias and increased MSE when compared against REML derived estimates (Table 5.9).

Estimates of the heritability tended to show greater bias than estimates of the covariance of full-sibs, and larger MSE (Table 5.9). When the environmental covariance of full-sibs is included in the model describing phenotype, the information for estimation of the heritability comes from the half-sib relationships only, while the information on the covariance comes from the full-sib groups versus the half-sibs. Since the resolution of half-sib relationships is poorer than the resolution of the full-sib relationships, poorer estimation properties of the heritability

might be expected. Moreover, inaccuracies in the additive genetic variance are magnified due to the factor of four needed to estimate the component from the unrelated and half-sib distributions.

The difference in the ability to resolve different relationships is most notable in the MCMC approach, where inspection of the reconstructed pedigrees showed that half-sib groups were often split into their component full-sib families. With larger amounts of marker information and larger numbers of half-sibs, parameter estimates based upon the MCMC approach were close to the REML derived estimates (e.g. Sets F and H, Table 5.9). This was because, with higher levels of marker information, both half and full-sib groups are reasonably accurately reconstructed. With the same populations the likelihood approach still showed large MSE for the estimates of the heritability. With lower numbers of half-sibs MCMC-based estimates of the heritability are positively biased and estimates of the covariance of full-sibs are negatively biased. The likelihood approach was very dependent upon the amount of marker information and exhibited large bias and large MSE with low amounts of marker information (e.g. Set A, Table 5.6).

5.4.4.2 The effect of the magnitude of the covariance of full-sibs simulated.

The bias of estimates of the covariance of full-sibs, obtained using the likelihood approach, becomes more negative as the simulated value of the covariance increases Bias is about –0.02 with a simulated covariance of 0.2, but was –0.17 with a simulated covariance of 1. The bias of the heritability shows the opposite trend, and becomes smaller as the simulated value for the covariance of full-sibs increases. Heritability has a bias of about –0.05 with a simulated covariance of 0.2 and is unbiased with a simulated covariance of 1.

Set	REML		Likelihood				MCMC	
			Known		Inferred			
	$\hat{h}^2$	$\hat{\sigma}_C^2$	$\hat{h}^2$	$\hat{\sigma}_C^2$	$\hat{h}^2$	$\hat{\sigma}_C^2$	$\hat{h}^2$	$\hat{\sigma}_C^2$
A	0.072	-0.009	0.023	-0.001	-0.889	0.420	0.005	-0.119
20 \ 2 \ 5 \ 5	(0.105)	(0.038)	(0.185)	(0.051)	(6.437)	(2.763)	(0.053)	(0.032)
B	0.001	0.011	-0.089	0.054	0.089	-0.024	0.213	-0.091
20 \ 2 \ 5 \ 10	(0.081)	(0.023)	(0.221)	(0.068)	(1.528)	(0.418)	(0.127)	(0.031)
C	0.119	-0.051	0.214	-0.058	0.214	-0.102	0.313	-0.142
20 \ 2 \ 5 \ 20	(0.124)	(0.027)	(0.248)	(0.055)	(0.490)	(0.110)	(0.192)	(0.039)
D	0.035	0.007	-0.173	0.010	-0.173	0.131	-0.149	-0.189
20 \ 5 \ 2 \ 5	(0.050)	(0.019)	(0.061)	(0.023)	(1.586)	(0.865)	(0.041)	(0.043)
E	-0.026	0.010	0.177	0.014	0.177	-0.120	-0.014	-0.130
20 \ 5 \ 2 \ 10	(0.042)	(0.017)	(0.050)	(0.018)	(0.734)	(0.402)	(0.047)	(0.029)
F	0.008	0.001	0.080	0.005	0.080	-0.033	0.046	-0.024
20 \ 5 \ 2 \ 20	(0.055)	(0.024)	(0.067)	(0.027)	(0.209)	(0.093)	(0.057)	(0.025)
G	0.057	-0.009	0.061	-0.006	0.061	0.008	0.176	-0.040
10 \ 4 \ 5 \ 10	(0.060)	(0.019)	(0.081)	(0.023)	(0.303)	(0.093)	(0.090)	(0.056)
H	0.002	-0.014	-0.082	-0.007	-0.082	0.041	0.017	-0.028
10 \ 5 \ 4 \ 10	(0.040)	(0.014)	(0.053)	(0.015)	(0.412)	(0.120)	(0.040)	(0.024)

Table 5.9: Summary of the bias and MSE of the heritability ($\hat{h}^2$) and environment covariance of full-sib ($\hat{\sigma}_C^2$) estimates for eight different sample structures (see Table 5.4). Known - Actual relationships were used in the analysis. Inferred – Inferred relationship information used in analysis.

The mean squared error of estimates of the covariance of full-sibs increases as the simulated value of that parameter increases (Figure 5.4). This is partly a scaling effect and partly due to the increasing downwards bias seen in parameter estimates. The mean squared error of the heritability estimates falls as the value of the covariance of full-sibs increases due to the effect of scaling (the additive genetic and environmental variances remained constant during the simulations).

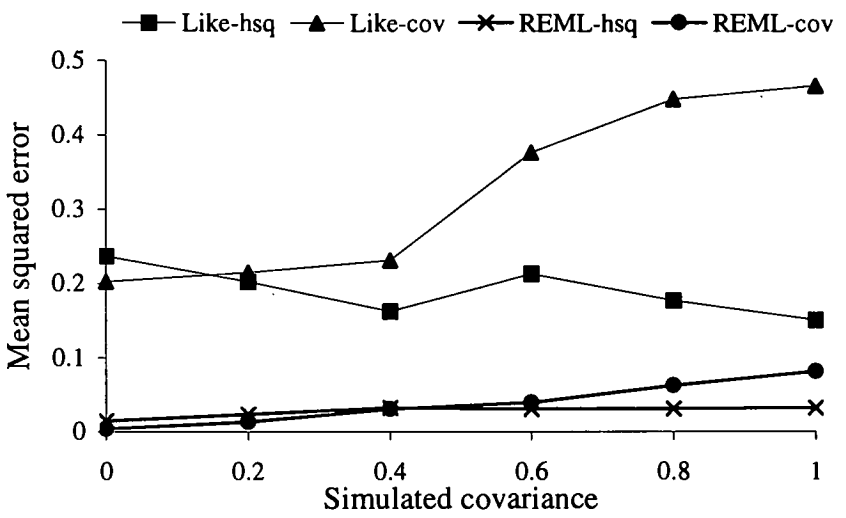

Figure 5.4: The mean squared error of the heritability and covariance of full-sib estimates obtained using the likelihood approach and REML techniques. Simulation conditions: Additive genetic variance = 0.25, environment variance 1, population structure 4 progeny within 4 dams within 40 sires (total size 320), 10 loci with 5 equally frequent alleles at each.

5.5 Discussion.

In this chapter the tolerance of the three different approaches was investigated in situations where the sample structure, marker information and phenotype model deviated from the idealised populations simulated in previous chapters.

The regression technique requires the least amount of prior information be available. The cost of this generality is large MSE in parameter estimates, indicating unreliability in the results. Samples require large numbers of related pairs, of a high

degree of relationship, before estimates become reliable. In addition, since the approach works through the estimation of the actual variance of the relationship, it is difficult to incorporate known information, such as maternal information or known relationships since this requires that related animals be regarded as a unit. Estimates of relatedness between animals would therefore be between groups and individuals and would require the introduction of parameters describing higher groups of relatives (e.g. triplets). Moreover, similar parameters describing the distribution of triplet-wise phenotypic similarity would be required. Such higher order parameters (describing higher order moments of the relatedness and phenotypic similarity) are difficult to interpret (Chapter 2).

Ritland (1996b) showed how the regression-based approach could be extended to estimate genetic variance due to dominant genetic effects. This involved estimates of four-gene coefficients of relatedness, where all four of the alleles at a locus in two individuals are considered at the same time rather than just two of the alleles. Since only half-sib and full-sib families were simulated in this study, estimates of the environmental covariance of full-sibs are entirely confounded with estimates of the dominance genetic variance. Ritland's approach to estimating dominance variance could therefore be used on the simulated samples studied here to estimate σ_C^2. Estimates made were very poor, showing extremely large MSE (often greater than one). Estimation of additional variance components using the regression approach requires considerably more marker information than simulated here before they become reliable.

In balanced hierarchical situations when relationships are known, the likelihood approach yields the REML estimates for the variance components, when the σ_A^2, σ_C^2 and σ_E^2 are partitioned from the total phenotypic variance. This is because σ_C^2 is estimated using the information contained in the full-sib relationship class and σ_A^2 is estimated using the information contained in the half-sib class.

The likelihood approach is less general since it requires that information on the sample structure be available prior to analysis. If incorrect prior information is used, resulting estimates are often biased, although bias decreases with increasing accuracy of the prior information. The MSE of the estimates reflected the size of the

bias. Simulations indicated that a larger amount of marker information and samples containing large numbers of related individuals (both half-sib and full-sib) are needed before both maternal effect and heritability can be estimated accurately. Analysis using the actual relationships showed that although the likelihood approach weights information from different relationships less efficiently than REML techniques, weighting is better than with the regression technique (Table 5.5).

The MCMC approach used in this study was the most basic form of the approach, using an uninformative distribution to describe the sib-ship sizes and not re-estimating population allele frequencies during reconstruction. Even so, estimates were often better than those from the likelihood technique, presumably through the more efficient weight of half-sib and full-sib data, but showed bias with low levels of marker information. In addition the MCMC approach allows more appropriate weights to be used in unbalanced populations (Chapter 4). When inaccurate assumptions were made about the sample structure, the MCMC gave biased results, unless large numbers of the assumed relationships were actually present in the sample and there was a large amount of marker information. In situations where a more complex hierarchical structure was assumed, when in fact the population contained only one class of sib, the MCMC approach gave similar estimates to analysis assuming the correct structure.

An added appeal of the likelihood and MCMC approaches is the ease with which they can be modified to include further information. For example, when maternal information is available it may be included, thereby improving parameter estimation. In the case of the MCMC approach, the inclusion of maternal information produces results almost identical to the REML-derived results regardless of the actual structure of the sib-ships.

The inclusion of data from different types of marker loci into the analyses has a different effect on the MCMC and likelihood approaches, where relationship classes are specified, than on the regression approach, where no relationship classes are specified. In the approaches specifying relationship classes, the likelihoods of some of those classes can be immediately set to zero based upon the information from a mitochondrial locus. The regression approach does not specify exact relationships and so mitochondrial haplotypes are not informative. This is because

allelic identity at a mitochondrial locus is independent of the additive genetic relationship. Simulations indicated that, given the choice between genotyping an autosomal locus or a mitochondrial locus, the number of alleles does not provide a good indication of information content. The sample structure, however, will affect the comparative level of information that can be obtained from a mitochondrial locus. In the study described here, full-sib families were simulated, and so a mitochondrial locus will be very informative since all members of a full-sibs family have the same mitochondrial haplotype. In other sample structures, a mitochondrial locus will be less informative. The most extreme example of this would be in a paternal half-sib structure, where the presence of a mitochondrial locus would be completely uninformative. Simulations specific to the allele distributions at the loci and the population structure would have to be run to determine which of a mitochondrial or autosomal locus should be typed. This type of question about the design of experiments is discussed further in the general discussion, Chapter 7.

Altering the likelihood equations in the likelihood and MCMC approaches would allow genotyping errors to be accounted for in the analysis. In the MCMC approach, calculation of the likelihood of the genotypes within a putative sib-ship would still be determined by summing over all parental genotypes. But summation would also include the probability that each genotype is the result of mistyping. As this would slow the algorithm considerably, some assumptions restricting the number of mistyped genotypes allowed in a single sib-ship may be required. The likelihood approach could be altered in a similar manner. However, simulations indicate that, provided that the number of mistyped loci is low, then the techniques are able to produce reasonable results without the need for any modification.

In the analysis reported in this chapter several modifications were used that made the MCMC mixing process more efficient and speeded up the time to convergence. Considering individuals sequentially rather than randomly speeds up convergence time since all individuals are mixed with equal frequency, resulting in better mixing of the sample. In addition, a step was added that periodically combined half-sib families together allowing larger steps across the likelihood surface, helping to prevent the chain becoming stranded on a false peak. The same technique could be applied to joining full-sib families within half sib-ships, further improving mixing.

When only some of the maternal genotypes are known, or when maternal genotype is not coupled with maternal identity, inclusion of maternal information in the MCMC approach becomes complex, involving part maternity inference, part sib-ship reconstruction. Mixing of individuals with unknown mothers would be over candidate mothers as well as full and half sib-ships (assuming hierarchical sample structure). Mixing of individuals with known mothers would be only over half sib-ships, with additional individuals (putative full-sibs of the known offspring) currently assigned to that mother moving too. Equations 4.4 and 5.5 may be modified to accommodate known maternal genotypes, with summation occurring over the known maternal genotypes, or over a combination of both known and unknown maternal genotypes. In this study, populations containing half and full-sibs were assumed to be in a hierarchical structure. However in practice, maternal half-sib families may be present in the population making reconstruction using the MCMC approach complex. Furthermore, sib-ships could become very extended, with individuals having half-sibs through both the mother and the father, as well as having full-sibs. If maternal data are known, this is less of a problem since summation would be only over the paternal data, but if maternal data is unknown then mixing would have to be between both fathers and mothers. Calculation of the likelihood of the genotypes seen within a sib-ship would become extremely slow, since summation would be across all possible parental genotypes. Assumptions would have to be made to speed up mixing and calculation. Likelihood calculations could be restricted to the immediate maternal and paternal half-sib families of the candidate individual, rather than including (for example) the maternal half-sib families of one of the candidate individual's paternal half-sibs. However valuable information from excluded genotype patterns might be lost using this assumption.

In the study of more complex sample structures presented in this chapter, the total population size was kept constant and at a rather low size. Several sets of simulations were run with larger total sample sizes, but using the same amount of marker information (10 loci with 5 equally frequent alleles) and the same family structure (e.g. 1 sire mated to two dams, producing 5 progeny). As in the analyses reported in previous chapters, when sample size was increased the estimation of the variance components generally improved. However inspection of the reconstructed

pedigrees in the case of the MCMC approach indicated that there was a slight increase in the proportion of incorrect sib-ships reconstructed. Increasing the number of individuals would eventually require that a greater amount of marker information be recorded to be able to distinguish related individuals from those similar due to chance. In the MCMC approach this would lead to increased bias in estimates of variance components as sample size increases, a trend that was observed in Figure 4.2a-i.

It is clear that the major drawback of all of the marker-based approaches is the need for large samples coupled with large amounts of marker information and large numbers of related individuals. The approaches designed to compensate for a lack of marker information and fewer related individuals in the sample require that assumptions be made about the sample structure and as a result their use in a practical situation is restricted. Moreover the need for assumptions can result in the introduction of bias into the estimates, when the assumptions are incorrect. When prior assumptions are correct, however, parameter estimates are improved. Furthermore, the likelihood and MCMC approaches allow for the simple inclusion of known relationships into the analysis thereby further improving parameter estimation, and so they may find a place in studies where only some proportion of the relationships are unknown.

Chapter 6
Estimates of the heritability of body weight in a feral population of Soay sheep.

6.1 Introduction

In recent years there has been increasing interest in estimating genetic variance components in natural populations, with estimates being used to address questions posed by both evolutionary biologists and conservationists. In evolutionary studies, estimates are important in the understanding of patterns of short-term evolution, the reconstruction of historical patterns of natural selection (Lande, 1979) and the prediction of genetic responses to selection. In addition they provide information on the target of selection in a set of correlated characters (Coyne and Beecham, 1987).

Estimates of variance components help provide information on the number of individuals required in order to maintain a viable population, and so are required for the management of captive populations (Storfer, 1996). Loss of genetic variation is a restricting factor in a species' ability to respond to natural selection, and hence a limitation on its potential to evolve (Lande, 1982; Falconer and Mackay 1996; Lande and Shannon 1996; Mousseau and Roff 1987). Variation is therefore critical for maintenance of species within a changing environment.

Central to the estimation of genetic variance parameters is the covariance of the trait between individuals of known relationship. In unmanaged populations, relationships are seldom known accurately and the estimation of parameters is therefore restricted. However, typing individuals at marker loci allows information to

be inferred about their relationships (Thompson, 1975; Lynch, 1988; Queller and Goodnight, 1989). Two approaches to dealing with marker-based relationship information may be adopted. Marker data may be used to assign specific relationships to the individuals of the sample and hence establish a pedigree. Alternatively the information may be used in a more general manner that avoids the assignment of exact relationships, and hence the need to specify a pedigree.

Two techniques have been introduced that do not require exact pedigrees to be specified: a regression approach (Chapter 2; Ritland, 1996b; Lynch and Walsh, 1998) and a likelihood approach (Chapter 3; Mousseau *et al.*, 1998). The main advantage of these pedigree-free approaches is that noise in the inferred relationship data may be accounted for in the analysis. The regression approach includes relationship information in the form of estimates of pair-wise relatedness. It uses a between and within locus ANOVA to remove the sampling error variance of relatedness estimation within pairs from the total variance of relatedness, thereby providing a 'noise-free' estimate of the actual variance of the relationships within the population for use in subsequent variance component analysis (Chapter 2; Ritland, 1996b). The likelihood approach also works on pairs, and accounts for the sampling error variance of the inferred relationship data by attaching a likelihood to each of a number of relationship classes into which the pair might be assigned (Chapter 3; Mousseau *et al.*, 1998).

The regression approach is the more general of the pedigree-free approaches since it requires no assumptions to be made about the population structure. However it does require that there is a large number of related individuals with a high degree of relationship before variance component estimates become reliable (Chapters 2 and 5). The likelihood approach requires the population structure to be known prior to study. As a result its application in natural situations is limited to situations where such information is available. Alternatively prior probabilities may be inferred from existing knowledge, such as the average life-time reproductive success and age structure of individuals in the study population. Most of the information on the genetic variance components comes from close relatives (e.g. full-sibs, half-sibs and parent-offspring), and accurate prediction of the prior probabilities of these relationship classes is important. If prior information is available, variance

component estimates are much more reliable (Chapters 3 and 5) than those made using the regression approach. The likelihood approach may be extended to incorporate known relationships by simply setting to one the probability a pair falls into a particular relationship class. In addition, knowledge of one relationship can provide information on the likelihood of other relationships. For example, the likelihood that Y sired X is affected by the knowledge that Z is the dam of X, since some of X's alleles may be traced to the mother.

The regression approach has been used previously to determine heritabilities in a wild plant population, *Mimulus guttatus* (Ritland and Ritland, 1996). Resulting estimates were larger than those determined under more controlled conditions. This result was contrary to expectation since, under controlled conditions, environmental variance might be expected to be lower (Coyne and Beecham, 1987; Ritland and Ritland, 1996). However the result may also be a reflection of the large sampling variance associated with this approach (Chapter 2). The likelihood technique was applied to a captive salmon population (*Oncorhynchus tshawytscha*), resulting in estimated heritabilities that were similar to previously derived estimates (Mousseau *et al.*, 1998). However the salmon population was set up under rather specific conditions so that prior information about the population structure could be determined.

Alternatively, marker information can be used to infer relationships, thereby reconstructing a pedigree suitable for use in traditional variance component analysis, e.g. REML techniques (Patterson and Thompson, 1971; Lynch and Walsh, 1998). The Markov-chain Monte Carlo (MCMC) approach (Chapters 4 and 5) is based upon relationship assignment. First, a likely set of sib-ships is reconstructed, and then, under the assumption that the pedigree is correct, REML techniques are used to estimate variance components.

For the purposes of clarification, the MCMC approach to sib-ship reconstruction is also based on likelihood techniques, but it will be referred to here as the MCMC approach and the likelihood-based pair-wise approach as the likelihood approach.

The main advantage of methods that assign relationships is that more traditional techniques of variance component estimation may be used, e.g. REML,

thus family-specific and relationship-specific information may be weighted more efficiently than in pair-wise analysis. In addition, known relationships may easily be incorporated into the analysis. However assignment can lead to large bias in variance component estimation through the assignment of incorrect relationships, especially when the amount of marker information is low and the relationships more distant (Chapter 5; Van Vleck, 1970a,b).

In a recent study, Milner *et al.* (2000) used a pedigree which was determined through field observation of mother-offspring pairs combined with paternity inference using genetic markers, to estimate the heritabilities of several traits in an unmanaged population of Soay Sheep (*Ovis ovaris*). In the study, paternities were determined using CERVUS 1.0 (Marshall *et al.*, 1998; Slate *et al.*, 2000) which attaches confidence values to an assigned paternity. Paternities achieving a confidence of at least 95% were used in variance component analysis (Milner *et al.*, 2000). Variance components were estimated using REML methodology with the data analysed under an 'animal' model (Lynch and Walsh, 1998). It was found that lower values of heritability were estimated when the pedigree was based upon paternity assigned with at least 80% confidence. This observed reduction might have been due to the bias introduced through inaccurate relationship information (chapter 5; Van Vleck, 1970a,b). In a second study, on a red deer (*Cervus elaphus*) population, the same approach to variance component estimation was adopted, although behavioural data was also used to assign paternities (Kruuk *et al.*, 2000).

In this study a Soay sheep data set is analysed using the marker-based systems of variance component estimation. The exact data set used is a modified form of the set used by Milner *et al.* (2000), and comprised animals born between 1995 - 1999 (inclusive), whose maternal identity was known. Body weight is used as an example trait and an attempt is made to address the question "which of the approaches produces a good (reliable) estimate of the heritability?" rather than addressing the question "how heritable is body weight in Soay sheep?". Estimates of the heritability of body weight are made using the pedigree free approaches, with analysis carried out assuming that the maternal identity is either unknown or known. Comparison is made with approaches that do specify a pedigree. Four approaches to the pedigree reconstruction are examined: i). MCMC reconstruction of half-sib

families in the absence of maternal information; ii). Paternity inference as in Milner *et al.* (2000); iii). MCMC reconstruction, including the maternal information; iv). MCMC reconstruction over individuals not assigned a father in the 95% confidence pedigree.

6.2 Materials and Methods

6.2.1 The Soay Sheep Population

Soay sheep were introduced to Hirta, the largest island (638 ha) in the St. Kilda group (57°49'N, 8°34'W), from a neighbouring island in 1932, and since that time have remained unmanaged. Study of the sheep has been restricted to a 170ha area that contains approximately one third of the island's population. Since 1985 over 95% of the animals born within the study area during the April-May lambing have been tagged and sampled for genetic analysis soon after birth (Clutton-Brock *et al.*, 1991, 1992). Behavioural observation of the lambs during this period allowed the identity of the mother, if tagged, to be established. Genotype information for all putative mother-offspring pairs was consistent with the inferred relationship (Pemberton *et al.*, 1999). Each August, over one-half of the population in the study area was caught, allowing body weight measurements to be taken. Further animals were caught during the November rut, when a number of rams immigrate into the study area. As many of these immigrants as possible were tranquillised, tagged and sampled for subsequent paternity analysis.

The study described here used a sub-set of the Soay sheep data comprising 529 animals, born after 1994, with 759 body weight measurements. The data set used was a modified form of the one used by Milner *et al.* (2000), and contained animals caught between 1995 to 1999 inclusive. Animals born before 1995 were excluded, since these had been typed using a different set of markers. In addition only animals with known mothers were included in the sample, regardless of whether the mother's genotype was known. Sampled animals were genotyped at twelve marker loci (Table 6.1). Across all animals only about 2 percent of genotype data was missing. A

pedigree determined using the likelihood-based paternity inference package Cervus 1.0 (Marshall *et al.*, 1998, Slate *et al.*, 2000), with confidence levels set to 95%, was available for the data set (Pemberton *et al.*, 1999). Details of the paternity inference are available from Pemberton *et al.*(1999) and Coltman *et al.* (1999).

Locus Name	Number of alleles	Observed heterozygosity
BM 1314	8	0.80
BM 203	11	0.78
INRA 5	9	0.69
TGLA 13	6	0.74
TGLA 263	7	0.78
DRB 3	8	0.82
OarFCB 304	4	0.60
MAF 35	4	0.57
MAF 45	6	0.74
OarCP 26	5	0.72
OarVH 34	6	0.54
Transferrin	7	0.74

Table 6.1: Marker loci used in the study. Further details can be found in Pemberton *et al.*, 1999.

6.2.2 Statistical Analysis

6.2.2.1 Analysis using pedigrees.

Restricted Maximum Likelihood (REML) techniques (Patterson and Thompson, 1971) were used to estimate variance components using an 'animal' model (Lynch and Walsh, 1998). This analysis was similar to that performed by Milner *et al.* (2000), although here only a single trait, body weight (live weight at catch in August), was examined and the sexes were analysed simultaneously. In addition, the number of fixed effects fitted was reduced and was varied to examine the influence of the model on marker-based analysis.

The fixed effects fitted were sex, age, twin status, year of measurement and day of measurement. The small number of fixed effects included was chosen to reflect those that might conceivably be known in a less well studied population. Age and sex were always fitted in the model, but for each model only two out of twin status, year of measurement, and day of measurement were fitted. Age was fitted as an interaction with sex, and was fitted as either a polynomial or as a categorical variable. When fitted as a categorical variable, age had three classes: lamb, yearling and adult. Models where age was fitted as a categorical variable were included to mimic natural situations where exact ages are unknown, but where some information about age can be inferred from the phenotype of other traits (e.g. teeth).

The phenotypic variance was partitioned into three components: the additive genetic variance (V_A), partitioned using the pedigree data; the specific environmental variance of an individual record (V_{Es}), partitioned using repeated measurements on the same individual; and the general environmental (V_{Eg}) common to all records of an animal (Falconer and Mackay, 1996). All the effects were estimated simultaneously using ASREML (Gilmour *et al.*, 1997), which also provides large sample estimates of the standard errors.

Five pedigrees derived using some form of pedigree reconstruction were analysed within the REML framework:

i). A pedigree based upon half-sib families only, not using known mother-offspring relationships. Half-sib families were reconstructed using the MCMC approach, under the assumption that each individual in the sample was a member of only one half-sib family and that there were no parent-offspring pairs in the sample. This was not a realistic model of the pedigree within the sample, but was included for completeness.

ii). A pedigree based upon sib-ships reconstructed when the maternal data were included.

iii). The 95% confidence pedigree, based upon known mother-offspring pairs and paternity inference (Table 6.2).

iv). The 95% confidence pedigree, but using sib-ship reconstruction on those animals not assigned a sire and thereby attempting to regain sib-ship information lost because the actual father had not been genotyped.

Relationship	Number
Mother-offspring	86
Father-offspring	19
Maternal half-sibs	362
Paternal half-sibs	271
Full-sibs	7
Unrelated	138959

Table 6.2: Summary table of the pair-wise relationships present in the data-set, determined using the 95% confidence pedigree.

v). A pedigree, derived from only the known mother-offspring links and no inferred relationships, was also analysed using the REML framework. The pedigree therefore contained only mother-offpring and maternal half-sib relationships.

For all the MCMC reconstructions uninformative distributions, where any family size was equally likely, were used to describe sib-ship size. Heritability estimates were used as a summary statistic for the variance components, and were calculated as:

$h^2 = V_A/(V_A + V_{Eg} + V_{Es})$.

6.2.2.2 'Pedigree-free' analysis.

In order to obtain a single phenotypic measure for each animal in the sample, the ASREML analysis was repeated, but with the pedigree data, either known or inferred, removed from the model. The fixed effects and the specific environmental variance were therefore estimated prior to analysis using relationship data inferred from marker genotypes. Residual deviations for each animal, equal to the average of the phenotypic values of the repeat readings on the animal after the fixed effects had been removed, were computed in ASREML. The variance among individual deviations is the sum of three components, the additive genetic variance, the common environmental component and the specific environmental variance divided by the number of repeat records for that individual. When averaged over animals the

coefficient of the specific environmental variance is the harmonic mean of the number of records for each individual. The additive genetic variance was then partitioned using the regression and likelihood approaches. The common environmental variance was estimated from the total variance of the residual errors and the estimates of the additive genetic and specific environmental variances..

Lynch and Ritland's (1999) estimator (see Chapter 2) of pair-wise relatedness was used to determine relationship information for inclusion in the regression approach. Simulation studies indicated that the use of this estimator in the regression framework yielded heritability estimates with the lowest mean squared errors (Chapter 2). The likelihood approach used the phenotypic function based on the difference between the phenotype of animals within a pair (see Chapter 3). Four classes of relationship were assumed present for the likelihood approach (Table 6.3). Two sets of prior information were used to describe the distribution of the relationships within the sample: a "flat" set, where 99% of pairs were assumed unrelated, but where the remaining categories were assumed equally frequent, and an "exact" set, where the actual percentage of each type of relationship was determined from the 95% pedigree.

Analysis was repeated using the likelihood approach, but with maternal data assumed known. Where a relationship was known exactly, the probability of that relationship was set to one in the likelihood calculation and that of other relationships set to zero. The likelihood of subsequent relationships for the offspring were updated to incorporate the extra information obtained through the knowledge of the source of one of its alleles. For example, if a mother-offspring pair was known, the parent-offspring relationship category was set to zero for all other comparisons between the offspring and an older female, and a pair already known to be maternal sibs were tested to see if they were half-sibs or full-sibs.

Analysis using the pair-wise frameworks was repeated using the relationships present in the 95% confidence pedigree, but restricting the relationship classes to those of the likelihood approach (since these represented the most common relationships within the sample). Other, more distant, relationships were assumed to be zero.

Pattern		Unrelated	Half sib	Full sib	Parent-offspring
$a_i a_i$	$a_i a_i$	p_i^4	$p_i^3(1+p_i)/2$	$p_i^2(1+p_i)^2/4$	p_i^3
$a_i a_i$	$a_i a_j$	$4p_i^3 p_j$	$p_i^2 p_j(1+2p_i)$	$p_i^2 p_j(1+p_i)$	$p_i^2 p_j$
$a_i a_i$	$a_j a_j$	$2p_i^2 p_j^2$	$p_i^2 p_j^2$	$p_i^2 p_j^2/2$	0
$a_i a_j$	$a_i a_j$	$4p_i^2 p_j^2$	$\frac{p_i p_j}{2}(p_i + p_j + 4p_i p_j)$	$\frac{p_i p_j}{2}(1+p_i + p_j + 2p_i p_j)$	$p_i p_j(p_i + p_j)$
$a_i a_i$	$a_j a_k$	$4p_i^2 p_j p_k$	$2p_i^2 p_j p_k$	$p_i^2 p_j p_k$	0
$a_i a_j$	$a_i a_k$	$8p_i^2 p_j p_k$	$p_i p_j p_k(1+4p_i)$	$p_i p_j p_k(1+2p_i)$	$p_i p_j p_k$
$a_i a_j$	$a_k a_l$	$8p_i p_j p_k p_l$	$4p_i p_j p_k p_l$	$2p_i p_j p_k p_l$	0
Distribution		$N(0, 2\sigma_A^2 + 2\sigma_E^2)$	$N(0, 1.5\sigma_A^2 + 2\sigma_E^2)$	$N(0, \sigma_A^2 + 2\sigma_E^2)$	$N(0, \sigma_A^2 + 2\sigma_E^2)$
'Flat' priors		0.9900	0.0333	0.0333	0.0333
'Exact' priors *		0.9946	0.0045	0.0001	0.0008

* Calculated from table 6.2.

Table 6.3: Summary of the relationship classes used in the likelihood technique showing the likelihoods for the genotype patterns, the distribution of phenotypic difference and the two sets of priors. Alleles are indexed i to l and are mutually exclusive. p_i is the allele frequency of i. $\sigma_E^2 = \sigma_{Eg}^2 + \sigma_{Es}^2/m$, where m is the harmonic mean of the number of phenotypic observations on each of the pair of animals.

Standard errors of the variance components were estimated by bootstrapping, which is a method of numerical resampling (Efron and Tibshirani, 1993; Weir, 1997). Bootstrapping could be done at two levels, either at the level of individuals before pair-wise analysis, or at the level of pairs. Simulations using known relationships analysed with the pair-wise techniques indicated that resampling using pairs greatly underestimates standard errors and that resampling individuals (with pairs made up of the same individual sampled twice excluded) tends to overestimate them (by between 10% to 100%). Standard errors were therefore calculated by resampling individuals and cannot be regarded as reliable. Bootstrap estimates of the standard error of the additive genetic variance were combined with the standard error of the total phenotypic variance estimated by ASREML to estimate the standard error of the heritability, using the approximation:

$$\frac{VAR\left[\hat{\sigma}^2_A/\hat{\sigma}^2_{TOT}\right]}{\left(\hat{\sigma}^2_A/\hat{\sigma}^2_{TOT}\right)^2} \approx \frac{VAR\left[\hat{\sigma}^2_A\right]}{\left(\hat{\sigma}^2_A\right)^2} + \frac{VAR\left[\hat{\sigma}^2_{TOT}\right]}{\left(\hat{\sigma}^2_{TOT}\right)^2} - \frac{2COV\left[\hat{\sigma}^2_A,\hat{\sigma}^2_{TOT}\right]}{\hat{\sigma}^2_A \cdot \hat{\sigma}^2_{TOT}}.$$

6.3 Results

Table 6.4 summarises the results for the analysis. Analyses using the pedigree determined through MCMC reconstruction of half-sibs only, and no known maternal information, resulted in no detectable additive genetic variance, and were excluded from the table. This was due to a large number of incorrectly assigned half-sib relationships causing substantial downward bias. Also, likelihood analyses that used 'flat' priors were excluded from the table since the use of flat priors resulted in all estimates of the additive genetic variance components being negative and thus fixed at the zero boundary of the parameter space. Several non-zero estimates were obtained when known mothers were incorporated into the flat prior analysis, but these were also biased downwards, due to the incorrect prior information. In situations where estimates of the additive genetic variance were fixed at zero, estimates of the genetic variance obtained from bootstrap samples also tended to be

fixed at zero. Meaningful standard errors could not therefore be found in those situations.

Estimates of the heritability obtained using the 95% confidence pedigree information ranged from about 0.2 to 0.4 regardless of the method of analysis (pair-wise or non-pair-wise), with only small deviation when different fixed effects were fitted (Table 6.4). REML-based estimates using this pedigree had smaller standard errors than the pedigree free estimates, reflecting the greater efficiency of REML techniques or poorer estimation of the sampling errors obtained using bootstrap methodology.

Estimates of heritability obtained using either of the pair-wise approaches and only inferred relationship data were unreliable and were very sensitive to the choice of fixed effects fitted (Table 6.4). Calculation of the actual variance of the relationship from the 95% confidence pedigree showed that, given the sample size, the variance of the relationships was low ($\approx$ 0.0005), reflecting a low number of related pairs. The population structure of the Soay sheep therefore makes analysis using just 12 marker loci very unreliable. Low levels of marker information and low relatedness may be partly compensated for through the inclusion of known data into the analysis. When maternal data or the 95% confident pedigree were included into the likelihood analysis estimates were improved, and approached estimates obtained using the 95% confidence pedigree, although with generally larger standard errors (Table 6.4).

Estimates made using REML techniques and the different pedigrees with assigned relationships indicated that the greater the number of assigned relationships there are in the pedigree, the lower the estimate of the heritability. At one extreme, when only inferred relationships were analysed (i.e. when only MCMC reconstruction of half-sibs was used) heritability estimates were either zero, or very low (not shown). At the other extreme, with only known relationships included, heritabilities were estimated as between 0.29-0.39 (Table 6.4). This pattern may be explained by downwards bias in the estimates derived from pedigrees determined using assigned relationships due to incorrectly assigned relationships. It may also be explained by the presence of a maternal effect, which would increase the similarity of sibs. Heritability estimates would therefore be biased upwards. The bias would be

	Pedigree free methods					Pedigree determined			
	Regression		Likelihood				Mother &		95% &
	Inferred	95%	No mother	Mother	95%	Mums only	MCMC	95%	MCMC
Known ages									
1 – year, twin	0.004	0.306	FIXED	0.365	0.417	0.309 *	0.234 *	**0.239** *	0.145
	(0.360)	(0.201)		(0.376)	(0.278)	(0.119)	(0.109)	(0.101)	(0.094)
2 – day, twin	-0.612	0.321	FIXED	0.115	0.405	0.342 *	0.253 *	**0.265** *	0.182
	(0.675)	(0.204)		(0.392)	(0.286)	(0.119)	(0.108)	(0.102)	(0.095)
3 – year, day	0.518	0.375	0.456	0.129	0.411	0.385 *	0.257 *	**0.302** *	0.167
	(0.662)	(0.224)	(0.715)	(0.193)	(0.212)	(0.125)	(0.113)	(0.107)	(0.098)
Inferred ages									
4 – year, twin	0.109	0.351	0.053	0.331	0.372	0.290 *	0.228 *	**0.239** *	0.153
	(0.689)	(0.227)	(0.800)	(0.479)	(0.306)	(0.112)	(0.103)	(0.097)	(0.091)
5 – day, twin	-2.814	0.363	FIXED	0.178	0.308	0.336 *	0.259 *	**0.281** *	0.202 *
	(6.212)	(0.289)		(0.342)	(0.318)	(0.109)	(0.106)	(0.095)	(0.090)
6 – year, day	1.453	0.376	FIXED	0.175	0.199	0.347 *	0.226 *	**0.291** *	0.159
	(0.984)	(0.257)		(0.268)	(0.248)	(0.118)	(0.107)	(0.102)	(0.093)

* Significantly different to zero ($p < 0.05$).

Table 6.4: Summary of the heritability estimates (top line) and standard errors (bracketed values) using the different estimators and under the different models. Model column includes summary of fixed effects fitted. The results for the likelihood analysis are those determined using the 'exact' priors. In analysis where pedigree was determined, REML techniques were used to estimate variance components. 95% indicates relationships taken from 95% confident pedigree (bold values represent traditional route to heritability estimation). No mother – no maternal information used. Mother – maternal information used. Mums only – pedigree based on mother-offspring pairs only. FIXED – standard error not determined as additive genetic variance fixed at lower boundary.

greatest when only mother-offspring relationships are used to form the pedigree, and would decrease as further (e.g. paternal) relationships are included in the analysis. Attempts to fit maternal effects into the model often resulted in REML analysis that failed to converge (mainly in cases where twin was included as a fixed effect). Maternal effects were therefore excluded from the analysis.

Visual comparison of the 95% pedigree and the MCMC approach with known mothers indicated that a number of the same half-sib ships were recovered, although some were specific to the method of reconstruction. Comparison of the MCMC recovered half-sib ships and a pedigree determined using paternity inference set at 80% showed the same pattern, but with more sib-ships in common. Hence greater numbers of inferred relationships were present in the pedigree when information from the 95% confidence pedigree and sib-ship reconstruction was combined than when information from each was analysed on its own. This helps explain why, when either the 95% confidence pedigree or a pedigree based on MCMC sib-ship reconstruction were used, the estimates of heritability were intermediate between estimates based only on mother-offspring links and estimates based on both the 95% confidence pedigree and MCMC reconstruction (Table 6.4).

6.4 Conclusions and Discussion.

The objective of this study was primarily to assess systems that use relationship information inferred from marker data to estimate variance components on an actual population. Two approaches to make use of the marker information were examined, either to gain non-specific relationship data or to specify exact relationships. When all the relationships in the sample were assumed unknown, in neither type of approach were variance components estimated successfully. The estimates of the heritability obtained from the 95% confidence pedigree using the pair-wise frameworks were regarded as the "best" achievable estimates using the pair-wise approaches. Deviations from these values when inferred relationship information was used were a result of the inaccuracies introduced through relationship inference. The regression approach, which operates using non-specific relationship data, gave very unreliable results, which deviated wildly when the fixed effects were changed. Low

amounts of marker data and low numbers of relatives in the sample resulted in poor estimates of the actual variance of the relationship, which was greatly under-estimated (by 100 times). Estimates of the heritability were therefore particularly sensitive to changes to the fixed effects, since small changes to the covariance of pair-wise relatedness and phenotypic similarity would be amplified when divided by the actual variance of the relationship (Equation 2.11).

The likelihood approach, which also does not specify relationships, gave negative estimates of the heritability and so estimates were fixed at the boundary, especially in the situation where the priors were inaccurate. Again this is due to insufficient amounts of marker data to gain useful relationship data, and low numbers of relatives in the sample upon which to partition the variance. The MCMC approach, which does specify relationships, also failed for similar reasons. For these techniques to operate successfully in a natural situation, much greater numbers of relatives are required in the sample as well as a greater amount of marker information. The likelihood and the MCMC approach were extended to include information on known relationships, which allowed more reliable estimates of the variance to be determined, that were less sensitive to the fixed effects present in the model.

In this study, the environmental covariance of full-sibs was not fitted into the model since there are few full-sibs in the Soay sheep population (Table 6.2), although the likelihood approach and the approaches that assign relationships allow for its inclusion (Chapter 5). When a pedigree was used that contained only the known mother-offspring links, estimates of the heritability were larger than when assigned relationships were included. This could be because of bias introduced through the inclusion of inferred relationships in the other pedigrees, or because of the existence of a maternal effect that would inflate the similarity between maternal sibs. When maternal effects were fitted in the model and the 95% confidence pedigree used, no appreciable additive genetic effect was found, presumably because of insufficient paternal links in the pedigree. Milner *et al.*(2000) also ran analysed the data using a model that included maternal effects. They found that heritabilities tended to be lower when maternal effects were included, although in body weight there was no observable change in the heritabilities. The majority of the information

contained within the sample therefore comes from the mother-offspring links; and as a result it is arguable whether any assigned relationships should be included in the pedigree.

A problem with all of these approaches is the calculation of the standard errors of variance component estimates. In this study, bootstrap methodology was used to estimate errors for the pair-wise approaches that did not specify a pedigree. In cases where a pedigree was specified, large sample estimates of the variance of the parameters (from ASREML) were used to calculate the standard errors. Neither of these approaches provided a reliable means of estimating the standard errors. In the case of estimates obtained using ASREML, no account is made of the inaccuracy of the pedigree and so estimates of sampling errors are likely to be underestimates. In the case of the bootstrap-derived estimates, simulated studies of balanced populations with known relationships indicated that the sampling errors were overestimated. Ideally the bootstrap would resample over independent data points, a condition clearly violated when resampling over pairs. The individuals within the sample are not independent either, since they share relationships, and so the conditions for the bootstrap are also violated. As a result, when the level of relatedness in the population increases, the accuracy of standard error estimates decreases while the accuracy of the parameter estimates increases.

In the Soay sheep population sib-ships could be reconstructed via paternity inference. In many cases paternities could be assigned with high certainty, and so would probably lead to the most reliable estimates of variance components. In the absence of information on candidate fathers, a distinct possibility in practice, MCMC reconstruction of half sibs using the known maternal information provides a means to recreate the lost sib-ships. Indeed, a number of the same paternal sib-ships were directly reconstructed using the MCMC approach including maternal data as were indirectly reconstructed through assignment of individuals to the same sire using CERVUS (Marshall *et al.*, 1998; Slate *et al.*, 2000) although a number of the sib-ships determined were specific to each approach. Increasing the number of assigned relationships led to a decrease in the size of the estimated heritability, probably due to an increase in the number of mis-assigned relationships, an effect also noted by Milner *et al.* (2000). Therefore only relationships assigned with a high degree of

confidence should be included in the analysis. Confidence levels may be determined for relationship assignments using sib-ship reconstruction and paternity inference by simulation (chapter 4; Marshall *et al.*, 1998). In the case of sib-ship reconstruction, however, some distribution describing sib-ship sizes is required to simulate the families.

Ideally, techniques that incorporate the confidence level of the unknown relationships into the standard methods of variance component estimation (e.g. REML) are needed, especially in cases where only few relationships can be assigned with high confidence, but many with lower confidence. Conceptually it is simple to write down the likelihood of the variance components given known and unknown relationships, but in practice such an equation is prohibitively difficult to maximise, unless there are very few unknown relationships. Foulley *et al.* (1987, 1990) presented an approach that allowed inclusion of uncertain paternities into a sire evaluation scheme. They attached probabilities to a small number of candidate paternities before maximising the likelihood with respect to the variance components and the assigned paternities using a Bayesian framework. This approach is conceptually very similar to the pair-wise likelihood technique, with relationships maximised through the inclusion of phenotypic information to form posterior probabilities for each relationship, except that candidate paternities for an individual are maximised rather than candidate relationships for a pair. As with the likelihood technique, maximising requires iterative procedures, and so estimates of parameters are fed back into the calculation as part of the posterior probabilities for each relationship, thereby introducing bias into subsequent estimates. In addition, the relationships between sires have to be known. It is unclear how their approach could be adapted to the animal model, since the observations (offspring phenotypes) attached to the sires are uncertain rather than the relationships. However, extension to a more general situation where the observations are attached to both sire and dam with some degree of uncertainty (a modified form of the 'gametic' model (Lynch and Walsh, 1998)) might be possible. The likelihood would be maximised over the sire and dam contribution to the offspring, with breeding values estimated for only the sire and dam rather than for every animal, as in the animal model. This would allow inclusion of the known maternal relationships as well as prior weights attached to

candidate fathers to be included in the analysis. However, some of the data from several generations of the same family would have to be excluded, e.g. a grandmother–mother-offspring group must be reduced to a single parent–offspring pair even if the relationships are known exactly, and a relationship matrix is still required for the sires and dams, unless they are assumed to be unrelated.

Chapter 7
General Discussion.

7.1 Thesis summary.

A number of approaches to estimating the genetic parameters, such as heritability, describing quantitative traits have been investigated in this thesis. The first approach, originally presented by Ritland (1996b), and described here in Chapter 2, is based upon regression theory and partitions the variance using relationship information inferred from pair-wise measures of relatedness. The simulation study of Chapter 2 indicated that the relatedness estimator of Lynch and Ritland (1999) yielded heritability estimates that showed the lowest sampling variance, although the differences between it and the next best relatedness estimator were marginal. The regression approach is the most general of the techniques studied in this thesis. It places no restrictions on the population structure within the sample. Since no prior restrictions or assumptions are required for the approach, estimates of the variance parameters are unbiased. The cost of this generality, however, is that estimates made show large sampling error, indicating unreliability in the parameters estimated. A notable property of the approach is a slower decrease in the sampling variance with increasing sample size than other techniques. This suggests inefficiency in the estimation procedure, possibly the result of using locus specific weights appropriate for unrelated pairs in the calculation of pair-wise relatedness, regardless of the true relationship between the pair.

The second main approach is based upon likelihood techniques (Chapter 3; Mousseau *et al.*, 1998), and expands upon the premise that some knowledge of the population structure is available. The approach works through calculation of the

likelihood that pairs of animals fall into a number of relationship classes based on their observed genotype and phenotype, and the prior knowledge about the population structure. The likelihood of pair-wise phenotype is a function of the variance components to be estimated, and so maximising the likelihood with respect to those parameters allows their estimation. In most situations no closed solution to the likelihood equations is available, and so iterative procedures must be adopted. Hence estimates of the variance components are fed back into the iterative procedure resulting in slightly biased parameter estimates. Simulation, however, showed that the bias was trivial compared to the sampling variance, which was considerably smaller than for the regression estimator. In addition, inclusion of prior information on the population structure compensates in part for having a smaller number of related individuals in the sample. Furthermore, likelihood techniques are easily adapted to include additional information, such as known maternal genotypes and known relationships, thereby further improving the accuracy of estimation. As a final attractive feature, additional variance components can easily be included in the analysis since relationship classes are specified. For example in the study of Chapter 5, the environmental covariance of full-sibs was included in the analysis when both full-sibs and half-sibs were present in the sample.

A disadvantage of both pair-wise approaches is that they weight the information in an inefficient manner. In the pair-wise approaches, family-specific weights are proportional to the number of pairs within which that family is represented, while in efficient estimation procedures, such as restricted maximum likelihood (REML; Lynch and Walsh, 1998), weights are proportional to the information content and size of the family. The study of Chapter 2 examined the problem using measures of relatedness directly in a relationship matrix for inclusion in REML analysis. Results indicated that the noise associated with the measures of the relatedness caused downwards bias in heritability estimates, which decreased when the amount of marker information increased, but increased when the sample size increased. This rather counterintuitive result is explained by examining the level of noise in the relationship matrix. The accuracy of relatedness estimation is independent of sample size (apart from a minor effect through increased accuracy of allele frequency estimation), and so adding a single new individual to a sample of

size n results in a less accurate relationship matrix due to the inclusion of n additional estimated relationships.

In addition to the loss of efficient family-specific weights, pair-wise analysis loses information only present in higher order groups. For example, if three individuals sampled from a single generation have genotypes a_ia_i, a_ja_j and a_ka_k (a_i, a_j and a_k being mutually exclusive alleles) they cannot be full sibs; but with pair-wise analysis such an exclusion is not possible. However triplet-wise analysis showed no observable advantage over pair-wise analysis (Chapter 3), and at lower levels of marker information showed large biases. With larger amounts of marker information estimated variance parameters approached those derived using REML and the known relationships. A possible explanation for the poor behaviour of triplet-wise analysis with low marker information is that the errors associated with assigning triplets into the five relationship categories are larger than the errors associated with assigning pairs into two relationship classes given the same level of marker information.

Alternatively, methods that actually assign relationships may be adopted. A Markov chain Monte Carlo (MCMC) procedure was developed for reconstructing sib-ships based upon their marker data (Chapters 4 and 5). Assigning relationships means that existing techniques for variance component estimation can be adopted e.g. REML, and that some of the efficient family-specific weights lost in pair-wise analysis are regained. In addition, the extension to multivariate analysis of multiple traits or the inclusion of further variance components and known relationships is straightforward. With low levels of marker information or with a small average family size, estimates of the additive genetic variance (or heritability) were biased downwards. This was mainly due to individuals that were genuinely unrelated being assigned a relationship rather than related individuals being assigned no relationship. Simulated studies of half-sib, full-sib and hierarchical population structures indicated that variance components estimated from sib-ships reconstructed with MCMC tended to show lower mean squared errors than other procedures, despite the bias evident at lower levels of marker information.

There are a few basic requirements of all of the approaches. They each require that there be a large number of related individuals within the sample. However, there also must be a spread of relationships, since there must be variation

of relationship upon which to partition the additive variance. In addition there must be sufficient marker information (e.g. 10 polymorphic marker loci) upon which to base relationship inference. The ability to resolve relationships is dependent upon the degree of relationship. For example resolving between unrelated and half-sibs is more difficult than resolving between unrelated and full-sib (Thompson, 1975; Blouin *et al.*, 1996; Lynch and Ritland, 1999). Therefore, when a greater number of relationship classes are included in the analysis, and in the absence of relationships of high degree, a greater amount of marker information is required before heritability estimates become reliable. These points were highlighted by the study estimating the heritability of body weight in a feral population of Soay sheep (*Ovis ovaris*) (Chapter 6). The study population has small, predominantly half-sib families, and as a result the samples studied had a low variance of relationship (Chapter 6, Table 6.2). Thus estimates of heritability obtained were unreliable when only inferred relationship data was used. Estimation was improved, however, when known mother-offspring relationships were included in the analysis. A number of different pedigrees that incorporated known and assigned relationships were investigated. It was noted that estimates of the heritability decreased with increasing numbers of assigned relationships were used (Chapter 6, Table 6.4; Milner *et al.*, 2000). This decrease may be due to bias introduced through the assignment of incorrect relationships, or may be due to upwards bias seen in estimates made using information gained from mother-offspring links only, due to maternal effects. It was suggested that only relationships assigned with a high degree of confidence should be included in variance component analysis. Alternatively procedures that account for uncertainty within just a few of the relationships, need to be developed (see section 7.3)

7.2 Study design.

7.2.1 The problem.

In a natural population the cost of capturing and sampling individuals, and of genotyping marker loci maybe constrains the design of the study. Is it better to capture a large number of individuals and genotype them at a few marker loci, or is it better to capture fewer individuals and genotype them at a larger number of marker loci ? Assuming constant population size, the optimum strategy will be affected by the variance of relatedness in the population. For example, sampling a small number from a population containing a small number of relatives will result in poor estimates of heritability since there will too few related pairs upon which to base estimates. Sampling the same number from a population with a larger number of relatives will result in better estimates of heritability. It is the objective of this section to discuss the problem of experiment design, using results compiled from simulations not previously included in the thesis.

7.2.2 Sample size versus number of loci genotyped .

Populations containing 1000 individuals were simulated using three full-sib family structures: 200 families of size 5, 100 of size 10 and 50 of size 20. Samples of size 100, 200, 400, 500 and 666 were randomly selected from the population and sampled individuals were assumed genotyped at either 20, 10, 5, 4, 3 or 40, 20, 10, 8, 6 loci respectively. The total number of loci genotyped was therefore either 2000 or 4000. Five equally frequent alleles were simulated at each locus and the heritability was set as 0.5. Each set of conditions was replicated 120 times. Heritabilities were estimated using each of the marker-based approaches, and the bias and mean squared errors of estimates calculated.

The simulations showed that for each set of conditions there was an optimum strategy for the collection of field data (Figure 7.1). The optimum varied depending upon which technique was used and upon the structure of the population. Bias was

the controlling factor in determining the optimum strategy when estimates were made from reconstructed sib-ships. In populations with lower full-sib family sizes, the optimum strategy is to select fewer individuals and genotype them at a larger number of loci (Figure 7.1a-i). This is because when a larger number of individuals are sampled there is insufficient marker information to accurately assign relationships and there are few exclusions due to incompatible genotypes within the putative sib-ships. This results in large downward bias being introduced into estimates of the heritability. With larger family sizes, however, there is an increase in the number of exclusions due to inconsistent marker genotypes. The sib-ships are therefore reconstructed more accurately, resulting in a substantial reduction in bias (to near zero) and a decreased mean squared error. The optimum strategy therefore favours a larger sample size than in populations with smaller family size. Increasing the amount of marker information also shifts the optimum to favour the collection of more individuals (Figure 7.1b-i). Again this is mainly due to a reduction in the bias introduced through the incorrect assignment of relationships.

The likelihood approach also favours the sampling of fewer individuals when family sizes are small (Figures 7.1a-i and 7.1b-i). Again, the optimum shifts to the collection of more individuals as simulated families size increases.

The regression approach shows the opposite pattern to the MCMC and likelihood approaches. In populations with a low family size the optimum scheme favours sampling more individuals while with larger family sizes the sampling of fewer individuals is better (Figure 7.1a-iii). The optimum number of individuals sampled increases with an increase in the total amount of genotyping (Figure 7.1b-iii). Bias is close to zero for the regression approach, and so MSE reflects the sampling variance. The sampling variance is large for all of the heritability estimates obtained from low family size structures, even at the optimum point. Under these circumstances the estimation of variance components is unreliable, unless a large sample is collected and a large number of loci are genotyped, or information on the population structure is inferred and the likelihood approach used.

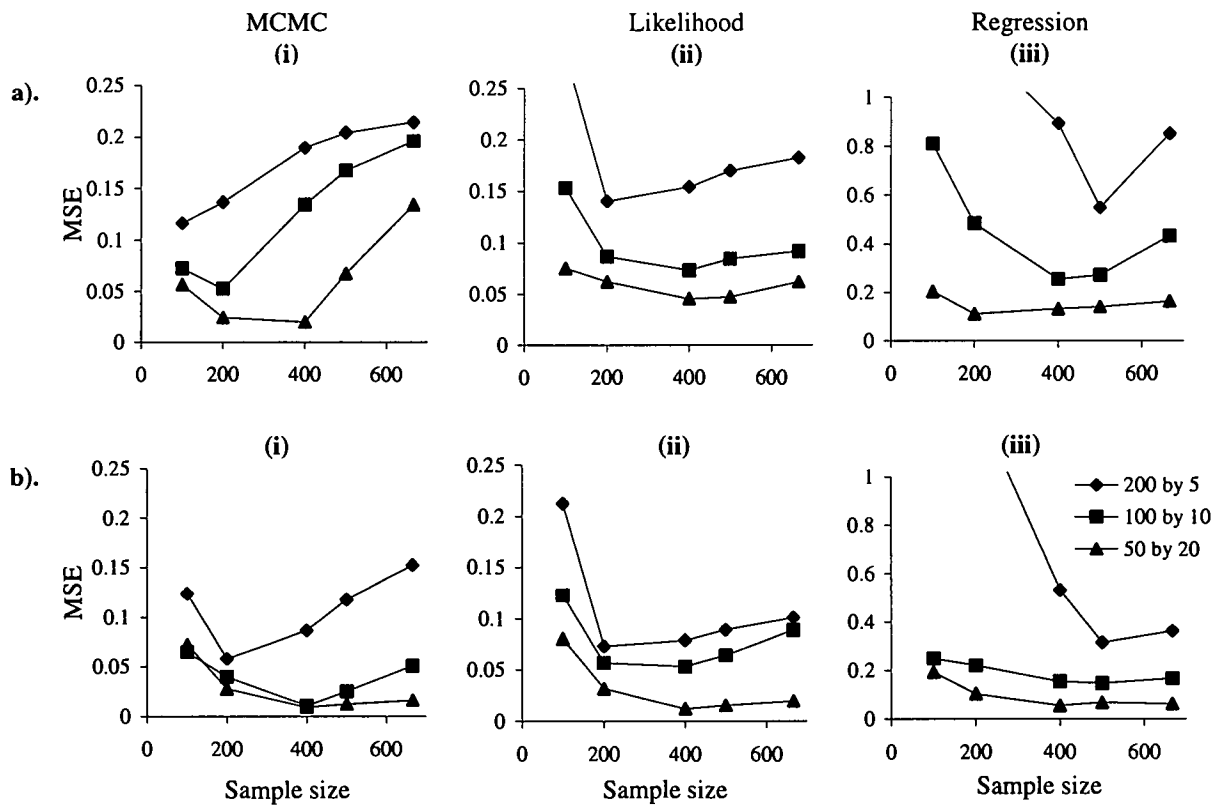

Figure 7.1: The mean squared errors of heritability estimates obtained using (i) the MCMC approach, (ii) the likelihood approach and (iii) the regression approach, when a). 2000 loci and b). 4000 loci in total were genotyped and when three population structures were simulated. Note the change of scale for the regression figures.

7.2.3 Discussion.

The MCMC and likelihood approaches place a greater weight on the accurate estimation of relationship information than the regression approach. If inferred relationship data is inaccurate the likelihood and MCMC techniques yield biased estimates of the heritability (Chapters 3 and 4).

In practice, when information on the structure is unavailable and cannot be accurately inferred, the regression approach must be used. Confirming results presented in Ritland (1996b) the optimum strategy in this situation is to sample as many individuals as possible and genotype them at about 4 to 8 loci. When information on the population structure is available, however, the likelihood and MCMC approaches become preferable and an optimum strategy can be determined using simulation (assuming that allele frequencies are known). In addition, with certain population structures inclusion of a mitochondrial locus in place of a more polymorphic autosomal locus can provide more information on the relationships (Chapter 5). Simulation allows the optimum combination of loci, as well as the optimum sample size to be determined.

The Soay sheep data set analysed in Chapter 6, had an extremely low variance of relationship, which was approximately 0.0005 (calculated from the 95% confidence pedigree). This figure was very low, and resulted in poor estimates of the heritability, when no information on known relationships was included. Low variation in relationship will always restrict analysis, even when the optimum allocation of genotyping is realised. For example at the optimal allocation scheme in Figure 7.1a-iii, for a population comprised of 200 families of size 5 analysed using the regression approach, the MSE was very high (approx. 0.6), indicating that estimates of heritability were not reliable.

Ritland (1996b) discusses additional considerations about the style of sample collection in plant populations, where physical distance has a stronger bearing on the level of the relationship than in animal populations. Sampling adjacent individuals in a plant population as well as more distanced individuals increases the variance of the relatedness within the sample, thereby improving estimates of heritability. However, adjacent individuals will be similar due to the effects of shared environment and so

Ritland (1996b) outlined a partial regression procedure to include the effects of shared environment by including variation due to the physical distance between pairs into the model. Using simulations Ritland found that inclusion of shared environment increased the bootstrap variance of heritability estimates by up to 5 percent. Estimated errors for the effects of physical distance were much smaller than for estimates of the heritability. In some animal populations individuals are less restricted by physical distance, and so increasing the variance of relationship by sampling strategy is not possible. Sub-divided populations also pose a problem for the sampling of individuals, especially if there are few individuals within each group. Allele frequencies within sub-groups may vary, due to increased localised relatedness and drift, making comparison between groups difficult, and introducing additional parameters describing between group variation as well as between family variation. On the other hand, if allele frequencies are assumed constant and the effect of localised environment included, sampling within and between sub-divided populations would increase the variance of the relatedness leading to improved estimates of the heritability. Clearly then, not all natural populations are suitable for marker-based analysis of heritability, and careful examination of each population is required before extensive sampling is undertaken. Ritland (2000) advocates the use of small scale pilot studies of local variation in candidate species, prior to larger scale study in those populations indicating suitable structure for analysis.

7.3 Conclusions and Discussion.

Direct comparison of the techniques is difficult since the likelihood and MCMC approaches require extra information or assumptions about population structure than the regression approach. In situations where prior information on the population structure is unavailable the regression approach must be used, and estimates of heritability are very prone to error due to sampling variance. The MCMC and likelihood approaches show much lower sampling variances, and as a result the extra effort required to determine information on the population structure prior to analysis may be warranted.

In the study presented in this thesis prior information was assumed known for use in the likelihood approach, which is almost certainly not the case in a natural population. Sensible approaches to estimating the prior probabilities from, for example, information collected on the age structure of the population, the average number of births per year and the average litter size are required to allow wide use of the likelihood approach in practice. Then, providing that inaccuracies in the estimated prior information are small and that a large amount of marker information is available, estimates of heritability will be reasonable. Several attempts were made to construct sensible prior probabilities for the Soay sheep data set assuming that no known relationship data was available. Priors were estimated using simulation based around the average number of sheep captured per year, the average population size, the sex-specific age structure of the population and the average life time reproductive success of males (Coltman *et al.*, 1999). In addition it was assumed that each year females had only one offspring (i.e. that the twinning rate was zero). Estimated priors all inflated the number of related individuals in the population (from about 1 percent Table 6.2) to 4 percent), reflecting inaccuracies in the assumptions made about the population dynamics. Inaccurate priors bias subsequent heritability estimates, although bias would decrease with increasing marker information. Further research into this problem is required to determine what type of life history and demographic data may be used to estimate reasonable prior probabilities. Emphasis should be placed on estimating the prior probabilities of close relationships (e.g. parent-offspring, full-sibs and half-sibs) since these categories provide the most information for estimating the heritability.

Including known relationship information (e.g. mother-offspring pairs) into the analysis decreases the sampling variance of estimates of the heritability (Chapters 5 and 6). Improvement is especially marked in the likelihood and MCMC approaches, where the additional information from the maternal genotype allows the source of one of the alleles present in the offspring to be identified (unless both mother and offspring are heterozygous for the same alleles). This knowledge modifies the likelihood of some of the other possible relationships involving the offspring. For example, if a second individual is known to share the same mother then only the likelihood of the offspring and the second individual being half-sib or

full-sib based on their paternal alleles need be calculated. The investigations presented in this thesis mainly examined the situation where none of the actual relationships in the sample were known. The approaches presented are not therefore necessarily the best approaches to use in situations where known and inferred relationship data are available. For example, Foulley *et al.*, (1987, 1990) presented an approach that allowed inclusion of uncertain paternities into a sire evaluation scheme. They used a Bayesian framework to maximise possible paternities with respect to the phenotypic data in an approach conceptually similar to the pair-wise likelihood approach. It is unclear how their approach would be adapted to the animal model, since it is the observations (offspring phenotypes) attached to the sires that are uncertain rather than the relationships. Analysis using a gametic model (Lynch and Walsh, 1998), however, would allow greater generality than under the sire model. The likelihood would be maximised over the sire and dam contribution to the offspring, with breeding values estimated for only the sire and dam rather than for every animal, as in the animal model. Analysis in this manner would also allow inclusion of the known maternal relationships as well as prior weights attached to candidate fathers to be included in the analysis. This approach still requires that the relationships between the mothers as well as fathers are known, unless they are assumed to be zero. In addition it does not provide any straightforward means of incorporating data from multiple overlapping generations.

The study presented in this thesis was primarily concerned with the estimation of additive genetic variance (heritability) in natural populations although the genetic covariance (correlation) between traits is also of interest to evolutionary biologists. For example the genetic covariance might provide information on the target of selection in a suite of correlated characters (Coyne and Beecham, 1987). Ritland (1996b) described how a straightforward extension of the regression approach may be used to estimate the genetic covariance between traits. Of note is the fact that an estimate of the actual variance of relationship is not required in Ritland's method since it cancels out during the calculation. Ritland's approach used the covariances between pair-wise relatedness and pair-wise phenotypic correlation (both between and within traits) to estimate the genetic covariance (Ritland, 1996b). Lynch (2000) showed how this method can be taken a step further, by examining the

pair-wise phenotypic correlations within and between traits directly, rather than their covariances with pair-wise relatedness. This removed the need for marker genotypes entirely, although resulting estimates showed large sampling variances. Lynch's approach may therefore be of theoretical interest only. Mousseau *et* al. (1998) also showed how the likelihood framework could also be extended to estimate genetic correlations. A drawback of these previous approaches is that the heritabilities and correlations for several traits are not estimated simultaneously, resulting in inefficient estimation of parameters.

A more general form of the pair-wise likelihood technique can be formulated, that does allow simultaneous multiple-trait analysis, through alteration of the probability density function describing the likelihood of the phenotypic observations (z_{blr} of equation 3.1). For example in an analysis of k traits the distribution of the pair-wise phenotypes would be described by:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix}, \mathbf{C} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \mathbf{E} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right), \tag{7.1}$$

for an unrelated pair. Where y_i is a 2×1 column vector containing the phenotypic observations of trait i for the pair; μ_i is a 2×1 column vector containing the average phenotypic value of trait i; $\mathbf{C}$ is a k×k additive genetic covariance matrix, with element c_{nm} equal to the additive genetic covariance between characters n and m; $\mathbf{E}$ is a k×k covariance matrix of within-individual environmental effects. The equivalent distribution for a full-sib pair is:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix}, \mathbf{C} \otimes \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} + \mathbf{E} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right). \tag{7.2}$$

Maximising of the likelihood would be with respect to the parameters of **C** and **E**, and would require iterative procedures (e.g. the Newton-Raphson algorithm (Edwards, 1972; Weir, 1996)). Again linear transformation of the pair-wise phenotypic observations for each trait would simplify the likelihood functions.

An interesting feature of multiple-trait analysis is information on the pair-wise similarity of all the traits is used in calculating the likelihood of a pair falling into each of the studied relationship classes. Estimation of parameters might therefore be improved, and the number of marker genotypes required per animal reduced. This is assuming that the bias introduced through inclusion of the phenotype in the likelihood function (Chapter 3) is decreased by studying many traits simultaneously, a possibility if the genetic covariances are low. On the other hand if the genetic covariances are high, bias introduced through inclusion of the phenotypic data might be increased, and more marker information might be required. At the extreme, analysis with no marker information can be imagined, with the phenotypic information on a number of traits to calculate the likelihood of each relationship category, although the approach would still require prior information on the population structure. The results from analyses using no marker information would likely be biased and would almost certainly have large sampling variance.

It is likely that the techniques studies in this thesis will prove more useful in the study of plant populations, where the scale of sampling has a greater effect on the variance of the relationship, than in free roaming animal population. In a recent review, Ritland (2000) described two situations where animal populations might be structured to allow reliable analysis, although in both situations there is a considerable potential for confounding common environmental and genetic effects. The first of these situations, is in populations where breeding groups are comprised of one breeding male and philopatric females. This would increase the variation of relationship between groups, improving heritability estimates. Although the optimum sampling scheme for the number of animals selected within a group versus the number of groups sampled requires study and depends on the variance of relatedness between and within groups. The second situation described is between and within founding groups. When each founding group is small and contains individuals of high relatedness the genetic distance between founding groups is increased.

Therefore optimum sampling within and between founder groups would increase the variance of the relationship within the sample. Further research into the type of population structure appropriate for marker-based analysis of quantitative genetic traits is required.

In conclusion, the marker-based approaches to variance component estimation show promise. This is especially true of the likelihood-based techniques (i.e. both pair-wise likelihood and MCMC approaches) which show considerable ease of expansion to include other types of marker loci and known relationships. In addition, further expansions to include age-specific priors, where the prior probability is affected by the age difference between the animals, or sex-specific information (i.e. separate the father-offspring and mother-offspring categories to allow estimation of maternal effects) are straightforward. The major restriction placed on the techniques is a basic need for large family sizes or, equivalently, a large variance of relationship in the sample. In addition, they require that about ten polymorphic marker loci are typed per individual before estimated heritabilities become reliable, unless known relationships are also included in the analysis. In consequence the techniques will not be appropriate for use on all natural populations of interest.

REFERENCES

ABRAMOWITZ, M. and I.A. STEGUN, 1965. *Handbook of mathematical functions.* AMS 55 U.S. Department of Commerce, Dover edition.

BERNADO, R., 1993 Estimation of coefficient of coancestry using molecular markers in maize. *Theor. Appl. Genet.*, **85**: 1055-1062.

BIRKHEAD, T.R. and A.P. MØLLER, 1992 *Sperm competition in birds: evolutionary causes and consequences.* Academic Press, London.

BLOUIN, M.S., M. PARSONS, V. LACAILLE and S. LOTZ, 1996 Use of microsatellite loci to classify individuals by relatedness. *Mol. Ecol.*, **5**: 393-401.

BOAG, P.T. 1983 The heritability of external morphology in Darwin's ground Finches (*Geospiza*) on Isla Daphne Major, Galapagos. *Evolution*, **37**: 877-894.

BOAG, P.T. and P.R. GRANT, 1978. Heritability of external morphology in Darwin's finches. *Nature*, **274**: 793-794.

BULMER, M.G., 1980 *The Mathematical Theory of Quantitative Genetics.* Oxford University Press, New York.

CLUTTON-BROCK, T. H., O. F. PRICE, S. D. ALBON and P. A. JEWEL, 1991 Persistent instability and population regulation in Soay sheep. *Journal of Animal Ecology.* **60**: 593-608.

CLUTTON-BROCK, T. H., O. F. PRICE, S. D. ALBON and P. A. JEWEL, 1992 Early development and population fluctuations in Soay sheep. *Journal of Animal Ecology.* **61**: 381-396.

COLTMAN, D.W., J.A.SMITH, D.R. BANCROFT, J.PILKINGTON, A.D.C. MACCOLL, T.H. CLUTTON-BROCK and J.M. PEMBERTON, 1999 Density-dependent variation in lifetime breeding success and natural and sexual selection in Soay rams. *American Naturalist*, **154**: 730-746.

COYNE, J.A. and E. BEECHAM, 1987 Heritability of two morphological characters within and among natural populations of *Drosophila melanogaster. Genetics,* **117**: 727-737.

DEKKERS, J.C.M. and J.P. GIBSON, 1998 Applying breeding objectives to dairy cattle improvement. *J. Dairy Sci.*, **81**: 19-35.

DILLON, W.R. and M. GOLDSTEIN, 1984 *Multivariate Analysis Methods and Applications.* John Wiley and Sons, New York.

DHONT, A.A., 1982 Heritability of blue tit tarsus length from normal and cross-fostered broods. *Evolution*, **36**: 418-419.

EDWARDS, A.W.F. 1972. *Likelihood.* Cambridge University Press, Cambridge.

EFRON, B. and R.J. TIBSHIRANI, 1993 *An Introduction to the Bootstrap.* Chapman and Hall, New York.

FALCONER, D.S. and T.F.C. MACKAY, 1996 *Introduction to Quantitative Genetics* (4th edition). Longman, Harlow, Essex.

FJERDINGSTAD, E.J., J.J. BOOMSMA and P. THOREN, 1998 Multiple paternity in the leafcutter and *Atta colombica* – a microsatellite DNA study. *Heredity*, **80**: 118-126.

FOULLEY, J.L., D. GIANOLA and D. PLANCHENAULT, 1987 Sire evaluation with uncertain paternity. *Genet. Sel. Evol.*, **19**: 83-102.

FOULLEY, J.L., R. THOMPSON and D. GIANOLA, 1990 On sire evalution with uncertain paternity. *Genet. Sel. Evol.*, **22**: 373-376.

GILKS, W.R., S. RICHARDSON and D.J. SPIEGELHALTER 1996 *Markov Chain Monte Carlo in Practice.* Chapman and Hall, New York.

GILMOUR, A.R., R. THOMPSON, B.R. CULLIS and S.J. WELHAM, 1997 *ASREML Manual.* New South Wales Department of Agriculture, Orange, 2800, Australia

JACQUARD, A., 1974 *The Genetic Structure of Populations.* Springer-Verlag, New York.

JEWELL, P.A., C. MILNER and J.M. BOYD, 1974 *Island Survivors. The ecology of the Soay sheep of St. Kilda.* The Athlone Press, University of London, London.

KRUUK, L.E.B., T.H. CLUTTON-BROCK, J. SLATE, J.M. PEMBERTON, S. BOTHERSTONE and F.E. GUINESS, 2000 Heritability of fitness in a wild mammel population. *Proc. Nat. Acad. Sci.* **97**: 698-703.

KUHNER, M.K., J.YAMATO and J.FELSENSTEIN, 1995 Estimating effective population-size and mutation-rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140(4)**: 1421-1430

LANDE, R., 1979 Quantitative genetic analysis of multivariate evolution, applied to brain: Body size allometry. *Evolution* 33: 402-416.

LANDE, R., 1982 A quantitative genetic theory of life history evolution. *Ecology* **63**: 607-615.

LANDE, R. and S. SHANNON, 1996 The role of genetic variation in adaptation and population persistence in a changing environment. *Evolution* **50**: 434-437.

LARGET, B. and D.L. SIMON, 1999 Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees. *Mol. Biol. Evol.* **16(6)**: 750-759

LI, C.C, D.E. WEEKS and A. CHAKRAVARI, 1993 Similarity of DNA fingerprints due to chance and relatedness. *Hum Hered.* **43**: 45-52.

LYNCH, M., 1988 Estimation of relatedness by DNA fingerprinting. *Mol. Biol. Evol.* **5(5)**: 584-599

LYNCH, M., 2000 Estimating genetic correlations in natural populations. *Genetical Research,* **74**: (3) 255-264.

LYNCH, M. and K. Ritland, 1999 Estimation of Pairwise Relatedness With Molecular Markers. *Genetics* **152**: 1753-1766.

LYNCH, M. and B. WALSH, 1998 *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Sunderland, MA.

MARSHALL, T.C., J. SLATE, L.E.B. KRUUK and J. M. PEMBERTON, 1998. Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol.*, **7**: 639-655.

MEAGHER, T.R., 1986 Analysis of paternity within a natural population of *Chamaelirium luteum.* 1. Identification of most-likely male parents. *American Naturalist*, **128**: 199-215.

METROPOLIS, N., A.W. ROSENBLUTH, M.N. ROSENBLUTH, A.H. TELLER and E. TELLER 1953 Equations of state calculations by fast computing machine. *J. Chem. Phys.*, **21**: 1087-1091.

MILLIGAN, B.G. and C.K. MCMURRY, 1993 Dominant vs. codominant genetic markers in the estimation of male mating success. *Mol. Ecol.*, **2**: 275-283.

MILNER, J.M., J.M. PEMBERTON, S. BOTHERSTONE and S.D. ALBON, 2000 Estimating variance components and heritabilities in the wild: a case study using the 'animal model' approach. *Journal of evolutionary Biology,* (in press).

MOUSSEAU, T.A. and D.A. ROFF, 1987 Natural selection and the heritability of fitness components. *Heredity* **59**: 181-197.

MOUSSEAU, T.A., K. RITLAND and D.D. HEATH, 1998 A novel method for estimating heritability using molecular markers. *Heredity* **80**: 218-224.

NORRIS, J.R. 1997 *Markov chains.* Cambridge University Press, Cambridge.

PACKER, C., D.A. GILBERT, A.E. PUSEY and S.J. O'BRIEN, 1991 A molecular genetic analysis of kinship and cooperation in African lions. *Nature*, **351**: 562-565.

PAINTER, I., 1997 Sibship reconstruction without parental information. *Journal of Agricultural, Biological, and Environmental Statistics,* **2**: 212-229.

PATTERSON , H. D. and R. THOMPSON, 1971 Recovery of interblock information when block sizes are unequal. *Biometrika* **58**:545-554

PEMBERTON, J.M., D.W. COLTMAN, J.A. SMITH and J.G. PILKINGTON, 1999 Molecular analysis of a promiscuous, fluctuating mating system. *Biological Journal of the Linnean Society*, **68**: 289-301.

PEMBERTON JM, J. SLATE, D.R. BANCROFT and J.A. BARRETT, 1995 Nonamplifying alleles at microsatellite loci – a caution for parentage and population studies. *Mol. Ecol.* **4**: 249-252.

QUELLER, D.C. and K.F. GOODNIGHT, 1989 Estimating relatedness using genetic markers. *Evolution* **43**: 258-275.

RAUW, W.M., E. KANIS, E.N. NOORDHUIZEN-STASSEN and F.J. GROMMERS, 1998 Undesirable side effects of selection for high production efficiency in farm animals: a review. *Livestock Production Science*, **56**: 15-33.

RITLAND, K., 1996a Estimators for pair-wise relatedness and individual inbreeding coefficients. *Genet. Res.* **67**: 175-185.

RITLAND, K., 1996b A marker-based method for inferences about quantitative inheritance in natural populations. *Evolution* **50**: 1062-1073.

RITLAND, K., 2000, Marker-inferred relatedness as a tool for detecting heritability in nature. *Mol. Ecol.* **9**: 1195-1204.

RITLAND, K. and C. RITLAND, 1996 Inferences about quantitative inheritance based upon natural population structure in the common yellow monkeyflower, *Mimulus guttatus. Evolution*, **50**: 1074-1082.

ROFF, D.A. and T.A. MOUSSEAU, 1987 Quantitative genetics and fitness – lessons from Drosophila. *Heredity*, **58**: 103-118.

ROFF, D.A. and A.M. SIMONS, 1997 The quantitative genetics of wing dimorphism under laboratory and 'field' conditions in the cricket *Gryllus pennsylvanicus. Heredity*, **78**: 235-240.

SEARLE, S.R., G. CASELLA and C.E.McCULLOCH, 1992 *Variance Components.* John Wiley and Sons, New York.

SLATE J., T. C. MARSHALL and J. M. PEMBERTON, 2000 A retrospective assessment of the accuracy of the paternity inference program CERVUS. *Mol. Ecol.* **9**: 801-808.

STEARNS, S.C., 1992 *The Evolution of Life Histories.* Oxford University Press, New York.

STORFER, A., 1996 Quantitative genetics: a promising approach for the assessment of genetic variation in endangered species. *Trends Ecol. Evol.* **11**: 343-348.

TAYLOR, A.C., A. HORSUP, C.N. JOHNSON, P. SUNNUCKS and B. SHERWIN, 1997 Relatedness structure detected by microsatellite analysis and attempted pedigree reconstruction in an endangered marsupial, the northern hairy-nosed wombat *Lasiorhinus krefftii*. *Mol. Ecol.*, **6**: 9-19.

THOMPSON E.A., 1975 The estimation of pairwise relationships. *Annals of Human Genetics* **39**: 173-188.

VAN. VLECK, L.D., 1970a Misidentification in estimating the paternal sib correlation. *J. Dairy Sci.*, **53**: 1469-1475.

VAN. VLECK, L.D., 1970b Misidentification and sire evalutaion. *J. Dairy Sci.*, **53**: 1697-1702.

WEIGENSBERG, I. and D.A. ROFF, 1996 Natural heritabilities: can they be reliably estimated in the laboratory? *Evolution*, **50**: 2149-2157.

WEIR, B.S. 1996. *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

YANG, Z., and B. RANNALA, 1997 Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Mol. Biol. Evol.* **14**: 717-724

Appendix 1

Analytical determination of the bias of SUM when relationships are known.

Consider a balanced sample containing f full-sib families with n progeny in each sib-ship. The expected sum of squares within (SSW) families is equal to:

$$E(\mathrm{SSW}) = E\left(\sum_{i=1}^{f}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_{i\cdot}\right)^2\right) = E\left(\sum_{i=1}^{f}\sum_{j=1}^{n} x_{ij}{}^2 - n\sum_{i=1}^{f}\bar{x}_{i\cdot}^2\right) = f(n-1)\sigma_{\mathrm{w}}^2, \qquad \text{(A1.1)}$$

where x_{ij} is the phenotype of animal j in family i, $\bar{x}_{i\cdot}$ is the mean phenotype within family i and σ_{w}^2 is the within family variance. The expected sum of squares between (SSB) families is equal to:

$$E(\mathrm{SSB}) = E\left(\sum_{i=1}^{f}\sum_{j=1}^{n}\left(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot}\right)^2\right) = E\left(n\sum_{i=1}^{f}\bar{x}_{i\cdot}^2 - nf\bar{x}_{\cdot\cdot}^2\right) = (f-1)\left(\sigma_{\mathrm{w}}^2 + n\sigma_{\mathrm{b}}^2\right), \qquad \text{(A1.2)}$$

where $\bar{x}_{\cdot\cdot}$ is the mean phenotype of the sample and σ_{b}^2 is the between family variance. The sample may also be expressed as $p = nf\,(nf - 1)/2$ pairs. The pairs can then be divided into $p_{\mathrm{fs}} = fn\,(n - 1)/2$ full sib pairs, and $p_{\mathrm{ur}} = fn\,(fn - n)/2$ unrelated pairs. The sum of the squares of SUM for full-sib pairs may be written:

$$\sum_{k=1}^{p_{\mathrm{fs}}}\left(x_{k1} + x_{k2} - 2\bar{x}_{\cdot\cdot}\right)^2$$

$$= \sum_{k=1}^{p_{\mathrm{fs}}} x_{k1}^2 + \sum_{k=1}^{p_{\mathrm{fs}}} x_{k2}^2 - 4\bar{x}_{\cdot\cdot}\sum_{k=1}^{p_{\mathrm{fs}}} x_{k1} - 4\bar{x}_{\cdot\cdot}\sum_{k=1}^{p_{\mathrm{fs}}} x_{k2} + 2nf(n-1)\bar{x}_{\cdot\cdot}^2 + 2\sum_{k=1}^{p_{\mathrm{fs}}} x_{k1}x_{k2}$$

$$= (n-1)\sum_{i=1}^{f}\sum_{j=1}^{n} x_{ij}^2 - 2nf(n-1)\bar{x}_{..}^2 + \sum_{i=1}^{f}\left(\sum_{j=1}^{n} x_{ij}\right)^2 - \sum_{i=1}^{f}\sum_{j=1}^{n} x_{ij}^2$$

$$= (n-2)\sum_{i=1}^{f}\sum_{j=1}^{n} x_{ij}^2 - 2fn(n-1)\bar{x}_{..}^2 + n^2\sum_{i=1}^{f}\bar{x}_{i.}^2$$

$$= (n-2)\mathrm{SSW} + (2n-2)\mathrm{SSB}, \qquad \text{(A1.3)}$$

where k indexes the pair, $k1$ is individual one in the pair and $k2$ is individual two. Substituting equations A1.1 and A1.2 into A1.3 and dividing by the number of pairs yields the estimate of the variance of the distribution of SUM for full-sib pairs:

$$\frac{(n-2)f(n-1)\sigma_w^2 + (2n-2)(f-1)\left(\sigma_w^2 + n\sigma_b^2\right)}{0.5nf(n-1)}$$

$$= 2\sigma_w^2 + 4\sigma_b^2 - \frac{4}{nf}\left(\sigma_w^2 + n\sigma_b^2\right), \qquad \text{(A1.4)}$$

but transformation of the multivariate normal distribution describing the pair-wise phenotype of full-sibs suggests that SUM should estimate $3\sigma_A^2 + 2\sigma_E^2 = 2\sigma_w^2 + 4\sigma_b^2$ (Table 3.2). The sum of the squares of SUM for unrelated pairs may be written:

$$\sum_{k=1}^{p_{ur}}(x_{k1} + x_{k2} - 2\bar{x}_{..})^2$$

$$= \sum_{k=1}^{p_{ur}} x_{k1}^2 + \sum_{k=1}^{p_{ur}} x_{k2}^2 - 4\bar{x}_{..}\sum_{k=1}^{p_{ur}} x_{k1} - 4\bar{x}_{..}\sum_{k=1}^{p_{ur}} x_{k2} + 2nf(nf-n)\bar{x}_{..}^2 + 2\sum_{k=1}^{p_{ur}} x_{k1}x_{k2}$$

$$= (nf-n)\sum_{i=1}^{f}\sum_{j=1}^{n} x_{ij}^2 - 2nf(nf-n)\bar{x}_{..} + \left(\sum_{i=1}^{f}\sum_{j=1}^{n} x_{ij}\right)^2 - \sum_{i=1}^{f}\left(\sum_{j=1}^{n} x_{ij}\right)^2$$

$$= (nf-n)\sum_{i=1}^{f}\sum_{j=1}^{n} x_{ij}^2 - nf(nf-2n)\bar{x}_{..}^2 + n^2\sum_{i=1}^{f}\bar{x}_{i.}^2$$

$$= (nf-n)\mathrm{SSW} + (nf-2n)\mathrm{SSB}. \qquad \text{(A1.5)}$$

Substituting equations A1.1 and A1.2 into A1.5 and dividing by the number of pairs yields the estimate of the variance of the distribution of SUM for unrelated pairs:

$$\frac{(nf-n)f(n-1)\sigma_w^2+(nf-2n)(f-1)\left(\sigma_w^2+n\sigma_b^2\right)}{0.5nf(nf-n)}$$

$$=2\sigma_w^2+2\sigma_b^2-\frac{4}{nf}\left(\sigma_w^2+n\sigma_b^2\right), \qquad \text{(A1.6)}$$

but transformation of the multivariate normal distribution describing the pair-wise phenotype of unrelated pairs suggests that SUM should estimate $2\sigma_A^2+2\sigma_E^2=2\sigma_w^2+2\sigma_b^2$ (Table 3.2).

σ_A^2 is equal to $2\sigma_b^2$, which under SUM is correctly estimated as (A1.4) – (A1.6). Therefore estimates of σ_A^2 under SUM are unbiased in a balanced full-sib design when known relationships are used. σ_E^2 is equal to $\sigma_w^2-\sigma_b^2$. But under SUM $\sigma_w^2-\sigma_b^2$ is estimated as 3/2(A1.6) – (A1.4), which is actually equal to $\sigma_w^2-\sigma_b^2-\frac{2}{nf}\left(\sigma_w^2+n\sigma_b^2\right)$. σ_E^2 is therefore underestimated by $\frac{2}{nf}\left(\sigma_w^2+n\sigma_b^2\right)$ and subsequent estimates of h^2 overestimated.

The reason for the discrepancy between the transformation and the analytical result is because in the transformation the population mean is assumed known without error. If the sum of squares of full-sibs and unrelated pairs under SUM are expanded with the actual mean (μ) rather than the sample mean the bias is removed. For example in Full-sibs:

$$\sum_{k=1}^{p_{fs}}(x_{k1}+x_{k2}-2\bar{x}_{..}+2\bar{x}_{..}-\mu)^2$$

$$=\sum_{k=1}^{p_{fs}}(x_{k1}+x_{k2}-2\bar{x}_{..})^2+\sum_{k=1}^{p_{fs}}(2\bar{x}_{..}-2\mu)^2+2(2\bar{x}_{..}-\mu)\sum_{k=1}^{p_{fs}}(x_{k1}+x_{k2}-2\bar{x}_{..}),$$

but $\sum_{k=1}^{p_{fs}}(x_{k1}+x_{k2}-2\bar{x}_{..})=0$, and so

$$\sum_{k=1}^{p_{fs}}(x_{k1}+x_{k2}-2\bar{x}_{..})^2+\sum_{k=1}^{p_{fs}}(2\bar{x}_{..}-2\mu)^2$$

$$=(nf-n)\text{SSW}+(nf-2n)\text{SSB}+2nf(n-1)(\bar{x}_{..}-\mu)^2.$$

Substituting equations A1.1 and A1.2 into the above and dividing by the number of full-sib pairs yields the estimate of the variance of the distribution of SUM for full-sib pairs:

$$\Rightarrow 2\sigma_w^2 + 4\sigma_b^2 - \frac{4}{nf}\left(\sigma_w^2 + n\sigma_b^2\right) + 4(\bar{x}_{..} - \mu)^2$$

But $(\bar{x}_{..} - \mu)^2$ is the sampling variance of the mean, which is also expressed as $\frac{1}{nf}\left(\sigma_w^2 + n\sigma_b^2\right)$ and so the last two terms cancel. Therefore when the actual mean is used in place of the sample mean, estimates of the variance of the distribution of full-sib pairs under SUM are unbiased. The estimates of the variance of the distribution of unrelated pairs under SUM may be shown to be unbiased in a similar manner.

Using the same analytical approach DIFF may be shown to be unbiased and NSUM biased (Equation 3.3).

Appendix 2

Modified forms of the sib-ship likelihood equations.

A2.1 Half-sib families.

A version of equation 5.4 which is faster to compute may be obtained by constraining the possible paternal genotypes. A specific allele is assigned to the sire by selecting an offspring at random from the putative paternal half-sib family and assigning one of its alleles to the sire. Summation of the likelihood of observing all the progeny in that half-sib family is then over all the other alleles that the sire may contain. This is repeated, but assigning the other allele in the offspring to the sire. As in chapter 2, indicator variables are used for ease of expression. S_{xy} represents an indicator variable. $S_{xy} = 1$ when allele x is identical to y and $S_{xy} = 0$ otherwise. If the randomly selected offspring has genotype (w, y) then the likelihood of the genotypes within a half-sib family may be expressed as:

$$L_{\text{genotypes}} = \prod_{\ell} L_{\ell}\,, \tag{A2.1}$$

where

$$L_{\ell} = \sum_{x=1}^{n_a} (2 - S_{wx}) p_w p_x \prod_{c=1}^{n_f} (d), \tag{A2.2}$$

when w is identical to y and

$$L_{\ell} = 0.5 \sum_{x=1}^{n_a} (2 - S_{wx}) p_w p_x \prod_{c=1}^{n_f} (d) + 0.5 \sum_{x=1}^{n_a} (1 - S_{xw})(2 - S_{yx}) p_y p_x \prod_{c=1}^{n_f} (e), \tag{A2.3}$$

when w is different from y. In equations A2.2 and A2.3:

$$d = 0.25\left(S_{wa_{c1}} p_{a_{c2}} + S_{wa_{c2}} p_{a_{c1}} + S_{xa_{c1}} p_{a_{c2}} + S_{xa_{c2}} p_{a_{c1}}\right)$$

$$e = 0.25\left(S_{ya_{c1}} p_{a_{c2}} + S_{ya_{c2}} p_{a_{c1}} + S_{xa_{c2}} p_{a_{c2}} + S_{xa_{c2}} p_{a_{c1}}\right)$$

L_ℓ is the likelihood for an individual locus, indexed by ℓ; n_ℓ is the number of alleles at locus ℓ; x indexes the unconstrained paternal allele; p_z is the allele frequency of z; c indexes an individual from the putative family of size n_f; a_{c1} and a_{c2} are the alleles one and two of individual c. In equations A2.2 and A2.3 terms of the type $(2 - S_{xw})$ convert the ordered genotype frequency of the sire into the unordered frequency. The $(1 - S_{xw})$ term in equation A2.3 prevents the (unordered) paternal genotype (w, y) being considered twice.

A2.2 Full-sib families.

A constrained version of the likelihood of a putative full-sib family may be calculated in a similar way. Here one allele from the randomly selected offspring is used as a constraint on one of the parents, and the other allele as a constraint on the other parent. The genotype of the offspring is again (w, y). The likelihood for a single locus is now:

$$L_\ell = S_{wy}\left[\sum_{x=1}^{n_\ell}\sum_{z=x}^{n_\ell} b\, p_w p_x p_y p_z \prod_{c=1}^{n_f}(d)\right] + (1 - S_{wy})\left[\sum_{x=1}^{n_\ell}\sum_{z=1}^{n_\ell} b\, p_w p_x p_y p_z \prod_{c=1}^{n_f}(d)\right] \qquad \text{(A2.4)}$$

where:

$$b = 8 \times 2^{-(S_{wx} + S_{yz} + S_*)}$$

$$d = 0.25\left(S_{wa_{c1}} S_{ya_{c2}} + S_{wa_{c1}} S_{za_{c2}} + S_{xa_{c1}} S_{ya_{c2}} + S_{xa_{c1}} S_{za_{c2}}\right) + 0.25\left(1 - S_{a_{c1}a_{c2}}\right)\left(S_{wa_{c2}} S_{ya_{c1}} + S_{wa_{c2}} S_{za_{c1}} + S_{xa_{c2}} S_{ya_{c1}} + S_{xa_{c2}} S_{za_{c1}}\right)$$

x and z index the unconstrained parental alleles; b is a term that adjusts the frequency of ordered genotypes to unordered genotypes; n_f is now the size of the full-sib family; S_* is also an indicator variable, with S_*=1 when the unordered genotype of parent 1 is the same as the unordered genotype of parent 2 and S_*=0 otherwise. For example, in the calculation of b: when the parental genotypes are (1, 2) and (3, 4), b = 8; when (1, 1) and (2, 3), b = 4; when (1, 1) and (2, 2), b = 2; and when (1, 2) and (2, 1), b = 4, etc. Multi-locus likelihoods are then calculated using equation A2.1.

A2.3 Hierarchical families.

Similar methods may also be employed to constrain the likelihood calculation for hierarchical families. Consider n_f full sib families nested within the paternal half-sib family, which are indexed by m. Each full-sib family contains n_m progeny which are indexed by c. The first individual from family one is used to constrain the paternal genotype, using each of the offspring's alleles in turn. Likelihoods must be weighted by ½, since either allele in the offspring could have come from the parent. The other offspring allele is used to constrain the maternal genotype of that full-sib family. The maternal genotypes for the remaining full-sib families are constrained by using the first offspring in each of those families. Again, since either allele could come from the mother this must be repeated and the likelihoods scaled by ½. For ease of expression, a_{mci} will denote allele i of individual c of full-sib family m. For example a_{232} is allele 2 of individual 3 of family 2. In each case, calculation of b follows the same from asA2.5. The likelihood for a single locus is:

$$L_\ell = S_{a_{111}a_{112}}\left[\sum_{i=1}^{1}\sum_{x=1}^{n_\ell} p_{a_{11i}} p_x \prod_{m=1}^{n_f}(d)\right] + \left(1 - S_{a_{111}a_{112}}\right)\left[0.5\sum_{i=1}^{2}\sum_{x=1}^{n_\ell} g\ p_{a_{11i}} p_x \prod_{m=1}^{n_f}(d)\right]. \quad \text{(A2.6)}$$

When $i = 1$, $g = 1$ and when $i = 2$, $g = 1 - S_{xa_{111}}$.

When $m = 1$,

$$d = \sum_{y=1}^{n_\ell} b p_{a_{m1j}} p_y \prod_{c=1}^{n_m}(f), \tag{A2.7}$$

where $j = 3 - i$.

When $m \neq 1$,

$$d = S_{a_{m11}a_{m12}}\left[\sum_{j=1}^{1}\sum_{y=1}^{n_\ell} b\, p_{a_{m1j}} p_y \prod_{c=1}^{n_m}(f)\right] + \left(1 - S_{a_{m11}a_{m12}}\right)\left[0.5\sum_{j=1}^{2}\sum_{y=1}^{n_\ell} h b\, p_{a_{m1j}} p_y \prod_{c=1}^{n_m}(f)\right]. \tag{A2.8}$$

When $j = 1$, $h = 1$ and when $j = 2$, $h = 1 - S_{ya_{m11}}$.

For equations A2.7 and A2.8

$$b = 8 \times 2^{-\left(S_{a_{11i}x} + S_{a_{m1j}y} + S_*\right)}$$

$$f = 0.25\left(S_{a_{mc1}a_{11i}} S_{a_{mc2}a_{m1j}} + S_{a_{mc1}a_{11i}} S_{a_{mc2}y} + S_{a_{mc1}x} S_{a_{mc2}a_{m1j}} + S_{a_{mc1}x} S_{a_{mc2}y}\right)$$
$$+ 0.25\left(1 - S_{a_{mc1}a_{mc2}}\right)\left(S_{a_{mc2}a_{11i}} S_{a_{mc1}a_{m1j}} + S_{a_{mc2}a_{11i}} S_{a_{mc1}y} + S_{a_{mc2}x} S_{a_{mc1}a_{m1j}} + S_{a_{mc2}x} S_{a_{mc1}y}\right)$$

S_* is an indicator variable, with S_*=1 when the unordered genotypes of the parents are the same and S_*=0 otherwise.

Multi-locus likelihoods are then calculated using equation A2.1. When $n_f = 1$ the constrained hierarchical equations reduce to the constrained full-sib equations. Likewise, when $n_m = 1$ for all m the constrained hierarchical equations can be shown to reduce to the constrained half-sib equations.

Published Papers

Chapter 3, excluding the triplet-wise analysis, and small parts of chapter 2 have been published as:

THOMAS S.C., J.M. PEMBERTON and W.G. HILL 2000 Estimating variance components in natural populations using inferred relationships. *Heredity* **84**: 427-436

Chapter 4 has been published as:

THOMAS S.C. and W.G. HILL 2000 Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics* **155**: 1961-1972