

Estimating variance components in natural populations using inferred relationships

STUART C. THOMAS*, JOSEPHINE M. PEMBERTON & WILLIAM G. HILL

Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, U.K.

Until recently, the estimation of the heritability of a trait has required knowledge of the pedigree within a population. In natural populations such knowledge is often unknown. Two techniques have been developed which use marker information to estimate heritabilities without reference to the exact nature of the relationships: a regression-based estimator that regresses phenotypic similarity for a pair of individuals against an estimate of their relationship and a likelihood-based estimator that maximizes the probability of the genotypic and phenotypic data given a known population structure. Computer simulation was used to compare the behaviour of these estimators. Bias in estimates of heritability decreased with increasing marker information, decreasing simulated heritability, increasing relatedness and increasing sample size. The techniques displayed reasonable tolerance to the percentage of missing data. The regression-based technique shows least average bias, but largest variance over simulations. Likelihood-based techniques show larger average bias, but smaller variances over estimates. A modified form of the likelihood technique, requiring fewer initial assumptions about population parameters, is presented. The modified form shows less bias in its estimates of heritability than the likelihood technique originally proposed.

Keywords: heritability, likelihood, marker information, pedigree information, regression, relatedness.

Introduction

A primary goal of many genetic studies is the estimation of variance components associated with individual traits and of covariance terms between traits. Heritability, the proportion of variation in a trait that is contributed by average effects of genes, may be calculated from variance components. The heritability of a trait gives an indication of the ability of a population to respond to selection, and thus, the potential of that population to evolve (Lande, 1982; Mousseau & Roff, 1987; Falconer & Mackay, 1996; Lande & Shannon, 1996).

Estimates of variance components and heritability are well documented in the field of animal breeding, where this information is used in the development of selection regimes to improve economically important traits (Falconer & Mackay, 1996; Lynch & Walsh, 1998). A requirement for estimating variance components is knowledge of the relationship structure of the population. Classically, relationships are calculated from

known pedigrees (Jacquard, 1974; Falconer & Mackay, 1996; Lynch & Walsh, 1998).

In natural populations, variance components are also of considerable interest for evolutionary studies (Lande, 1982; Boag, 1983; Lande & Shannon, 1996) and, increasingly, for conservation purposes (Storfer, 1996). In natural populations, however, information on relationships may be unreliable or unavailable. Molecular marker information [restriction fragment length polymorphism (RFLP), minisatellites, microsatellites, RAPDs, etc.] from the population of interest provides a means of circumventing this problem, by allowing estimation of relationships on a pair-wise basis (Thompson, 1975; Ritland, 1996a; Queller & Goodnight, 1989), without the need for pedigree reconstruction. These estimates of relationships may be combined with phenotypic information gathered from the same individuals, allowing inferences to be made about variance components (Ritland, 1996b; Mousseau *et al.*, 1998).

Two methods that allow the estimation of variance components and, more specifically, the heritability of a trait in natural populations without reference to the exact pedigree have previously been described

*Correspondence. E-mail: sthomas@srv0.bio.ed.ac.uk

(Ritland, 1996b; Lynch & Walsh, 1998; Mousseau *et al.*, 1998). Molecular data are used to infer relationships between individuals on a pair-wise basis, because this provides the least complex level at which relationships may be estimated, while still allowing a population to be partitioned into different relationship classes. Estimates of pair-wise relationships are then combined with a pair-wise measure of phenotypic information. This combination may be through regression-based procedures (Ritland, 1996b; Lynch & Walsh, 1998), in which a measure of the pair's phenotypic similarity is regressed against their estimated relationship. Alternatively, likelihood-based procedures (Mousseau *et al.*, 1998) may be used, in which pairs are placed into relationship classes of predetermined structure, according to the probability of observing their genotype and phenotype.

To date there has been no formal comparison of the properties of these techniques. The objectives of this study are: (i) to compare the techniques using a full-sib family structure, to determine biases in estimates of heritability; (ii) to examine the sampling variance of the estimators; and (iii) to present modifications to the maximum likelihood-based procedure that make more efficient use of the data and to include these refinements in the comparison.

Methods

The two techniques are designed for use on a sample where l marker loci are scored for allele type (denoted i) and phenotypic information (y) has been measured on the trait of interest.

Method 1: the regression-based procedure

A number of methods for calculating estimates of pair-wise relatedness are available (Lynch & Ritland, 1999). Here we use the correlation-based procedure derived by Ritland (1996a) which shows similar sampling variances to other techniques under the simulated conditions investigated (Lynch & Ritland, 1999). The correlation procedure equates the observed identity of alleles between a pair with the expected identity calculated using the allele frequencies and the relationship of the pair. Rearrangement then yields an estimator for the relationship. Ritland (1996a) calculated a weighted average over loci and alleles giving an estimator efficient at low levels of relatedness:

$$\hat{r} = \left[\sum_l (n_l - 1) \right]^{-1} \cdot \left[\sum_{jl} (S_{jl} - P_{jl}^2) / P_{jl} \right], \quad (1)$$

where n_l denotes the number of alleles at locus l , S_{jl} is the observed probability of sampling two identical alleles of type j , one from each individual, and P_{jl} is the allele frequency of allele j at locus l , calculated omitting the individuals under consideration.

Ritland's method relies on the estimation of the 'actual' variance of \hat{r} , because mean \hat{r} given by the estimator is close to zero; relationship estimates from this estimator are not bounded between zero (unrelated) and half (identical twin) (Ritland, 1996a). The 'actual' variance is calculated on the assumption that each locus provides an independent estimate of the relationship. An ANOVA is used to eliminate the within-individual variance in relationship from the total variance in relationship, thereby yielding an estimate of the 'actual' variance of relatedness between individuals in the population.

The phenotypic similarity for a trait is calculated for all possible pairs:

$$Z_i = [(y_i - \bar{y})(y'_i - \bar{y})] / \hat{\sigma}^2, \quad (2)$$

where: Z_i is the measure of the phenotypic similarity between pair i ; y_i and y'_i are the trait values for pair i ; \bar{y} is the mean of the trait and $\hat{\sigma}^2$ its variance, each estimated from the sample. Z_i may also be expressed as:

$$Z_i = 2r_i h^2 + e_i, \quad (3)$$

where e_i is the residual error term, and h^2 the heritability. Regression theory then yields an estimator for the trait heritability:

$$\hat{h}^2 = C_{ZR} / 2V_R, \quad (4)$$

where C_{ZR} is the covariance of the relatedness and phenotypic similarity and V_R is the estimate of the 'actual' variance of relatedness for the population.

Method 2: the likelihood-based estimators

Mousseau *et al.* (1998) described a likelihood-based procedure applicable in situations where some prior knowledge of population structure is known. In the case of a population comprising only full-sib families this prior knowledge would be the probability that a pair of individuals randomly selected from the population are full-sibs.

There are seven possible genotype patterns observable at a single locus from a pair of diploid individuals. The likelihood of observing these patterns given the relationship between the pair may be calculated for a given relationship (Thompson, 1975). These patterns are given

Table 1 The relative likelihoods for the seven possible pair-wise genotypes observable in diploid individuals in unrelated, half-sib and full-sib pairs, with i, j, k and l denoting mutually exclusive alleles

Pair-wise genotype	Unrelated	Half-sib	Full-sib
$A_iA_i-A_iA_i$	$4p_i^2$	$2p_i(1 + p_i)$	$(1 + p_i)^2$
$A_iA_i-A_iA_j$	$4p_i$	$1 + 2p_i$	$1 + p_i$
$A_iA_i-A_jA_j$	4	2	1
$A_iA_j-A_iA_j$	$8p_i p_j$	$p_i + p_j + 4p_i p_j$	$1 + p_i + p_j + 2p_i p_j$
$A_iA_i-A_jA_k$	4	2	1
$A_iA_j-A_iA_k$	$8p_i$	$1 + 4p_i$	$1 + 2p_i$
$A_iA_j-A_kA_l$	4	2	1

in Table 1, with examples of the likelihood for each pattern for three different classes of relationship. These likelihood patterns are multiplied across loci to give the total likelihood of the observed molecular data in a pair given a particular relationship.

The phenotypic information also provides information on the likelihood of the relationship, because the distribution of some function of a pair’s phenotypes will be 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:

$$L = \prod_i \left(\sum_r a_r m_{i|r} z_{y_i, y'_{i|r}} \right), \tag{5}$$

where L is the total likelihood for the population, i 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 , $m_{i|r}$ is the likelihood of the molecular data of pair i given relationship r and $z_{y_i, y'_{i|r}}$ is the probability density of the phenotypic data for pair i given relationship r and the population parameters. A function that combines the phenotypic data of a pair is required to allow the probability of the observed phenotypes, given a particular pair-wise relationship, to be calculated.

Assuming that the trait under consideration is normally distributed, then any linear function of a pair’s phenotypes is also normally distributed. A number of different linear functions may be defined, and their expected distributions, in terms of the within-family (V_w) and between-family (V_b) variance, for given relationship classes derived. Three functions, termed methods 2a, 2b and 2c, are defined in Table 2. The variance of the expected distributions is dependent on the relationship between the pair, because the phenotypic correlation between any given pair is dependent only upon their relationship, assuming no environmental correlation.

Each function requires that different numbers of population parameters be estimated prior to the likelihood calculation (Table 2). Method 2a, proposed by Mousseau *et al.* (1998), is equivalent to the sum of the normalized trait values and requires the greatest number of parameter estimates prior to calculation. Methods 2b and 2c use the sum and difference of the observed trait values, respectively, and are modified forms of the likelihood technique that require fewer parameters to be estimated prior to calculation. It is desirable to have fewer parameters requiring estimation prior to calculation as they lead to bias in subsequent estimates of the variance. Method 2a estimates heritability directly, whereas methods 2b and 2c use the formula:

$$\hat{h}^2 = 2\hat{V}_b / (\hat{V}_w + \hat{V}_b). \tag{6}$$

Table 2 Summary table showing the linear functions used in the likelihood procedure, the parameters required prior to estimation and expected distributions of unrelated and full-sib pairs. $N(0, 2 + h^2)$ denotes a normal distribution with mean 0 and variance $2 + h^2$

	Function	Parameters	Distribution (unrelated)	Distribution (full-sib)
Method 2a	$(y_i + y'_i - 2\bar{y})/\hat{\sigma}$	a, \bar{y} and $\hat{\sigma}$	$N(0, 2)$	$N(0, 2 + h^2)$
Method 2b	$y_i + y'_i - 2\bar{y}$	a and \bar{y}	$N(0, 2V_w + 2V_b)$	$N(0, 2V_w + 4V_b)$
Method 2c	$y_i - y'_i$	a	$N(0, 2V_w + 2V_b)$	$N(0, 2V_w)$

Maximization of likelihood eqn (5) 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 2). Further functions may be defined which combine a pair's phenotypic information; for example both phenotype sum and difference can be combined into the one estimator, because these are uncorrelated in both full-sib and unrelated pairs. However, this would require that the same population parameters as method 2b be estimated prior to calculation.

Bias in pair-wise techniques with full pedigree information

Variance components were estimated using correct pedigree information (i.e. known relationships) using the pair-wise techniques and restricted maximum likelihood (REML; Lynch & 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. Balanced and unbalanced family structures were examined.

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 method 1, 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 equally weighted in the calculation. With complete pedigree information the likelihood techniques become equivalent to an ANOVA that uses the mean square of the linear functions to estimate the appropriate sums of the variance components for full-sib and for unrelated pairs (Table 2). This equivalence can be used to show that methods 2a and 2b are biased, whereas method 2c yields the REML-derived estimate.

For method 2b the bias affects only the estimate of V_w , which has an expected value less than the REML-derived estimate of V_w by:

$$(2/nf)(V_w + nV_b), \quad (7)$$

where n is the family size and f is the number of families. Estimates of heritability are therefore upwardly biased

(eqn 6). This bias is a result of using the estimate of the sample mean in the likelihood calculation, and is removed if the actual population mean is known and used.

The estimate of heritability from method 2a is less in expectation than the REML-derived estimate of heritability by:

$$\frac{4\left(\frac{1}{nf} - 1\right)V_w + 2\left(\frac{n}{f} - 2\right)V_b}{(nf - 1)V_w + n(f - 1)V_b} - \frac{2V_b}{V_w + V_b}. \quad (8)$$

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. In methods 2a and 2b the bias decreases with larger sample sizes.

The unbalanced case REML estimates of variance components weight families according to the phenotypic correlation between the offspring rather than the family size (Falconer & Mackay, 1996). Because the weights given to each family depend upon the estimates of the parameters made, REML techniques yield slightly biased estimates of variance components in the unbalanced case.

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

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. This source of bias is difficult to assess analytically and so simulation is required.

Simulations

Phenotypic data for full-sib data sets were generated using the infinitesimal model (Bulmer, 1980). The phenotypic mean was set to 0, with a variance of 1. An individual's phenotype was simulated as:

$$Y_{ij} = [(a_{i1} + a_{i2})/2] + N[0, (\sigma_A^2/2)] + N(0, \sigma_E^2) \quad (9)$$

where Y_{ij} is the phenotypic value of sib j in family i , and a_{i1} and a_{i2} are the uncorrelated additive genetic values of the parents of family i [simulated as $N(0, h^2)$]. Because

the total variance was equal to 1, h^2 was equal to the additive genetic variance.

Populations were simulated with 5, 10 and 20 equally frequent alleles at each of 2, 4, 6, 10, 15, 20 and 30 loci. Heritabilities were simulated at five levels: 0, 0.25, 0.5, 0.75 and 1.0. Four full-sib family structures containing 150 individuals were simulated: 15 families of size 10, 30 families of size five, 75 families of size two and an unbalanced structure that followed a Poisson distribution with a mean family size of five. Simulations were repeated 1000 times for each set of starting parameters. Simulations were also run where family size remained at 10, but the number of families in the population was varied.

Simulations were run in which various percentages of marker information were simulated as missing, to investigate the case where some loci are not scored for genotype. Loci were selected at random from a generated data set and classed as missing. Locus-specific weights for the method of moments relationship estimator were altered so that weights summed to one. Simulations where half-sib families were used in place of full-sib families were also run.

In all cases estimates of heritability were compared against the REML estimates of heritability, which were calculated using the simulated pedigree (i.e. using actual relationships). Four statistics were calculated for each set of simulations: the deviation of each pair-wise estimate from the REML estimate; the sampling variance over simulations; the bias of the pair-wise estimate from the simulated parameter; and the mean squared error. Mean squared error is equal to the sampling variance plus the squared mean bias.

Results

Marker information

In general, the estimates approached the REML-derived estimate as marker information increased. This trend is observable for both an increase in the number of marker loci (Fig. 1a) or an increase in allele number per locus (not shown). In method 1, the regression-based estimator yielded estimates that are, on average, greater than the REML-derived estimates across the range of locus numbers. As marker information increases, the likelihood-based estimates approach the deviation predicted by theory (eqns 7 and 8).

The relative importance of phenotypic information in the likelihood techniques has a large effect at lower numbers of marker loci. With few marker loci, the posterior probabilities of the relationships become more dependent on the phenotypic information, whereas with larger numbers of loci the dependency is less strong.

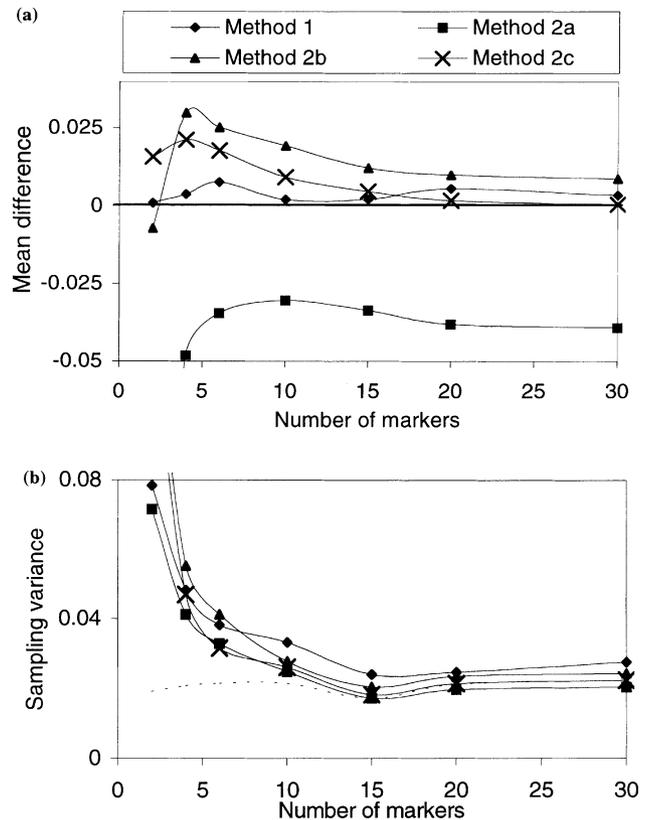


Fig. 1 (a) The mean difference between pair-wise heritability estimates and restricted maximum likelihood (REML)-derived estimates for the four estimators, in relation to the number of marker loci used to calculate relationships. Simulated heritability was 0.25, number of alleles per locus was 10, family structure was 15 full-sib families of size 10. Method 2a showed a large negative bias of 0.21 with two simulated loci. (b) The sampling variance of heritability estimates in relation to the number of marker loci. Simulated parameters as for (a). The dotted line denotes the sampling variance for the REML-derived estimates.

At lower marker numbers, the phenotypic information causes a slight upward bias in the heritability estimates because a pair that is phenotypically similar will also have a higher probability of being classed as full-sibs; this includes pairs that are not actually full-sibs.

With small numbers of markers (<5) and small numbers of alleles per locus (<10) different relationship classes cannot be distinguished, 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 (Fig. 1a). Similar graphs are obtained when the plots of mean deviation against marker number are made under different family structures, heritabilities and allele numbers (results not shown).

As marker information increases, the sampling variance of the heritability estimates decreases, approaching the sampling variance for REML estimates (Fig. 1b). As might be expected, method 1 shows the largest sampling variances because additional information about population structure (the prior probability that a pair are full-sibs) is only used in the likelihood estimators. The sampling variance of method 2b is larger than that of method 2c. Method 2c approaches the sampling variance of the REML-derived estimates because both yield unbiased estimates of the same two population parameters (V_w and V_b) in the balanced case. The method 2a heritability estimates are smaller than the REML-derived estimates (eqn 8); it is therefore expected that their sampling variance will fall below that of the REML-derived estimate as marker information increases. With increased numbers of alleles per locus, relationships are also more accurately estimated, resulting in increased accuracy of variance component estimation (results not shown).

Because REML gives unbiased estimates of variance components in the balanced case, average bias is very close to average difference. 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 true parameter value are more likely to be a result of sampling rather than of bias in the technique.

The simulated heritability

Method 2a yields smaller estimates of heritability than the REML-derived estimates (Fig. 2a). The magnitude of the mean difference decreases as simulated heritability increases. The remaining three estimates do not differ, on average, from the REML-derived estimates at zero heritability, but have increasingly positive difference as simulated heritability increases (Fig. 2a). This trend is especially marked for method 2b. The sampling variance of the estimates of heritability increases with simulated heritability for all the techniques (Fig. 2b), with the exception of method 2c, where sampling variance falls relative to the variance 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.

Missing data

The techniques showed little change in bias or mean difference, even with a high percentage (40%) of genotypes not scored. Sampling variance increased in

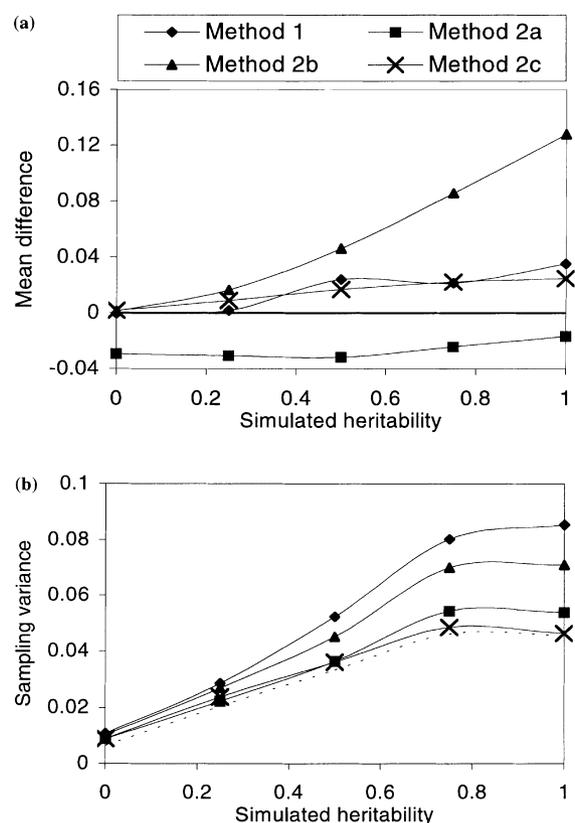


Fig. 2 (a) The mean difference of pair-wise heritability estimates from restricted maximum likelihood (REML)-derived estimates in relation to the simulated heritability. Number of loci was 10, number of alleles per locus was 10 and family structure was 15 full-sib families of size 10. (b) The sampling variance of heritability estimates in relation to the simulated heritability. Parameters as for (a). The dotted line denotes the REML estimate.

an approximately linear manner with increasing percentage of missing data. With 40% of the loci unscored the sampling variances were approximately double those of the sampling variances for data sets with the same simulated parameters and no unscored loci.

The sample size

The mean difference between pair-wise estimates and REML-derived estimates decreases as sample size increases (Fig. 3a). Sampling variance also decreases with increased sample size. The sampling variance for REML-derived estimates falls in proportion to the inverse of the sample size (Fig. 3b), as does that for the likelihood-based estimators. The sampling variance of the regression-based estimator, however, falls at a slower rate, reflecting an inefficient use of the data and the importance of having a sufficiently large variance of relatedness in this technique. An increase in the sample size, while

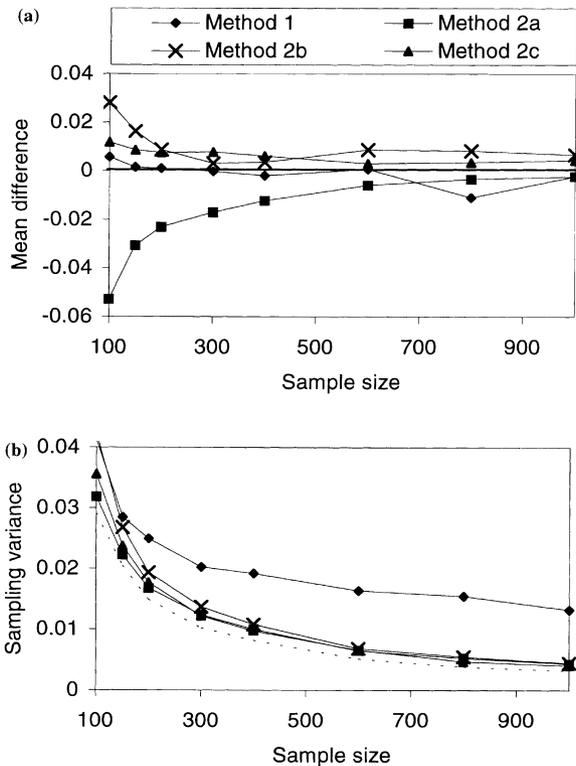


Fig. 3 (a) The mean difference of the pair-wise heritability estimates from the restricted maximum likelihood (REML)-derived estimates with different sample sizes. Number of loci was 10, allele number per locus 10, family size was 10 and simulated heritability 0.25. (b) The sampling variance of heritability estimates with different sample sizes. Simulated parameters as for (a). The dotted line denotes the REML estimate.

maintaining the same offspring number per family, 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. Again mean squared error is dominated by the sampling variance, indicating that the source of error is through sampling rather than bias.

Population structure

Different population structures also give rise to different variances in the relationships. Average difference is little affected by change in the simulated variance of relationship (Fig. 4a), except in situations where the variance of relationship was low, e.g. with 75 families of size two. The higher the variance of relationship in the simulated population the lower the sampling variance

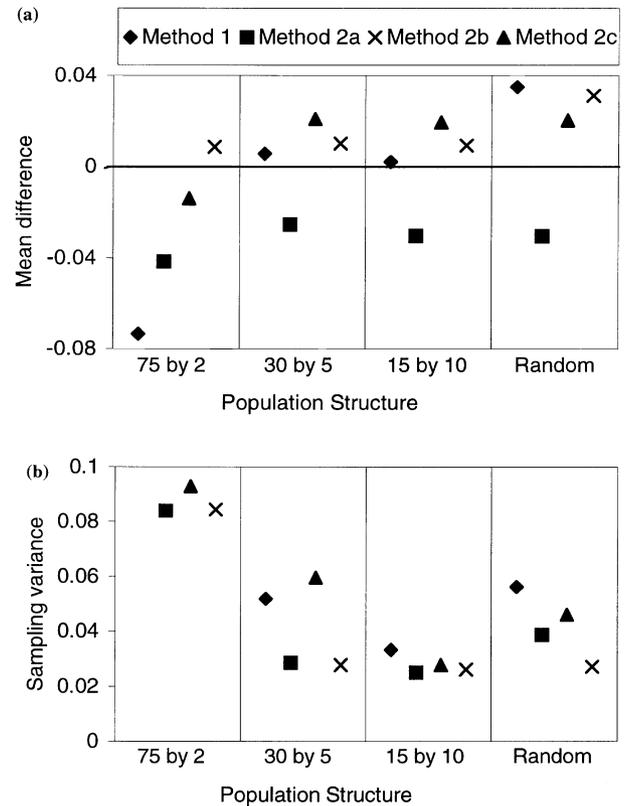


Fig. 4 (a) The mean difference of the pair-wise heritability estimates from the restricted maximum likelihood (REML)-derived estimates with different population structures. Number of loci was 10, allele number per locus 10, and simulated heritability 0.25. Random family structure signifies those simulations run with family sizes following a Poisson distribution with mean five. (b) The sampling variance of heritability estimates with different population structures. Simulated parameters as for (a). The variance was 0.4 for method 1 with the 75 families of size two.

and the smaller the bias of the estimate (Fig. 4b). With the family structure 75 by two, giving the smallest variance of relationship, the regression-based estimator (method 1) performed considerably worse than the likelihood-based estimators, having a sampling variance of 0.4. Estimates in populations with random family size showed slightly larger mean difference and sampling variances than those of populations with balanced families of the same mean size (Fig. 4a,b).

Half-sib families

The mean difference between pair-wise and REML estimates increased at each level of marker information when using half-sib families in place of full-sib families (Figs 1a and 5a). This may be caused by the decreased ability to resolve between half-sib pairs and unrelated

pairs compared with full-sib pairs and unrelated pairs in the case of the likelihood-based procedures (Thompson, 1975; Blouin *et al.*, 1996). In the regression-based technique this bias increase is caused by the reduction (by a factor of one-quarter) in the actual variance of the relationship. Additional increase in bias comes through the calculation of heritability in methods 2b and 2c, where bias in estimates of V_b is multiplied by 4, rather than 2. The likelihood-based procedures follow the same trend as Fig. 1(a), although shifted to the right, indicating a relatively lower precision even with larger amounts of marker information. Method 1 shows least bias over the range of simulated marker numbers, except at the highest levels of marker information, where method 2c shows least bias.

With a half-sib family structure compared to a full-sib structure there is increased sampling variance over that of the REML-based estimates in the estimates of heritability (Fig. 5b). Again this reflects the lower actual variance of relationship, and the poorer ability to distinguish between more distant levels of relationship.

Discussion and conclusions

In all cases the average deviation of the 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.

Method 1, 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. Method 2c, the likelihood-based technique that uses the phenotypic differences within pairs to calculate variance components, performs best out of the likelihood-based procedures, presumably because fewer assumptions about population parameters are required prior to calculation. At higher levels of marker information, method 2c shows less bias than the regression-based technique (Figs 1a and 5a). Simulations, run where both the sum and the difference of phenotype of pairs were combined into one estimator, indicate that there is little further information to be gained over the difference or over the sum, alone. This is despite the sum and the difference being uncorrelated for both full-sib and unrelated pairs. Results indicated that estimates made were more biased than the estimator using the difference only (because more population parameters require estimation prior to analyses), but yielded slightly lower sample variances at larger sample sizes.

All the methods share a number of basic properties. For accurate results they require that adequate amounts

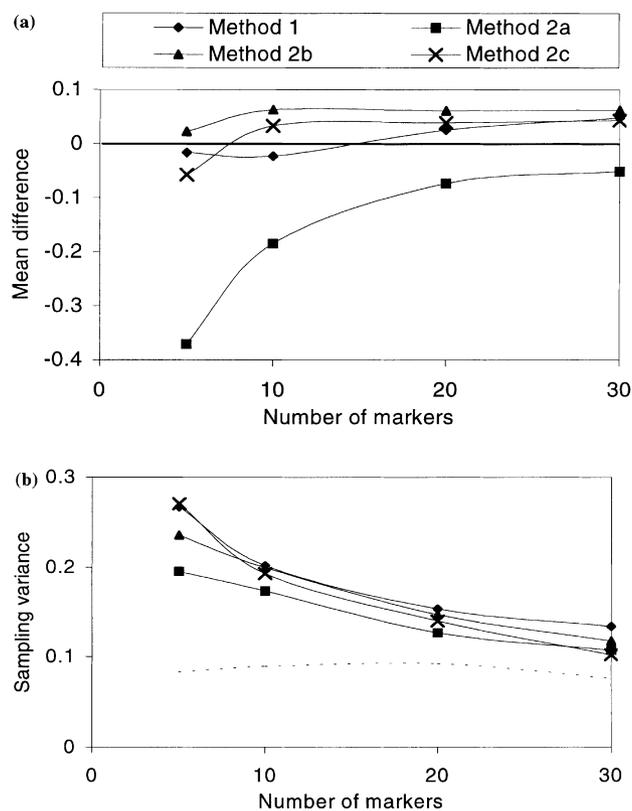


Fig. 5 (a) The mean difference of the pair-wise heritability estimates from the restricted maximum likelihood (REML)-derived estimates in relation to the number of marker loci. Simulated heritability was 0.25, allele number per locus was 10, and family structure was 15 half-sib families of size 10. (b) The sampling variance of heritability estimates in relation to the number of marker loci. Simulated parameters as for (a). The dotted line denotes the variance for the REML-derived estimates.

of marker information be used in the estimation of relationships. They also require that a sufficiently large variance of relationship be 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 is available from which to draw 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. Sampling variances 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 a less efficient method.

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, in all of the likelihood techniques there is upwards bias arising from the inclusion of phenotypic information in the likelihood calculation because a pair that is phenotypically similar has a higher probability of being classed as full-sibs.

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 an increase in the sampling variance 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 comprising small half-sib 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 could 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 procedure to estimate the relationships and therefore does not require extension to deal with combinations of relationships, provided there is sufficient variation in actual relatedness within the population. 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. The law of invariance means that distributions for phenotypic information may be maximized with respect to the additive genetic (V_a) and environmental (V_e) variance components rather than the within-family and between-family variances, because this allows increased generality in the relationship classes under consideration; for example, the expected distributions of the differ-

ence between unrelated pairs may be rewritten as $N(0, 2V_a + 2V_e)$, half-sib pairs as $N(0, 1.5V_a + 2V_e)$ and full-sib and parent-offspring pairs as $N(0, V_a + 2V_e)$. However, for both types of technique the ability to distinguish between more distant relationship classes falls rapidly with the increase in relationship distance (Thompson, 1975; Blouin *et al.*, 1996). This results in poorer estimates of variance components in populations with low variance in relatedness.

An additional consideration concerns 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. 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 likelihood 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 the relationship through the father may be based on the remaining allele. In practice, errors in genotyping must be accounted for if this technique is to be adopted in a practical situation (Marshall *et al.*, 1998).

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 random effects (e.g. maternal effects) as well as 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). Of particular note is the between-family environmental effect. This would increase the similarity of individuals within a family, increasing estimates of heritability. Because the methods investigated here are based on pair-wise computations rather than on individuals it is harder to incorporate these effects into the model. One solution to this problem might be to use the relationship estimation procedures to construct an estimated relationship matrix that might be used in the REML procedure. This would allow inclusion of additional effects into the model in

the traditional manner, and might improve estimation of parameters in a heterogeneous population through accommodation, at least in part, of appropriate weightings for families. However, the result of using an estimated relationship matrix, assumed known in conventional studies, on the accuracy of estimates of variance components is unclear.

Acknowledgements

Stuart Thomas was funded by a BBSRC Ph.D. studentship. Peter Keightley and Andy Peters gave constructive comments on the manuscript.

References

- BLOUIN, M. S., PARSONS, M., LACAILLE, V. AND LOTZ, S. 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.
- BULMER, M. G. 1980. *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, New York.
- EDWARDS, A. W. F. 1972. *Likelihood*. Cambridge University Press, Cambridge.
- FALCONER, D. S. AND MACKAY, T. F. C. 1996. *Introduction to Quantitative Genetics*, 4th edn. Longman, Harlow, Essex.
- JACQUARD, A. 1974. *The Genetic Structure of Populations*. Springer-Verlag, New York.
- LANDE, R. 1982. A quantitative genetic theory of life history evolution. *Ecology*, **63**, 607–615.
- LANDE, R. AND SHANNON, S. 1996. The role of genetic variation in adaptation and population persistence in a changing environment. *Evolution*, **50**, 434–437.
- LYNCH, M. AND RITLAND, K. 1999. Estimation of pairwise relatedness with molecular markers. *Genetics*, **152**, 1753–1766.
- LYNCH, M. AND WALSH, B. 1998. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.
- MARSHALL, T. C., SLATE, J., KRUIK, L. E. B. AND PEMBERTON, J. M. 1998. Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol.*, **7**, 639–655.
- MOUSSEAU, T. A. AND ROFF, D. A. 1987. Natural selection and the heritability of fitness components. *Heredity*, **59**, 181–197.
- MOUSSEAU, T. A., RITLAND, K. AND HEATH, D. D. 1998. A novel method for estimating heritability using molecular markers. *Heredity*, **80**, 218–224.
- QUELLER, D. C. AND GOODNIGHT, K. F. 1989. Estimating relatedness using genetic markers. *Evolution*, **43**, 258–275.
- 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.
- STORFER, A. 1996. Quantitative genetics: a promising approach for the assessment of genetic variation in endangered species. *Trends Ecol. Evol.*, **11**, 343–348.
- THOMPSON, E. A. 1975. The estimation of pair-wise relationship. *Ann. Hum. Genet.*, **39**, 173–188.
- WEIR, B. S. 1996. *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.