Genet. Res., Camb. (2003), **81**, *pp.* 145–156. *With* 1 *figure.*
DOI: 10.1017/S0016672302006055 Printed in the United Kingdom

Marker-assisted estimation of quantitative genetic parameters in rainbow trout, *Oncorhynchus mykiss*

A. J. WILSON[1]*, G. McDONALD[1], H. K. MOGHADAM[1], C. M. HERBINGER[2] AND M. M. FERGUSON[1]
[1] *Department of Zoology, University of Guelph, Guelph, ON, N1G 2W1, Canada*
[2] *Department of Biology, Dalhousie University, Halifax, NS, B3H 4J1, Canada*

(*Received 29 July 2002 and in revised form 7 November 2002*)

Summary

Estimation of quantitative genetic parameters conventionally requires known pedigree structure. However, several methods have recently been developed to circumvent this requirement by inferring relationship structure from molecular marker data. Here, two such marker-assisted methodologies were used and compared in an aquaculture population of rainbow trout (*Oncorhynchus mykiss*). Firstly a regression-based model employing estimates of pairwise relatedness was applied, and secondly a Markov Chain Monte Carlo (MCMC) procedure was employed to reconstruct full-sibships and hence an explicit pedigree. While both methods were effective in detecting significant components of genetic variance and covariance for size and spawning time traits, the regression model resulted in estimates that were quantitatively unreliable, having both significant bias and low precision. This result can be largely attributed to poor performance of the pairwise relatedness estimator. In contrast, genetic parameters estimated from the reconstructed pedigree showed close agreement with ideal values obtained from the true pedigree. Although not significantly biased, parameters based on the reconstructed pedigree were underestimated relative to ideal values. This was due to the complex structure of the true pedigree in which high numbers of half-sibling relationships resulted in inaccurate partitioning of full-sibships, and additional unrecognized relatedness between families.

1. Introduction

Knowledge of the genetic architecture underlying quantitative traits is critical to our understanding of phenotypic determination and evolution. Within the conceptual framework of quantitative genetics, this architecture is described in terms of genetic parameters such as trait heritabilities and genetic correlations (Falconer & Mackay, 1996). Conventionally these parameters are most readily estimated by comparison of phenotype between individuals of known relationship, and thus there is a requirement for a known pedigree. This requirement can pose difficulties for the study of natural populations in which relationships between individuals are unknown. In natural animal populations pedigree may sometimes be determined by observation, though this typically requires long-term study with intensive sampling. In some taxa (notably birds and mammals), the common provision of parental care has facilitated this approach and allowed estimation of heritabilities in the field (e.g. Merilä *et al.*, 1999; Qvarnström, 1999; Cadée, 2000). However, in many taxa such observational pedigree determination is not feasible. As a result, estimation of quantitative genetic parameters in natural populations has generally been limited. Various possible alternatives have been suggested including; the use of phenotypic measures of variance and covariance to infer underlying genetic architecture (Cheverud, 1988), regressing phenotypic trait values of lab-reared offspring on their wild-caught parents (Riska *et al.*, 1989), or extrapolation of lab-based estimates to natural populations (Weigensberg & Roff, 1996). However heritability is dependent on both genetic and environmental components of phenotypic variance, and as such it is a population specific parameter. Thus the validity of any

* Corresponding author. Fax: +1 (519) 767 1656. e-mail: awilso00@uoguelph.ca

ex situ estimate, even if based on wild-caught parents to control genetic factors (Riska *et al*., 1989), will be flawed since environmental factors in the lab will not duplicate those in the field.

More recently, attention has been focused on the potential use of molecular markers to infer relationships between individuals, and hence circumvent this problem of unknown pedigree (see Ritland, 2000 for a review). In some cases partial pedigree information might be available and genotypic data can then be used to derive supplemental information. For example, maternity might sometimes be determined by observation, whilst paternity remains unknown. In this context paternity analysis has been used to provide additional pedigree information that can result in improved estimation of genetic parameters (e.g. Kruuk *et al*., 2000; King *et al*., 2001). Where there is no pedigree information available, several approaches to marker-assisted estimation of quantitative genetic parameters have been proposed. These approaches can be separated into those that do not depend on explicit pedigree reconstruction, and those that do.

A maximum likelihood-based procedure for estimating trait heritabilities without using explicit pedigree information was described by Mousseau *et al*. (1998) with subsequent development by Thomas *et al*. (2000). This approach is based on using the joint probability of observed phenotypic and genotypic data to determine likelihoods for assigning pairs to each of several specified relationship classes. However, this maximum likelihood estimator requires *a priori* knowledge of the distribution of relatedness (e.g. all individuals are either full-sibs or unrelated), and is therefore not appropriate where this distribution is unknown (Mousseau *et al*., 1998). A more generally applicable estimator was presented by Ritland (1996*b*), in which trait heritabilities are estimated from a linear regression of pairwise phenotypic similarity on pairwise relatedness. This approach therefore relies on estimators of pairwise relatedness (e.g. Queller & Goodnight, 1989; Ritland, 1996*a*; Lynch & Ritland, 1999), but again does not require specification of an explicit pedigree. In addition to its wider applicability, simulation-based studies found that this method resulted in decreased bias in estimates of heritability (h^2) as compared to the maximum likelihood estimator, though variance was higher (Thomas *et al*., 2000). High sampling variance is a feature of all pairwise relatedness estimators (Van de Casteele *et al*., 2001) and may reduce the utility of this latter method.

Alternatively genotypic information might be used to explicitly reconstruct relationships. In natural populations pedigree reconstruction from marker data might involve determination of parent-offspring relationships (e.g. Marshall *et al*., 1998) or of sibships (e.g. Painter, 1997; Almudevar & Field, 1999; Thomas & Hill, 2000; Smith *et al*., 2001). These pedigrees can subsequently be used to estimate quantitative genetic parameters using conventional methods such as parent-offspring regression, sib analysis or restricted maximum likelihood (REML). In particular REML estimators do not require balanced data sets and can be obtained for any arbitrary pedigree that might occur in a natural population (Lynch & Walsh, 1998). Whilst parentage analysis represents a useful approach, it may be limited in many natural populations where there is likely to be incomplete sampling of candidate parents. Although likelihood-based methodologies can be used with incomplete sampling of parents, successful parentage assignment decreases rapidly as a function of the proportion of candidate parents sampled (Marshall *et al*., 1998). Thus efficient sampling for parentage assignment may be problematic in systems lacking extensive overlap of generations (on either spatial or temporal scales). In such situations sibship reconstruction might provide a more practical approach. Thomas & Hill (2000) demonstrated the use of Markov chain Monte Carlo (MCMC) techniques to partition populations into full-sibships that were then used in an animal model to estimate genetic parameters. Based on simulation studies, this method was found to be superior to those methods described above that do not rely on explicit pedigree reconstruction (Thomas & Hill, 2000).

To date there has been comparatively little empirical work done in this area of marker-assisted parameter estimation (but see Ritland & Ritland, 1996; Mousseau *et al*., 1998; Thomas *et al*., 2002). Furthermore, success has been mixed, and the generation of reasonable estimates would seem to be dependent on the population genetic structure of the system being considered. In particular it may be necessary to focus on systems in which there is high variation in relatedness between individuals sampled (Ritland, 2000), or equivalently in which family sizes are likely to be large (Thomas *et al*., 2002). Despite such limitations, the advantages of being able to generate field-based estimates of genetic parameters are such that there is a clear need for further empirical work to test the utility of these methods.

The current study attempts to address this need for empirical study by examining marker-assisted estimation of trait heritabilities and genetic correlations in rainbow trout, *Oncorhynchus mykiss*. The data consist of genotypic and phenotypic information from an aquaculture strain development project (McDonald, 2001; Quinton, 2001). In this project parental fish from three strains of rainbow trout (denoted B, G and O) were used in both intra- and inter-strain crosses to generate a progeny generation. Parental individuals were used in multiple crosses such that the progeny generation contains many full-sib families that share one parent with other families. Given this prior knowledge of the manner in which crosses were performed,

Table 1. *Variability of microsatellite loci used, showing number of alleles* (n) *and expected heterozygosity* (H_e) *for data sets ALL, B, G and O. Values are based on allele frequencies for both generations combined*

	ALL		B		G		O	
Locus	*n*	H_e	*n*	H_e	*n*	H_e	*n*	H_e
OmyRGT41TUF[1]	16	0·869	8	0·800	9	0·796	6	0·788
OmyFGT5TUF[2]	9	0·703	5	0·541	5	0·697	4	0·542
OmyFGT10TUF	5	0·530	4	0·614	2	0·433	2	0·428
OmyFGT12TUF[1]	18	0·909	16	0·894	10	0·884	6	0·736
OmyFGT14TUF	4	0·468	4	0·612	3	0·239	2	0·438
OmyFGT15TUF[2]	5	0·432	5	0·494	5	0·298	4	0·482
OmyFGT23TUF	12	0·821	9	0·824	8	0·695	6	0·755
OmyFGT34TUF[1]	17	0·919	12	0·911	8	0·812	5	0·726
Omy27DU	6	0·654	5	0·588	4	0·618	3	0·630
Omy77DU[2]	10	0·744	9	0·834	7	0·666	4	0·686
Omy325UoG	12	0·808	8	0·823	7	0·767	3	0·661
One18ASC	7	0·705	5	0·728	6	0·684	3	0·538
Ots1BML[2]	12	0·759	11	0·817	5	0·680	7	0·803
Ssa85DU[1]	14	0·875	11	0·880	5	0·757	5	0·768
Ssa20.19NUIG	5	0·770	5	0·774	3	0·521	4	0·708
SSOSL439	14	0·826	11	0·821	9	0·804	4	0·537
Mean	9·76	0·694	7·53	0·703	5·65	0·609	4	0·602

[1] Only parental generation genotyped at these loci.
[2] Not all progeny were genotyped at these loci.

and that family sizes produced by such crosses may potentially be large, it is anticipated that this captive population will have a structure amenable to marker-assisted estimation of quantitative genetic parameters. By using microsatellite data, we estimate such parameters for phenotypic traits of body weight and spawning time according to two alternative methods. Firstly we apply the regression-based method of Ritland (1996*b*) that requires no explicit pedigree structure, and secondly estimates are made using a pedigree obtained from reconstruction of intra-generational sibships. Performance of these methods is examined by comparison of estimated genetic parameters to those obtained from the true pedigree obtained from parentage analysis.

2. Methods

(i) *Data sets*

Data were obtained from a rainbow trout strain development project involving three aquaculture strains. Seventy-one parental fish were used to obtain a progeny generation containing 595 individuals derived from both intra- and inter-strain crosses (see McDonald, 2001 and Quinton, 2001 for details). Phenotypic trait data describing 2-year weight (WT) and female spawning time at ages 3 (ST3) and 4 (ST4) were available (see Quinton, 2001 for details). Spawning time is defined as the number of days after October 1st in the given spawning season.

Tissue samples were taken from both parental and progeny generations, and DNA was extracted using a phenol chloroform procedure (Bardakci & Skibinski, 1994). Twelve microsatellite loci were then amplified in the PCR using the multiplexed reactions of Fishback *et al*. (1999). PCR products were visualized using an ABI 377 DNA sequencer and alleles were scored using GENESCAN™ 2.0.0 and GENOTYPER™ 1.1r8. An additional three loci were amplified in single locus PCR reactions as described in McDonald (2001). Products of these reactions were separated using gel electrophoresis in a 6% polyacrylamide-7-M urea matrix. Alleles were visualized using the Hitachi FMBIOII fluorescence imaging system, and sized by comparison with 350-TAMRA lane standards loaded on each gel.

Genetic data are thus comprised of genotypes for up to sixteen moderately- to highly-variable loci (Table 1). Individuals of the parental generation were genotyped at all sixteen loci, while all of the progeny were genotyped at a common subset of eight of these. For some progeny individuals, genotypic information was available at an additional subset of four of those loci used in the parents.

Parentage analysis was carried out to determine the true pedigree structure, using the microsatellite data to assign progeny to parental pairs on the premise that a parent and offspring must share at least one allele at every codominant locus. Using this purely exclusion-based approach as implemented in PROBMAX (Danzmann, 1997), 97% of the progeny were

https://doi.org/10.1017/S0016672302006055 Published online by Cambridge University Press

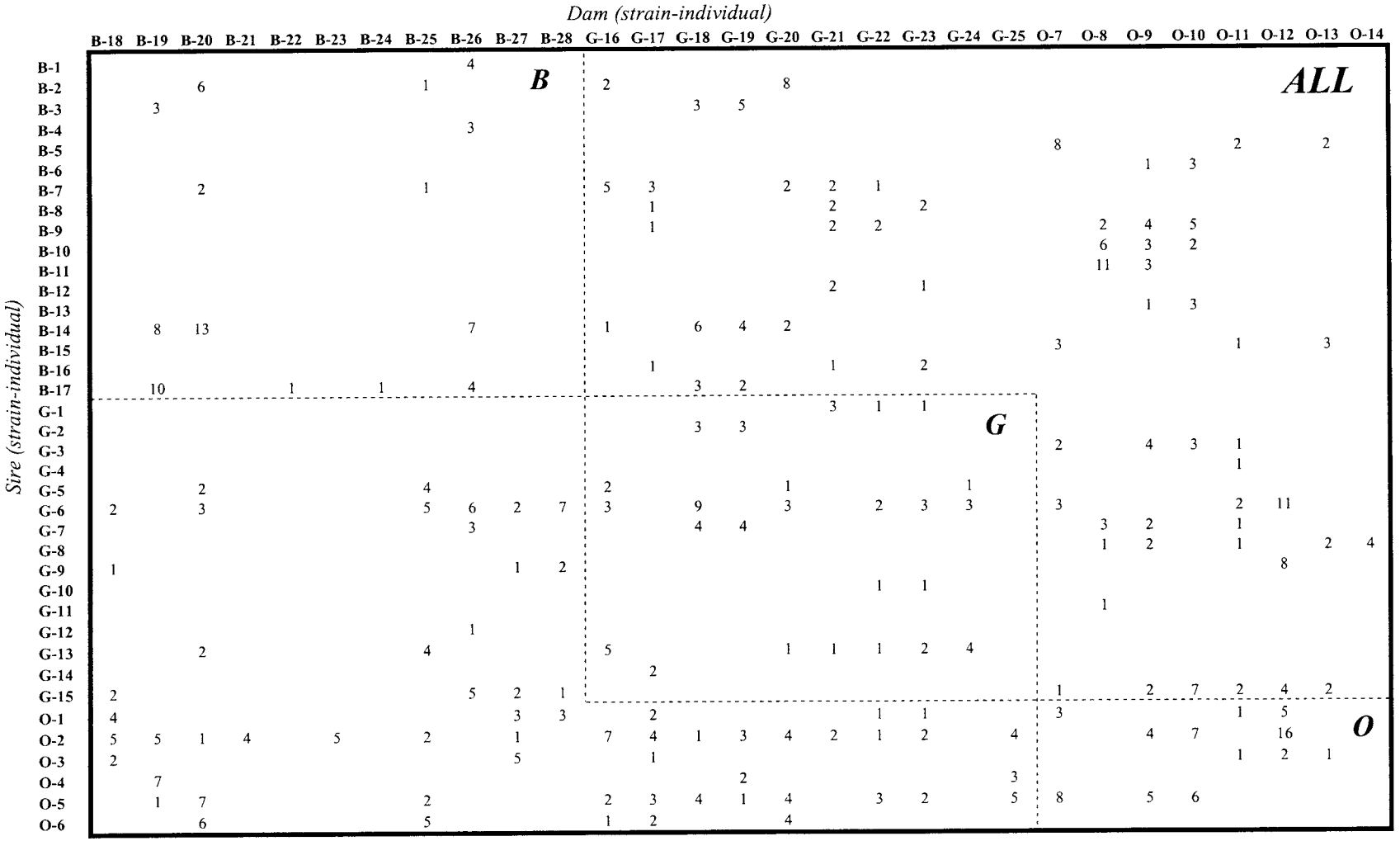

Fig. 1. True pedigree as determined by parentage analysis, indicating dam, sire and size for families in data sets ALL, B, G and O.

assigned to a single parental pair consistent with Mendelian segregation of alleles. The remaining 3% (18 individuals) could be assigned to more than one possible parental pair and were therefore excluded from the data set such that this pedigree is considered correct. Parentage analysis in this data set indicated the presence of 182 full-sibling families, with family size ranging from 1 to 16, and furthermore revealed the complex nature of the pedigree structure that results from the use of parental individuals in multiple crosses (Figure 1).

In addition to the complete rainbow trout data set, subsequently denoted ALL, three pure-strain subsets are considered in the subsequent analysis. The subsets are henceforth denoted B, G and O, corresponding to the three strains used in the strain development project (McDonald, 2001). Each subset is comprised of the parental generation individuals from a given strain, together with all progeny resulting from crosses between males and females of that strain (Figure 1). These subsets are examined to investigate relative performance of methodologies with smaller sample sizes. Since sample size is considerably less for spawning time traits as opposed to weight (only females can be scored for the former), estimation of genetic parameters for ST3 and ST4 may not be informative but is nevertheless included for completeness.

(ii) *Data treatment*

(a) *Regression method*

The procedure of Ritland (1996*b*) relies on a simple regression of pairwise trait similarity on pairwise relatedness to estimate genetic parameters. For example, heritability (h^2) is estimated from the model:

$$Z_{ij} = 2r_{ij}h^2 + r_e,$$

Z_{ij} is a measure of trait similarity between individuals i and j, calculated as the product of the trait values of the individuals (after normalization to variates of mean zero and variance one). The residual error term, r_e, can be interpreted as the average environmental correlation between the two individuals. We follow the notation of Ritland (1996*b*) in using r_{ij} to denote the pairwise relatedness, defined as the probability that two genes, one drawn at random from each individual, will be identical by descent. The estimator for r_{ij} was calculated from microsatellite data according to Ritland (1996*a*). Relative performance of different relatedness estimators may vary such that there is no single best-performing estimator (Van de Casteele *et al*., 2001). Here the estimator of Ritland (1996*a*) was selected on the basis of preliminary simulations (using allele frequencies as estimated from ALL), to compare performance of several estimators (results not shown).

Since the model requires no explicit pedigree structure it can be used in samples that span multiple generations. Thus both parental and progeny information was combined for estimating trait heritabilities and genetic correlations. This regression-based procedure was implemented using the program MaRQ version 1.0 (written by K. Ritland, available from http://genetics.forestry.ubc.ca/ritland/programs.html). Estimates were made of trait heritabilities, genetic correlations, and the actual variance of relatedness in each data set. It should be noted that in the absence of significant actual variance of relatedness, heritability estimates made under this model cannot be significant. For each pair of traits, the sign of a genetic correlation was determined as the sign of $\mathrm{Cov}(r_{ij}, Z_{ij})$, with Z_{ij} being the "similarity" between trait 1 in individual i and trait 2 in individual j. Significance of all parameters estimated was determined by bootstrapping over individuals with a bootstrap number of 1000. Estimates were deemed significant if 95% of the bootstrap values were found to be greater than zero.

In order to examine the performance of the relatedness estimator underlying the estimates of genetic parameters we define the test statistic Δ_{ij}, where:

$$\Delta_{ij} = \text{estimated } r_{ij} - \text{true } r_{ij},$$

with true r_{ij} being defined as 0 between unrelated individuals, 0·125 between half-sibs, and 0·25 between full-sibs or between parents and offspring. True r_{ij} was therefore determined from the true pedigree (determined as described above), under the assumption that all parental individuals were unrelated. For each data set the mean value of Δ_{ij}, was then used to describe the bias of the estimator. Normality of Δ_{ij} was assumed to calculate 95% confidence limits to the mean and hence test for significant deviation from zero. Variance of Δ_{ij} was also calculated as a further measure of estimator performance. These statistics were calculated using all possible pairwise values of Δ_{ij}, and for each data set, mean and variance of Δ_{ij} were also obtained for each value of true relatedness (i.e. true r_{ij} equal to 0·25, 0·125 or 0). In each data set, all pairwise values of true r_{ij} were also used to calculate the actual variance of relatedness (for comparison to the estimates made using MaRQ).

(b) *Sibship reconstruction*

For each data set, individuals were partitioned into groups of putative full siblings using the MCMC approach described by Smith *et al*. (2001), which maximizes an overall likelihood score on the basis of pairwise likelihood ratios of being full siblings or unrelated. The algorithm is constrained such that within a group of putative siblings the genotypes at each locus must be able to be derived from a single parental

https://doi.org/10.1017/S0016672302006055 Published online by Cambridge University Press

Table 2. *Estimates of trait heritabilities and genetic correlations* (± SE). *For the regression method 95% confidence limits obtained from the bootstrap percentile method are indicated in parentheses. Significant deviation from 0 is denoted by *. The number of individuals scored for each trait/pair of traits in the progeny generation is denoted by n for the both generations combined, and by n_p for the progeny alone*

				Heritability/Genetic correlation		
Data set	Trait(s)	*n*	n_p	*Regression method*	*Reconstructed pedigree*	*Ideal*
ALL	WT	628	558	1·78* (1·39, 2·26)	0·382 ± 0·031*	0·455 ± 0·044*
	ST3	268	247	2·41* (1·68, 3·33)	0·456 ± 0·049*	0·500 ± 0·063*
	ST4	279	199	2·35* (1·61, 3·19)	0·402 ± 0·058*	0·498 ± 0·069*
	WT–ST3	265	244	−0·770* (−0·606, −0·951)	−0·698 ± 0·066*	−0·861 ± 0·052*
	WT–ST4	275	195	−0·711* (−0·532, −0·911)	−0·617 ± 0·056*	−0·882 ± 0·055*
	ST3–ST4	265	195	0·994* (0·965, 1·08)	0·951 ± 0·017*	0·991 ± 0·008*
B	WT	91	62	0·746 (−0·253, 2·21)	0·278 ± 0·085*	0·296 ± 0·114*
	ST3	35	31	2·56 (−0·149, 5·87)	0·120 ± 0·164	0·179 ± 0·154
	ST4	35	22	0·952 (−0·861, 3·59)	0·238 ± 0·178	0·160 ± NE
	WT–ST3	32	28	−0·0390 (−2·50, 1·36)	−0·866 ± 0·462	−1·000 ± NE
	WT–ST4	31	18	−0·194 (−7·23, 7·58)	−0·848 ± 0·268*	−1·000 ± NE
	ST3–ST4	29	18	1·11 (−3·60, 3·50)	1·000 ± NE	1·000 ± NE
G	WT	90	62	0·489* (0·0169, 1·93)	0·137 ± 0·108	0·000 ± 0·000
	ST3	27	26	0·115 (−1·82, 2·71)	0·502 ± 0·171*	0·000 ± 0·000
	ST4	30	24	0·853 (−0·750, 3·84)	0·238 ± 0·371	0·000 ± 0·000
	WT–ST3	24	23	−0·274 (−7·26, 4·69)	0·389 ± 0·549	0·986 ± 0·999
	WT–ST4	26	20	−0·0750 (−4·33, 2·49)	1·000 ± NE	0·000 ± 1·000
	ST3–ST4	20	20	3·57 (−12·9, 12·7)	0·925 ± 0·094*	0·000 ± 1·000
O	WT	77	56	−1·29 (−17·3, 12·6)	0·213 ± 0·165	0·225 ± 0·028*
	ST3	38	26	−0·652 (−20·0, 25·3)	0·000 ± 0·000	0·004 ± 0·000*
	ST4	31	18	−2·06 (−35·8, 23·1)	0·135 ± NE	0·120 ± NE
	WT–ST3	35	26	10·5 (−47·4, 75·3)	1·000 ± 1·000	1·000 ± NE
	WT–ST4	27	18	−1·08 (−27·1, 19·5)	1·000 ± NE	−1·000 ± NE
	ST3–ST4	25	18	1·81 (−23·6, 30·2)	0·135 ± NE	−1·000 ± NE

pair. Multiple runs were performed to assess solution convergence, and reconstructed pedigrees were based chain lengths of 10 million (ALL) and 3·9 million (B, G and O) iterations. Phenotypic data was recoded using PEST (Groeneveld & Kovac, 1990), and then used in conjunction with reconstructed pedigree to estimate quantitative genetic parameters with standard errors. Estimates were made using a REML procedure implemented in VCE4 (Groeneveld, 1994) under the sire model:

$$y_i = \mu + a_{si} + e_i,$$

where y_i is a phenotypic observation for individual i, μ is the population mean (or strain-specific population mean for the data sets B, G and O), a_{si} is the random sire effect, and e_i is a residual error term. This simple model was used since, in comparison to an animal model, it is expected to yield smaller standard errors associated with estimated genetic parameters. This will permit more effective comparison of results to "ideal" parameter estimates (described below).

In order to examine the performance of the sibship reconstruction algorithm, partitioned sibships were examined visually to assess homology with true sibship groups (from the known pedigrees) in all four data sets. Summary statistics were used to describe the number and mean size of families in the reconstructed (and true) pedigrees. The number of correctly and incorrectly reconstructed full-sib pairs were counted for each reconstructed pedigree and used to calculate the accuracy statistic of Thomas & Hill (2000) in which:

$$\text{accuracy} = (S_{fs/fs} - S_{fs/nfs})/Tot_{fs},$$

where $S_{fs/fs}$ and $S_{fs/nfs}$ are the numbers of correctly and incorrectly reconstructed full-sib pairs, and Tot_{fs} is the number of full-sib pairs in the true pedigree. Furthermore since the true pedigree contained a mix of full-sib, half-sib and unrelated pairs, it was of interest to examine the extent to which incorrect partitioning of half-sibs as full-sibs accounted for observed error in pedigree reconstruction. To this end we define the statistic half-sib error, where:

$$\text{half-sib error} = S_{fs/hs}/S_{fs/nfs},$$

where $S_{fs/hs}$ is the number of incorrectly partitioned full-sib pairs in the reconstructed pedigree that were actually true half-sibs.

To further assess the performance of the pedigree reconstruction algorithm, 1000 pedigrees were simulated for each data set by assigning individuals to families at random. From these simulated pedigrees,

https://doi.org/10.1017/S0016672302006055 Published online by Cambridge University Press

mean values (and standard errors) for accuracy and half-sib error were determined. These were used to test whether the observed values (i.e., those derived from the reconstructed pedigrees) were larger than might be expected with random assignment of individuals to families. The distribution of family sizes (which will affect both accuracy and half-sib error) was kept constant and identical to the distribution of family sizes in the reconstructed pedigree. This was carried out for each of the four data sets.

(c) *Estimation of "ideal" parameters from true pedigree*

The true pedigree (as determined by full parentage analysis) was used to generate values of genetic parameters to which the marker-assisted estimates can be compared. All parental and progeny information was used to generate trait heritabilities and genetic correlations under the sire-model approach as described above.

Table 3. *Performance of relatedness estimator in each data set and class of true r_{ij}. n_{ij} indicates number of pairwise estimates made, * denotes significant departure from zero*

Data set	true r_{ij}	n_{ij}	mean Δ_{ij}	var Δ_{ij}
ALL	all pairs	210 276	−0·0129*	0·00536
	0·25	2322	−0·0168*	0·0426
	0·125	15 141	−0·0371*	0·0146
	0	192 813	−0·0110*	0·00414
B	all pairs	4278	−0·0469*	0·00526
	0·25	334	−0·112*	0·00910
	0·125	696	−0·0837*	0·00573
	0	3248	−0·0324*	0·00384
G	all pairs	4371	−0·0347*	0·00719
	0·25	229	−0·0649*	0·0222
	0·125	533	−0·0693*	0·0100
	0	3609	−0·0277*	0·00555
O	all pairs	2775	−0·0624*	0·00992
	0·25	332	−0·158*	0·00899
	0·125	559	−0·0988*	0·0101
	0	1884	−0·0348*	0·00728

3. Results

(i) *Regression method*

In the full data set (ALL), actual variance of relatedness was estimated as 0·002 and was found to be significantly greater than zero. The regression-based approach resulted in estimates of heritability that were significant for all three traits (Table 2). Furthermore significant negative genetic correlations were determined between WT and both spawning time traits, while between ST3 and ST4 there was a significant positive genetic correlation. These results are qualitatively consistent with the "ideal" estimates obtained using the REML methodology in the true pedigree (Table 2). However compared to the ideal estimates, regression-based estimates of trait heritabilities all exhibit significant upward bias. It should be noted that under this model heritability is not constrained to lie between 0 and 1. In contrast, no bias is seen in the estimation of genetic correlations, which are comparable in magnitude to the optimal estimates. In all cases the sign of the estimated genetic correlations is consistent with ideal results. For all parameters, precision of regression-based estimators is low. This is particularly true for heritability estimates where 95% confidence intervals are an order of magnitude greater than those of the optimal estimates.

In the pure-strain subsets of the data, estimated actual variances of relatedness were significantly greater than zero (with values of 0·003, 0·005 and 0·003 corresponding to B, G and O respectively). However, small sample sizes led to difficulties in estimating standard errors for ideal genetic parameters relating to spawning time traits, (especially genetic correlations; Table 2). As such comparison between ideal and regression-based estimates of genetic parameters is problematic for these traits. Nevertheless examination of heritabilities for WT suggests that performance of this method is lower in the pure-strain subsets of the data. In particular, while the ideal estimates suggest significant heritabilities for WT in data sets B and O but not G, the regression-based estimator indicates the opposite (Table 2). In B, heritability was considerably overestimated (0·746 as opposed to the ideal estimate of 0·296), but remains non-significant due to the wide confidence interval. In O, the regression-based estimate of h^2 was negative, a result that is interpreted as zero.

Significant bias was found to occur in the estimation of pairwise relatedness that is used in the regression method (Table 3). For all data sets and all classes of true relatedness, mean Δ_{ij} is significantly less than zero, indicating pairwise relatedness is systematically underestimated. In all data sets bias varies across relatedness classes, being generally lowest for the true $r_{ij}=0$ relatedness class, and highest for true $r_{ij}=0{\cdot}25$. Among data sets bias has a greater magnitude in the smaller pure-strain subsets than in the complete data set. Estimates of var Δ_{ij} are of comparable magnitude across the data sets. Actual variances of relatedness were calculated as 0·00156 for ALL, and 0·00584, 0·00438, and 0·00759 for B, G and O respectively.

(ii) *Sibship reconstruction*

In the full data set, sibship reconstruction resulted in estimates of quantitative genetic parameters that also

https://doi.org/10.1017/S0016672302006055 Published online by Cambridge University Press

Table 4. *Summary data for true and reconstructed pedigrees. Accuracy and half-sib error are reported as observed in reconstructed pedigrees, and as expected with random assignment of individuals to families. * Denotes observed value significantly greater than expected value at* $\alpha = 0{\cdot}001$

	ALL	B	G	O
Number of progeny	578	64	63	59
True pedigree				
no. full-sib families	182	14	25	12
mean family size	3·18	4·57	2·52	4·92
no. full-sib pairs	1166	206	86	214
no. half-sib pairs	15 141	696	533	559
no. unrelated pairs	150 446	1114	1334	938
Reconstructed pedigree				
no. full-sib families	133	17	30	20
mean family size	4·34	3·76	2·10	2·81
no. full-sib pairs	1476	142	57	142
(correct)	550	118	29	32
(incorrect)	926	21	28	43
Accuracy				
Observed (reconstructed pedigree)	−0·322*	+0·471*	+0·012*	−0·051*
Expected (with random assignment)	−1·248	−0·536	−0·604	−0·263
Half-sib error				
Observed (reconstructed pedigree)	0·640*	1·000*	0·714*	0·605*
Expected (with random assignment)	0·092	0·384	0·256	0·372

showed qualitative agreement with the ideal values (Table 2). Furthermore, quantitative agreement is relatively good between the marker-assisted and ideal estimates, though the sibship-reconstruction method did result in all parameters being underestimated. Based on standard errors presented, differences from ideal values are not significant except for the genetic correlation between WT and ST4. It should be noted that the standard errors associated with the marker-assisted estimates are based on the assumption that the reconstructed pedigree is correct, and thus as a measure of precision they do not adequately capture all sources of error.

In the pure strain subsets, comparisons are again complicated by problems in estimating standard errors for both marker-assisted and ideal estimates. However, marker-assisted estimates for heritability of WT show close agreement with ideal values in B and O, and there is no evidence of significant bias. The same is true for heritabilities of ST3 and ST4, though for the latter traits standard errors could not be calculated so that the significance of deviation in these data sets cannot be assessed. In G, trait heritabilities were significantly overestimated for all three traits, and for ST3, h^2 was significantly greater than zero, a result not corroborated by the ideal values.

Visual comparison of the reconstructed sibships with the true pedigree showed that there was considerable homology between reconstructed and true full-sib families. Nevertheless discrepancies were common and errors included both splitting of true full-sibships among two or more families, and inclusion of non-full sibs into partitioned families. In the reconstructed ALL pedigree, many partitioned families contained multiple true-sib groupings. This amalgamation of separate true families resulted in a smaller number of larger families in the reconstructed pedigree as compared to the true pedigree (Table 4). In contrast, the reconstructed pedigrees for all three pure strain data sets (B, G and O) had larger numbers of smaller families than the corresponding true pedigrees. These sources of error in the pedigree reconstruction are reflected by the comparatively low accuracy scores for all four reconstructed pedigrees (Table 4). Note that the negative score obtained for ALL and O show that in these cases more of the reconstructed full-sib relationships were false than true. Nevertheless, in all cases accuracy was found to be considerably (and significantly) higher than would be expected under random assignment of individuals to families (Table 4). This was also true for half-sib error. Furthermore, half-sib error explained the majority of incorrectly partitioned full-sib relationships in the reconstructed pedigrees, actually accounting for all of them in data set B (Table 4).

4. Discussion

The results of our analyses demonstrate the utility of marker-assisted approaches for the estimation of quantitative genetic parameters with unknown pedigree. In many instances, detection of significant variance and covariance components might be sufficient to test biological hypotheses regarding the genetic basis

https://doi.org/10.1017/S0016672302006055 Published online by Cambridge University Press

of phenotype. In the complete data set, both the regression-based method, and the use of reconstructed pedigree information, were successful in this respect. Specifically, both approaches detected significant heritabilities for all traits, a significant positive genetic correlation between the ST3 and ST4, and significant negative genetic correlations between weight and both spawning time traits. These qualitative findings were corroborated by the ideal results obtained from the known pedigree. Nevertheless, differences between the two approaches are apparent from a quantitative comparison of the results.

In general the regression-based method was the less successful of the two marker-assisted approaches employed, a finding that is consistent with previous comparative studies based on both simulated and empirical data (Thomas & Hill, 2000; Thomas *et al*., 2000). For example, heritability estimates exhibited large upward bias, with all values being greater than one. Thus these estimates actually lie outside of the true parameter space, a result that presents a challenge for biological interpretation. Furthermore confidence intervals were large for all estimated parameters. This low accuracy and precision reflects both bias and high sampling variance that are often associated with the estimation of pairwise relatedness (Van de Casteele *et al*., 2001). Here, underestimation of pairwise relatedness is more pronounced for pairs with true r_{ij} of 0·25 and 0·125 than it is for unrelated pairs, an effect that will likely contribute to the observed upward bias in h^2. While some bias is expected as an inherent property of the pairwise relatedness estimator (Ritland, 1996*a*; Ritland, 2000), underestimation of r_{ij} can also result from high levels of relatedness among the individuals used to estimate population allele frequencies (e.g. Hansson *et al*., 2000). In the current work population allele frequencies were estimated from the data set itself, which is known to include large numbers of relatives. Since estimation of heritability (but not genetic correlation) under the regression-based model requires division by the actual variance of relatedness (Ritland, 1996*b*), underestimation of this parameter will also introduce upward bias. This has been reported elsewhere (Thomas *et al*., 2002), though in this case estimated actual variance of relatedness is more than the true value as determined from the known pedigree. Nevertheless, the large confidence limits associated with the regression-based h^2 estimates may reflect low precision in estimating the actual variance of relatedness. This is a source of error that does not affect the estimation of genetic correlations.

Genetic parameters estimated from the reconstructed pedigree show closer agreement with ideal values than those obtained using the regression method. The downward bias of all estimates relative to the ideal values can be attributed to errors in pedigree reconstruction and to unrecognized relatedness between full-sib families. The number of incorrectly assigned full-sibling relationships is high (actually exceeding the number correctly assigned). Despite the use of considerably more genotypic information in this case, accuracy of the reconstructed pedigree is much lower here than has been reported elsewhere (e.g. Thomas & Hill, 2000; Smith *et al*., 2001). This can be attributed to the presence of half-sibs in the pedigree structure that clearly pose problems to the pedigree reconstruction, and account for most of the incorrectly assigned full-sibships. While a full-sibling relationship will not be assigned if multi-locus genotypes cannot be attributed to a single parental pair, this constraint is not expected to exclude all true half-sib pairs. Elsewhere, simulations following the extension of the MCMC approach to include nested full- within half-sib families, have suggested the reconstruction methods to be inherently conservative such that true half-sibs are unlikely to be reconstructed as full-sibs (Thomas & Hill, 2002). However this conclusion was based on an assumed structure of pairs of maternal full-sib families nested within paternal half-sib families, while in this case the half-sib structure is considerably more complex and extended. Furthermore this type of error is made more likely here because the number of true half-sib pairs present in the data set is an order of magnitude greater than the number of true full-sib pairs. Any incorrectly assigned full-sib pairs will depress true relatedness within reconstructed families, resulting in downward bias of trait heritabilities. However less bias is introduced if incorrectly assigned full-sib pairs are actually true half-sibs (as opposed to unrelated individuals). Thus, while the presence of half-sibs poses a challenge to pedigree reconstruction that results in low accuracy, the high level of half-sib error in this case also tempers the effect of this low accuracy on parameter estimates.

Splitting true full-sibling families into multiple groups represents a second type of error in pedigree reconstruction. This will also contribute to downward bias in estimated genetic parameters because reconstructed families are assumed to be unrelated. A tendency for reconstructed pedigrees to underestimate family size has been reported and attributed to this type of error (Thomas & Hill, 2000; Smith *et al*., 2001). In the present study, mean family size was actually larger in the reconstructed pedigree than in the true pedigree, though some instances of true families being partitioned into two or more groups were seen. Furthermore, even in the complete absence of this type of error, unrecognized relatedness between families is unavoidable due to the fact that single sires were used to produce multiple families. Therefore with half-siblings present and no parental information available, some downward bias of genetic parameters is expected even if full-sib families are reconstructed without error.

In the pure-strain subsets of the data, insufficient sample size was the primary constraint on effective estimation of trait heritabilities and genetic correlations. Even when the true pedigree was available, difficulties were encountered in generating meaningful estimates of genetic parameters, and this largely prevents the evaluation of marker-assisted estimation procedures in these data subsets. Sample size is critical to both the accuracy and precision of estimated genetic parameters (Falconer & Mackay, 1996), and this constraint applies to both marker-assisted and conventional procedures. In natural populations, larger sample sizes might be possible using the regression model since, with no pedigree structure assumed, application to samples drawn from multiple generations is valid. This consideration might be particularly important in systems with overlapping generations, or for organisms in which age determination is difficult. However, in the current work the pedigree reconstruction method was superior despite the larger sample sizes under the regression model. Furthermore, under the regression model, results were similar (both qualitatively and quantitatively) if parental individuals were excluded from the data set (results not shown).

In addition to the direct effects of sample size, features of the pure-strain data subsets have further implications to marker-assisted methodologies in particular. For example, smaller sample sizes will result in less reliable estimates of population allele frequencies, a problem exacerbated by the presence of relatives in the sample. While this latter point applies to the full data set (as discussed above), B, G, and O contain higher proportions of known relatives than does ALL. Furthermore, the set of microsatellite marker loci used herein is more informative for the complete data set, because allelic diversity is less in the pure-strain subsets (presumably as a consequence of hatchery management practices). It should be noted that for estimation of pairwise relatedness, loci provide information roughly in proportion to the number of alleles at each locus (Ritland, 1996*a*), while the accuracy of pedigree reconstruction is expected to increase with allelic diversity (Thomas & Hill, 2000; Smith *et al*., 2001).

These features of the data result in decreased accuracy and precision of the relatedness estimates in the pure-strain data subsets. It is noticeable that performance of the pairwise relatedness estimator is worst in data set O which has the lowest sample size, the highest proportion of pairwise relationships with true r_{ij} greater than zero, and the lowest mean number of alleles per locus. The particularly poor performance in this strain is consistent with previous findings based on the relatedness estimator of Queller & Goodnight (1989) (McDonald, 2001). In contrast, the accuracy of pedigree reconstruction was actually higher for B, G and O than it was for ALL, suggesting that this method may be less vulnerable to the difficulties posed in the smaller data sets. Nevertheless it is noticeable that reconstructed pedigrees for B, G and O had larger numbers of families with smaller mean sizes that the true pedigrees. This effect has been attributed to splitting of large families into smaller ones, a partitioning error that occurs in particular when estimates of population allele frequencies are made from a target sample containing a few large families (Smith *et al*., 2001).

The application of marker-assisted approaches to estimate quantitative genetic parameters is dependent upon the relationship structure of the system examined. In particular there is a requirement for a large variance of relationship in the sample (Ritland, 1996*b*; Thomas *et al*., 2000). This will occur when the sample contains high numbers of related individuals. In the current work the presence of family structure within the data sets was known *a priori*, but in natural populations this will not typically be the case. For example, Thomas *et al*. (2002) attributed disappointing results in estimating heritability of body weight in Soay sheep to a lack of relatedness structure in the sample, as well as to insufficient amounts of genotypic information. Application to natural populations therefore requires careful consideration of biological processes such as social grouping (Ritland, 2000), as well as appropriate design of sampling strategy (Wilson & Ferguson, 2002), such that significant numbers of relatives are sampled. Since such structure can lead to bias and inaccuracy in the molecular pedigree analysis (discussed above), efficient estimation of genetic parameters in small, highly structured samples may require estimation of allele frequencies from a larger sample more representative of the whole population. Alternatively, approaches such as the iterative process for estimating allelic frequencies described in Smith *et al*. (2001) may partially correct that problem.

In conclusion, we find that both the regression-based method, and use of pedigree reconstruction allow useful analysis of quantitative genetic architecture in a qualitative sense. While the regression-based method requires fewer assumptions regarding family structure, and can also be applied across generations, this generality was associated with low accuracy and precision of estimated genetic parameters. However, superior estimates were obtained using pedigree reconstructed by the MCMC method, despite low accuracy of pedigree reconstruction that was caused to a large extent by failure to distinguish between full- and half-sibling relationships. Even if more genotypic information was used so as to reduce this error rate, some downward bias of genetic parameters is expected in the presence of half-siblings due to unrecognized relatedness between families. Nevertheless, in this case reconstructing pedigree under the assumption that only full-sib family structure was present resulted in estimates of quantitative genetic parameters very similar to those

obtained when the true complex pedigree was known. Simulation-based studies would provide a useful approach for further testing of the generality of this result. As many natural populations will likely contain a complex mix of relationship classes, extension of the MCMC methodology to include these classes also represents a useful direction for research, though it is likely to be a complex problem (Thomas & Hill, 2002). Thus the results herein are particularly encouraging in that currently available methodology might be applied to natural populations in which a mix of both full- and half-sibs is possible or even likely.

This research was supported by NSERC Research grants to MMF and CMH. We are grateful to Cheryl Quinton for providing access to data, Philippe Fullsack for much appreciated help with programming, and to two anonymous reviewers for helpful comments on the manuscript.

References

Almudevar, A. & Field, C. (1999). Estimation of single generation sibling relationships based on DNA markers. *Journal of Agricultural, Biological, and Environmental Statistics* **4**, 136–165.

Bardakci, F. & Skibinski, D. O. F. (1994). Applications of the RAPD technique in tilapia fish; species and subspecies identification. *Heredity* **73**, 117–123.

Cadée, N. (2000). Genetic and environmental effects on morphology and fluctuating asymmetry in nesting barn swallows. *Journal of Evolutionary Biology* **13**, 359–370.

Cheverud, J. M. (1988). A comparison of genetic and phenotypic correlations. *Evolution* **42**, 958–968.

Danzmann, R. G. (1997). PROBMAX: A computer program for assigning unknown parentage in pedigree analysis from known genotypic pools of parents and progeny. *Journal of Heredity* **88**, 333.

Falconer, D. S. & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*. 4th edn. New York: Longman.

Fishback, A. G., Danzmann, R. G., Sakamoto, T. & Ferguson, M. M. (1999). Optimization of semi-automated microsatellite multiplex polymerase chain reaction systems for rainbow trout (*Oncorhynchus mykiss*). *Aquaculture* **172**, 247–255.

Groeneveld, E. (1994). REML VCE – a multivariate multimodel restricted maximum likelihood (co)variance component estimation package. In: Proceedings of an EC symposium on application of mixed linear models in the prediction of genetic merit in pigs. Ed. Groeneveld, E.

Groeneveld, E. & Kovac, M. (1990). A generalized computing procedure for setting up and solving mixed linear models. *Journal of Dairy Science* **73**, 513–531.

Hansson, B., Bensch, S., Hasselquist, D., Lillandt, B. G., Wennerberg, L. & Von Schantz, T. (2000). Increase of genetic variation over time in a recently founded population of great reed warblers (*Acrocephalus aruninaceus*) revealed by microsatellites and DNA fingerprinting. *Molecular Ecology* **9**, 1529–1538.

King, R. B., Milstead, W. B., Gibbs, H. L., Prosser, M. R., Burghardt, G. M. & McCracken, G. F. (2001). Application of microsatellite DNA markers to discriminate between maternal and genetic effects on scalation and behavior in multiply-sired garter snake litters. *Canadian Journal of Zoology* **79**, 121–128.

Kruuk, L. E. B., Clutton-Brock, T. H., Slate, J., Pemberton, J. M., Brotherstone, S. & Guinness, F. E. (2000). Heritability of fitness in a wild mammal population. *Proceedings of the National Academy of Sciences of the USA* **97**, 698–703.

Lynch, M. & Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Heredity* **80**, 218–224.

Lynch M. & Walsh B. (1998). *Genetics and analysis of Quantitative Traits*. Sunderland: Sinauer Associates.

Marshall, T. C., Slate, J., Kruuk, L. E. B. & Pemberton, J. M. (1998). Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology* **7**, 639–655.

McDonald, G. (2001). Relatedness determination and detection of spawning time QTL in rainbow trout (Oncorhynchus mykiss). M.Sc. thesis, University of Guelph, Guelph, ON.

Merilä, J., Przybylo, R. & Sheldon, B. C. (1999). Genetic variation and natural selection on blue tit body condition in different environments. *Genetical Research* **73**, 165–176.

Mousseau, T.A., Ritland, K. & Heath, D. D. (1998). A novel method for estimating heritability using molecular markers. *Heredity* **80**, 218–224.

Painter, I. (1997). Sibship reconstruction without parental information. *Journal of Agricultural, Biological, and Environmental Statistics* **2**, 212–229.

Queller, D. C. & Goodnight, K. F. (1989). Estimating relatedness using genetic markers. *Evolution* **43**, 258–275.

Quinton, C. D. (2001). Growth rate and spawning time in diallele crosses of three strains of rainbow trout (*Oncorhynchus mykiss*). M.Sc. thesis, University of Guelph, Guelph, ON.

Qvarnström, A. (1999). Genotype-by-environment interactions in the determination of the size of a secondary sexual character in the collared flycatcher (*Ficedula albicollis*). *Evolution* **53**, 1564–1572.

Riska, B., Prout, T. & Turelli, M. (1989). Laboratory estimates of heritabilities and genetic correlations in nature. *Genetics* **123**, 865–871.

Ritland, K. (1996*a*). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research* **67**, 175–185.

Ritland, K. (1996*b*). 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. *Molecular Ecology* **9**, 1195–1204.

Ritland, K. & Ritland, C. (1996). Inferences about quantitative inheritance based on natural population structure in the yellow monkeyflower, *Mimulus guttatus*. *Evolution* **50**, 1074–1082.

Smith, B. R., Herbinger, C. M. & Merry, H. R. (2001). Accurate partition of individuals into full-sib families from genetic data without parental information. *Genetics* **158**, 1329–1338.

Thomas, S. C. & Hill, W. G. (2000). Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics* **155**, 1961–1972.

Thomas, S. C. & Hill, W. G. (2002). Sibship reconstruction in hierarchical population structures using Markov chain Monte Carlo techniques. *Genetical Research* **79**, 227–234.

Thomas, S. C., Pemberton, J. M. & Hill, W. G. (2000). Estimating variance components in natural populations using inferred relationships. *Heredity* **84**, 427–436.

Thomas, S. C., Coltman, D. W. & Pemberton, J. M. (2002). The use of marker-based relationship information to estimate the heritability of body weight in a natural population: a cautionary tale. *Journal of Evolutionary Biology* **15**, 92–99.

Van de Casteele, T., Galbusera, P. & Matthysen, E. (2001). A comparison of microsatellite-based pairwise relatedness estimators. *Molecular Ecology* **10**, 1539–1549.

Weigensberg, I. & Roff, D. A. (1996). Natural heritabilities: can they be reliably estimated in the laboratory? *Evolution* **50**, 2149–2157.

Wilson, A. J. & Ferguson, M. M. (2002). Molecular pedigree analysis in natural populations of fishes: approaches, applications, and practical considerations. *Canadian Journal of Fisheries and Aquatic Sciences*. In press.