Accurate Partition of Individuals Into Full-Sib Families From Genetic Data Without Parental Information

Bruce R. Smith,* Christophe M. Herbinger† and Heather R. Merry*

**Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia B3H 3J5, Canada and †Marine Gene Probe Laboratory, Department of Biology, Dalhousie University, Halifax, Nova Scotia B3H 4J1, Canada*

Manuscript received July 14, 1998
Accepted for publication April 12, 2001

ABSTRACT

Two Markov chain Monte Carlo algorithms are proposed that allow the partitioning of individuals into full-sib groups using single-locus genetic marker data when no parental information is available. These algorithms present a method of moving through the sibship configuration space and locating the configuration that maximizes an overall score on the basis of pairwise likelihood ratios of being full-sib or unrelated or maximizes the full joint likelihood of the proposed family structure. Using these methods, up to 757 out of 759 Atlantic salmon were correctly classified into 12 full-sib families of unequal size using four microsatellite markers. Large-scale simulations were performed to assess the sensitivity of the procedures to the number of loci and number of alleles per locus, the allelic distribution type, the distribution of families, and the independent knowledge of population allelic frequencies. The number of loci and the number of alleles per locus had the most impact on accuracy. Very good accuracy can be obtained with as few as four loci when they have at least eight alleles. Accuracy decreases when using allelic frequencies estimated in small target samples with skewed family distributions with the pairwise likelihood approach. We present an iterative approach that partly corrects that problem. The full likelihood approach is less sensitive to the precision of allelic frequencies estimates but did not perform as well with the large data set or when little information was available (*e.g.*, four loci with four alleles).

WITH the development of numerous informative nuclear DNA markers, particularly microsatellites, there is a growing interest in the possibility of inferring relatedness among individuals when part or all of the pedigree information is missing. Paternity inference or parentage assignment using such polymorphic markers is becoming common in the study of natural populations (Double *et al.* 1997; Fitzsimmons 1998; Marshall *et al.* 1998; O'Reilly *et al.* 1998). However, these studies require parental data. Recently, Blouin *et al.* (1996), Painter (1997), Herbinger *et al.* (1997), Almudevar and Field (1999), and Thomas and Hill (2000) explored the possibility of reconstructing sibships from genetic data without parental information. The ability to reconstruct a pedigree and estimate kinship relationships among individuals in natural populations would have obvious applications to the management of small threatened wildlife populations. It would also be important in such fields as behavioral genetics, ecological genetics, and evolutionary genetics, where it would allow for the estimation of gene flow, mating behavior, or inclusive fitness of natural populations, for example.

The pairwise likelihood score approach developed by Herbinger *et al.* (1997) allows for the estimation of which pairs of individuals (dyads) are potentially related from the patterns of allele sharing. In some instances, that information is all that is needed. For example, that approach was used to avoid mating related individuals in various aquaculture breeding experiments (Herbinger *et al.* 1995). However, in most cases one is interested in further reconstruction of the pedigree of the population by allocating the individuals into various genetic groupings. Generally, the pedigree of a group of individuals of unknown relatedness will potentially include a variety of relationships (*i.e.*, full-sibs, half-sibs, cousins, parent-offspring, uncle-niece, unrelated, etc.). In theory, there could thus be an infinite number of genealogies (pedigree) giving rise to the observed pattern of shared alleles (Thompson 1991). However, there are many cases in which clusters of individuals are known or supposed to be a mixture of full-sibs or half-sibs (Reeve *et al.* 1990; Apostol *et al.* 1993). Similarly, Painter (1997) tested his approach on a group of peregrine falcons that were supposed to be an unknown mixture of full-sibs. Our article presents two methods for partitioning individuals into full sibships when no parental information is available and illustrates their accuracy on various real and simulated data sets.

MATERIALS AND METHODS

Partitioning algorithm: A full sibship configuration of N individuals is a partition of the individuals into full-sib families and is therefore equivalent to a partition of the set {1, . . .,

Corresponding author: Bruce Smith, Department of Mathematics, Statistics and Computer Science, Dalhousie University, Halifax, Nova Scotia B3H 3J5, Canada. E-mail: bsmith@mscs.dal.ca

$N\}$. If a configuration has K full-sib families, the partition is written as $c(1), \ldots, c(K)$, where $c(j)$ is the subset of $\{1, \ldots, N\}$ that identifies those individuals in the jth family. Any partition is equivalent to the associated collection of pairwise indicators c_{ij}, $i, j = 1, \ldots, N$, where $c_{ij} = 1$ specifies that i and j are full-sibs and $c_{ij} = 0$ that they are unrelated. The pairwise relations must be consistent. For example, if 1 and 2 are full-sibs and 1 and 3 are full-sibs, then 2 and 3 must be full-sibs. That is, $c_{1,2} = 1$ and $c_{1,3} = 1$ imply $c_{2,3} = 1$.

The space of all possible data configurations consisting only of full-sibs or unrelated individuals is denoted by $\mathcal{C}$. Let S be a function that assigns a score to each configuration $C \in \mathcal{C}$. We want to find a function S whose maximizing value provides a good estimate of the true configuration underlying a set of genotypic data. In addition, we want to have efficient methods for maximizing $S(C)$. One approach is direct enumeration, but the enormous size of $\mathcal{C}$, for even moderate N, precludes this method.

Another approach is to sample the space of configurations using a Monte Carlo approach. The Metropolis-Hastings algorithm is a general tool to sample from a state space, in this case $\mathcal{C}$. The idea is to define a Markov chain having stationary distribution $p(C)$, $C \in \mathcal{C}$. Where C_t denotes the tth configuration generated, the algorithm proceeds by simulating a candidate or proposal value C from a transition distribution $q(C_t, C)$. At the next step, C_{t+1} is randomly assigned to be either C with probability $r(C_t, C)$ or C_t with probability $1 - r(C_t, C)$, where

$$r(C_t, C) = \min\left(\frac{p(C)\,q(C, C_t)}{p(C_t)\,q(C_t, C)}, 1\right).$$

For appropriate choices of q this algorithm is guaranteed to generate samples from $p(C)$ for t large. Conditions under which this is the case are discussed in Hastings (1970).

In the examples presented, we set the initial configuration C_0 to "all unrelated" in which there are N families each containing one individual. For the distribution $q(C_t, C)$, we choose two individuals I and J independently according to a uniform distribution on $\{1, \ldots, N\}$. Let c_I and c_J denote the full-sib groups in C_t to which I and J belong. If $c_I \neq c_J$, then the proposed configuration C is obtained by moving individual I from c_I to c_J. If $c_I = c_J$ then I is removed from c_I to a new full-sib group of size one. This choice of q satisfies the necessary conditions under which the algorithm will generate samples from the desired distribution $p(C)$. It also ensures that $q(C_t, C) = q(C, C_t)$, in which case the Metropolis-Hastings algorithm is the original Metropolis algorithm (Metropolis *et al.* 1953).

We experimented with two distributions $p(C)$ on the configuration space. The first, denoted by $p_S(C)$, is based on a combination of pairwise likelihood ratios. Let g_i be the genotype of individual i, and $c_{ij} \in \{0,1\}$ denote the relationship between individuals i and j, restricting the possible pairwise relationships to "full-sib" ($c_{ij} = 1$) or "unrelated" ($c_{ij} = 0$). Let $P(g_i, g_j|c_{ij})$ be the probability of the genotypes of individuals i and j given their relationship and the population allele frequencies. The pairwise likelihood ratio for i and j is

$$\log \frac{P(g_i, g_j|c_{ij})}{P(g_i, g_j|1 - c_{ij})}. \tag{1}$$

If c_{ij} specifies the true relationship between i and j this should be relatively large, and otherwise small, so that its magnitude can be taken as evidence for the proposed relationship c_{ij}. Assuming that the parents of i and j are unrelated and noninbred, the probabilities in (1) can be derived in a straightforward fashion, as in Thompson (1991), and are functions of the (usually unknown) allele frequencies. We define a pairwise likelihood score for configuration C as

$$S(C) = \sum_{i \neq j} \log \frac{P(a_i, a_j|c_{ij})}{P(a_i, a_j|1 - c_{ij})}. \tag{2}$$

When C specifies the relationship between i and j correctly the associated summand will be positive, on average, while if the relationship is incorrectly specified, the summand will typically be negative. Therefore a configuration maximizing $S(C)$ should be a reasonable estimate of the true underlying configuration. The score function S gives rise to the family of probability distributions

$$p_S(C) = Ke^{S(C)/T}, \quad C \in \mathcal{C},$$

where K is a normalizing constant and T parameterizes the distribution.

For fixed T the Hastings-Metropolis algorithm is used to generate samples from p_S, with T governing the rate at which $\mathcal{C}$ is sampled. As $t \to \infty$, the algorithm is guaranteed to sample all of $\mathcal{C}$ for any fixed T, with new configurations being accepted more frequently for larger values of T. However, sampling with fixed T may not be the most efficient method of finding the configuration that maximizes S. An alternative is to apply a stochastic optimization method. One such algorithm is simulated annealing (Kirkpatrick *et al.* 1983), in which T (the annealing temperature) is allowed to decrease to 0 as $t \to \infty$. Simulated annealing has the desirable theoretical property that it samples only from the set of global maximizers of $p(C)$ as $T \downarrow 0$.

The second distribution $p(C)$ that we considered is the full joint distribution of the observed alleles given the configuration C [proportional to the likelihood $L(C)$ of the configuration], conditional on the allele frequencies. This approach was extensively investigated by Painter (1997). Where $g_m(j)$ and $g_p(j)$ are the maternal and paternal genotypes for the jth full-sib group, $c(j)$ is the collection of offspring in the jth full-sib group, $g_i(j)$ is the observed genotype of the ith individual in the jth full-sib group, and p denotes the unknown allele frequencies, we set down the single-locus likelihood for a configuration consisting of K full-sib groups as

$$\prod_{j=1}^{K} \sum_{g_m(j)} \sum_{g_p(j)} \left(\prod_{i \in c(j)} P(O_i(j)|g_m(j), g_p(j)) \right) P(g_m(j)|p) P(g_p(j)|p). \tag{3}$$

With several unlinked loci, the likelihood is a product of such terms over loci.

Where relationships are restricted to full-sib or unrelated, the single-locus single-family likelihood can be written down directly as a polynomial function of the allele frequencies, which effects a substantial computational saving. These single-family likelihoods are given in Table 1 for each of the 14 possible single-locus genotype configurations of a full-sib family. For example, if a family consists of n_{AA} individuals of genotype AA and n_{BB} individuals of genotype BB, the single-locus likelihood of the genotypes is $4p_A^2p_B^2(1/4)^n$, where $_{n =} n_{AA} + n_{BB}$. The derivation of these formulas is straightforward and the results have appeared elsewhere, for example in Table A.2 of Painter (1997). In that case, however, there are misprints in his entries 10 and 13.

In our study we restricted ourselves to the analysis of data sets without mutations or scoring errors, and so all proposed genotype configurations are required to be feasible in that they are compatible with the genotypes of two parents followed by independent segregation of alleles to offspring.

Data sets and goodness-of-fit of estimated partitions: Four measures were used to describe the fit between true and predicted full-sib groups. The number of moves is simply the smallest number of individuals that need to be moved from their predicted full-sib groups to their real full-sib groups to get a perfectly matched configuration. This number of moves is identical to the number of individuals that must be removed

TABLE 1

Single-locus full-sibship likelihoods

Full-sib genotypes	Sibship likelihood
AA	$p_A^4 + 4p_A^3p_X(1/2)^n + 4p_A^2p_X^2(1/4)^n$
AB	$2[p_A^2p_B^2 + (1/2)^{n_{AB}}\,2(p_A^3p_B + p_A^2p_Bp_X + p_Ap_B^3 + p_Ap_B^2p_X) + (1/4)^{n_{AB}}\,4(p_A^2p_Bp_X + p_Ap_B^2p_X + p_Ap_Bp_X^2)] + (1/2)^{n_{AB}}\,4p_A^2p_B^2$
AA BB	$4p_A^2p_B^2(1/4)^n$
AA AB	$4p_A^3p_B(1/2)^n + 4p_A^2p_B^2(1/4)^{n_{AA}}(1/2)^{n_{AB}} + 8p_A^2p_Bp_X(1/2)^n$
AA AB BB	$4p_A^2p_B^2\,(1/4)^{n_{AA}+n_{BB}}\,(1/2)^{n_{AB}}$
AA BC	$8p_A^2p_Bp_C(1/4)^n$
AB AC	$4p_A^2p_Bp_C(1/2)^n + 8p_A^2p_Bp_C(1/4)^n + 8p_Ap_Bp_Cp_X(1/4)^n$
AA AB AC	$8p_A^2p_Bp_C(1/4)^n$
AA AB BC	$8p_A^2p_Bp_C(1/4)^n$
AB AC BC	$8p_Ap_Bp_C(1 - p_X)(1/4)^n$
AA AB AC BC	$8p_A^2p_Bp_C(1/4)^n$
AC BD	$8p_Ap_Bp_Cp_D(1/4)^n$
AD AC BC	$8p_Ap_Bp_Cp_D(1/4)^n$
AC AD BC BD	$8p_Ap_Bp_Cp_D(1/4)^n$

n_{AB} is number of individuals with genotype *AB*; n is total number of individuals in the full sibship; p_X is the probability of all unspecified alleles. For example, for the family consisting of n_{AA} individuals with genotype *AA*, $p_X = 1 - p_A$, and for the second last full-sib genotype configuration listed (*AB, AC*), $p_X = 1 - p_A - p_B - p_C$.

from the true and estimated configurations to make them agree. The number of block moves is a refinement of that notion, which distinguishes the movement of N individuals that were incorrectly assigned to N different groups from moving a block of N individuals that were not assigned to their correct group but were nonetheless correctly predicted to be sibs. For example, consider the case of two full-sib families (A's and B's) of 10 individuals each:

True configuration: (A A A A A A A A A A)
(B B B B B B B B B B)

Prediction 1: (A A A A A A A A A B)
(B B B B B B B B A) (A) (B)

Prediction 2: (A A A A A A A A A A)
(B B B B B B) (B B B B)

In each prediction four moves are needed to get to the correct grouping. However, prediction 2 is only one block move from the real configuration, which reflects that this prediction uncovered more of the full-sib structure than the first prediction.

Finally, the number of full-sib pairs incorrectly classified as unrelated pairs (E_1) and the number of unrelated pairs incorrectly classified as full-sib pairs (E_2) distinguish between these two types of error. In the prior example, with 20 individuals for two full-sib families of 10, there are 190 ($20 \times 19/2$) dyads (pairs) of which 90 are full-sib pairs and 100 are unrelated pairs. In prediction 1, 34 of the 90 true full-sib pairs are incorrectly classified as unrelated (proportion $E_1 = 0.378$) while 16 of the 100 true unrelated pairs are incorrectly classified as full-sibs (proportion $E_2 = 0.16$). In prediction 2, 24 of the 90 true full-sib pairs are incorrectly classified as unrelated (proportion $E_1 = 0.276$) while none of the 100 true unrelated pairs are incorrectly classified as full-sibs (proportion $E_2 = 0$).

The following data sets were used in this study to illustrate the performance of the partitioning algorithms.

Example 1: This is a large data set consisting of 759 Atlantic salmon comprising 12 families typed at four microsatellite loci, with 11, 14, 10, and 8 alleles per locus in the offspring. Specific information on the fish, and particularly on parentage assignment to these offspring based on parental DNA information, can be found in O'Reilly *et al.* (1998) while details of the microsatellites used in this study can be found in O'Reilly *et al.* (1996). The parental genotypes and family sizes are listed in Table 2, but this parental information was not used in our estimation of the family configuration. The empirical allele frequencies of the offspring were used as estimates of the allele frequencies, which are required to calculate the likelihood and pairwise likelihood score.

Example 2: This data set consists of genotypes at nine loci for nine peregrine falcons. Painter (1997) carried out a full-likelihood analysis for these data, with a variety of assumptions on the population allele frequencies, and an analysis in which the likelihood is integrated over a distribution on the population allele frequencies. The resulting marginal distribution was maximized approximately using similar algorithms to those described above. For this data set we sample both the likelihood and the pairwise likelihood score (with $T = 10$) using sample allele frequencies and running for 100,000 iterations.

Example 3: Several large-scale simulation studies were carried out to better assess the sensitivity of the procedures to various data configurations. A factorial design was set up to investigate the importance of the number of loci, the number of alleles at each locus, the distribution of alleles at each locus, and the independent knowledge of population allele frequencies. Three levels were used for the number of loci (two, four, or eight), two levels for number of alleles per locus (four or eight), and three levels for allele distribution (equifrequent, nonequifrequent, or mixed). The equifrequent allele distribution assigns equal probability to each allele at a locus. The nonequifrequent allele distribution with K alleles assigns the relative weights of 1, 2, 3, . . . , K to the alleles at a locus, and the mixed-allele distribution assigns an equifrequent distribution to one-half of the loci and a nonequifrequent distribution to the other one-half. The fourth factor has two levels indicating the use of independent population allelic frequencies or allelic frequencies estimated on the target sample.

Because of the computational requirements of the simulation study, we restricted our simulations to 50 individuals each and ran the sampling algorithms for 30,000–100,000 iterations. The optimal configuration was taken to be that which maximized S_C or the full likelihood over the prespecified number of iterations. For a given combination of number of loci, number of alleles, and allele distribution, five unrelated families of 10 full-sibs each were generated by first randomly generating five pairs of parents according to the stipulated genotype distribution and then randomly sampling the parental alleles according to the rules of Mendelian inheritance. The data were then fit by the algorithm described, leading to two estimated family configurations, one assuming knowledge of the population allele frequencies and the other based solely on sample allele frequencies. The process was repeated 100 times at each combination of design parameters, leading to a $3 \times 2 \times 3 \times 2$ factorial simulation design, with 100 replicates in each cell.

A second set of simulations was carried out to assess the effect of family distribution. To reduce the overall computational burden the number of loci was fixed at four, as the

TABLE 2

Parental genotypes and family sizes of the Atlantic salmon in Example 1

Family	No.	M/F	*Ssa202*	*Ssa171*	*Ssa197*	*Ssa85*
1	51	F	(296, 308)	(234, 234)	(203, 207)	(127, 133)
		M	(280, 300)	(244, 254)	(167, 167)	(127, 133)
2	31	F	(272, 296)	(240, 246)	(167, 175)	(131, 133)
		M	(248, 280)	(240, 240)	(171, 183)	(127, 133)
3	54	F	(304, 328)	(240, 244)	(167, 175)	(133, 133)
		M	(280, 328)	(238, 244)	(175, 183)	(111, 129)
4	64	F	(272, 300)	(234, 240)	(175, 183)	(133, 137)
		M	(280, 296)	(258, 262)	(183, 191)	(127, 131)
5	59	F	(256, 272)	(236, 240)	(163, 171)	(125, 133)
		M	(300, 300)	(234, 240)	(167, 199)	(111, 127)
6	91	F	(300, 300)	(234, 242)	(171, 175)	(111, 133)
		M	(280, 328)	(234, 246)	(171, 175)	(127, 127)
7	69	F	(272, 300)	(234, 266)	(171, 183)	(125, 137)
		M	(280, 308)	(234, 246)	(175, 191)	(127, 133)
8	10	F	(276, 300)	(222, 234)	(175, 175)	(127, 133)
		M	(280, 292)	(234, 238)	(175, 175)	(127, 127)
9	8	F	(296, 304)	(234, 248)	(163, 207)	(127, 127)
		M	(280, 308)	(234, 244)	(175, 207)	(127, 133)
10	107	F	(296, 296)	(234, 244)	(183, 191)	(129, 133)
		M	(280, 308)	(246, 248)	(175, 175)	(127, 131)
11	140	F	(248, 308)	(236, 240)	(167, 187)	(127, 131)
		M	(300, 300)	(234, 244)	(167, 171)	(135, 137)
12	75	F	(300, 308)	(246, 276)	(167, 207)	(127, 133)
		M	(272, 276)	(236, 240)	(175, 175)	(129, 131)

No., family size; M, male parent; F, female parent.

first simulation indicated this to be sufficient to differentiate among full-sib families, at least with uniform family distributions and a moderately large number of alleles per locus. All simulations in this second set were carried out with equifrequent allele distribution, as the results of the first simulation showed that the effect of the type of allelic distribution was consistent but very small. Three family configurations (10, 10, 10, 10, 10), (20, 10, 10, 5, 5), and (30, 5, 5, 5, 5) were used, each having 50 individuals in five families. Other factors were as before.

Example 4: The simulation studies showed that with the pairwise likelihood score approach the estimated configuration is more accurate in many cases when population allele frequencies are known. Example 4 illustrates a way to improve the accuracy of the partitioning when population allelic frequencies are not known and allelic frequencies have to be estimated on the target sample itself. A partial solution to this problem is to estimate full sibship in an iterative way. The target group allelic frequencies are used in lieu of the true population allelic frequencies. The likelihood ratios are constructed and the partitioning algorithm is run, allowing estimation of sibship groups (step 1). Weighted allelic frequencies are then recalculated giving a weight of 1 to individuals that are estimated to belong to a family subgroup on their own and weight of $1/m$ to the m individuals that are estimated to belong to the same family subgroup. The likelihood ratios are reconstructed, and the partitioning algorithm is run again (step 2). Weighted allelic frequencies are further refined and the process repeats until the predicted partition stabilizes. This process is illustrated in example 4, where predicted partitions using both population allelic frequencies and iterative sample allelic frequencies are compared for a simulated family configuration of (30, 5, 5, 5, 5) generated with four loci and four equifrequent alleles at each locus. This family configuration was chosen to ensure bias in the allelic frequencies estimated on the sample alone.

RESULTS AND DISCUSSION

Example 1: We applied the methods described above to a large data set consisting of 759 individuals comprising 12 full-sib families, with data at four loci.

To illustrate the effect of the parameter T when sampling from $p_S(C)$, we ran the Markov chain Monte Carlo (MCMC) sampler three times using fixed T equal to 3, 10, or 30, running for 10^6 iterations in each case. As well, we ran the simulated annealing algorithm five times with an initial temperature $T_0 = 30$ and an annealing schedule that decreased T by 10% every 10^5 iterations, terminating the search on reaching $T = 0$ after 10^6 iterations. Each run began at the all-unrelated configuration having 759 full-sib groups of size 1. We carried out four runs in which we sampled from the full likelihood (as described in MATERIALS AND METHODS), of which three began in the all-unrelated configuration and one was started at the true configuration having 12 full-sib groups. In addition we ran two simulated annealing type runs described more fully below. In all cases we ran for 10^6 iterations. The results are summarized in Table 3, which shows the maximized values

TABLE 3

Error counts and optimization summaries for Example 1

Estimated configuration	Method	Optimized criterion	Iteration
(13, 2, 2, 135, 91)	$S(C)$ $T = 30$	1,523,081	734,507
(12, 2, 2, 135, 101)	$S(C)$ SA $T = 30$	1,522,899	308,452
		1,522,899	469,098
		1,522,899	532,671
	$S(C)$ $T = 10$	1,522,899	533,039
(13, 29, 3, 1134, 101)	$S(C)$ $T = 3$	1,522,864	495,801
	$S(C)$ SA $T_0 = 30$	1,522,864	734,234
(13, 27, 4, 790, 125)	$S(C)$ SA $T_0 = 30$	1,519,256	680,734
True configuration		1,522,637	
(17, 108, 7, 4130, 168)	ML SA $T_0 = 30$	−61.1	957,193
(24, 294, 14, 15701, 139)	ML	−415.6	226,453
(27, 372, 18, 19022, 96)	ML	−478.6	921,908
(28, 375, 17, 19062, 4)	ML SA $T_0 = 1$	−491.3	273,616
(25, 323, 15, 17470, 110)	ML	−509.1	762,518
(12, 22, 99, 197) at true	ML started	−27.8	362
(12, 0, 0, 0, 0)	True configuration	−89.0	

Optimized criterion is maximized value of log-likelihood or score iteration (iteration at which optimum was first encountered). Configuration is (no. f.s., no. moves, no. group moves, E_1, E_2), where no. f.s. is number of full-sib groups; no. moves is number of moves to correct configuration; no. g.m. is number of group moves; E_1 is number of full-sib pairs misclassified as unrelated; E_2 is number of unrelated pairs misclassified as full-sibs; ML is the maximum-likelihood method; $S(C)$ is the pairwise scoring method; and SA T is simulated annealing with starting temperature T.

found over 10^6 iterations as well as the iteration number at which the maximum was first achieved ("iteration"), the number of full-sib groups in the best configuration achieved ("no. f.s."), and the four error criteria evaluated at the best configuration. At the beginning of the process, no. f.s. = 759 and at the end no. f.s. should hopefully be 12, with all four error criteria optimally being 0.

Four of the five best runs judged on the basis of maximized $S(C)$ resulted in the same estimated configuration, differing by only two individuals from the correct configuration. In each of these four runs, the same individual was incorrectly assigned from family 7 to family 6, and a second individual was incorrectly assigned from family 7 to family 8.

In general, the configurations estimated using $S(C)$ are excellent, considering that this is a large real data set with information from only 4 loci and with very unbalanced family sizes, ranging from 8 individuals in family 9 to 140 individuals in family 11. Three of the estimated configurations are somewhat less accurate than the others, with 27 or 29 moves necessary to bring the estimate to the true configuration. However, in these cases one of the true-sib groups was split into two smaller groups of roughly equal size so that the number of group moves from the true configuration was still small. A comparison to the score of the true configuration (1,522,637) shows that all but one of the estimated configurations have higher scores. This suggests that the MCMC optimization procedure is doing a reasonable job in maximizing the score. On the other hand, it appears that there may be a large number of configurations whose scores exceed that of the true configuration, although these seem to be fairly close to the true configuration, at least in terms of number of group moves. It is also clear that the best configuration is not typically found until a very large number of iterations were performed, so that, for this data, 10^6 iterations may not be enough to provide an adequate coverage of the configuration space.

Of the five simulated annealing runs using $S(C)$, four terminated in the best configuration found during the prior 10^6 iterations. The fifth annealing run ended one move away from the best configuration found, with the addition of a 14th group consisting of a single individual. The annealing schedule that we have chosen is very simplistic, with temperature being decremented by a constant 10% every 10^5 iterations. Aarts and Van Laarhoven (1993) provide guidelines for the choice of an annealing schedule that has some desirable theoretical properties and may be an appropriate choice here.

The relationship between the two error types can be seen, for example, by considering the four runs with equal outcome, where one individual from group 7 (sample size $n = 69$) was improperly assigned to group 6 ($n = 91$) and another individual from group 7 was assigned to group 8 ($n = 10$). Thus the number of unrelated pairs incorrectly identified as full-sibs is $E_2 = 91 + 10 = 101$ and the number of full-sib pairs incorrectly identified as unrelated is $E_1 = 67 + 67 + 1 = 135$. These numbers are quite small considering that there are 31,588 full-sib pairs and 256,073 unrelated

pairs in the data set. Even in the worst run, the proportion of error type E_1 is only 3.5% (1134/31,588) while the proportion of error type E_2 is a very low 0.04% (101/256,073).

When carrying out a statistical analysis it is generally considered desirable to make inferences on the basis of the full joint distribution of the data (proportional to the likelihood) as opposed to some function of lower dimensional projections, such as the pairwise likelihood score $S(C)$. However, in the present case, sampling from the full joint distribution [equivalently the likelihood $L(C)$] did not provide accurate estimates of the true configuration and led to estimated numbers of full-sib groups equal to 24, 25, and 27 (entries labeled $T = 1$).

Our supposition is not that the likelihood is a bad criterion to optimize but rather that the likelihood sampler needs some tuning appropriate to the problem at hand. We considered an annealing version of the likelihood with acceptance probability

$$r(C_t, C) = \min\left(\left[\frac{L(C)}{L(C_t)}\right]^{1/T}, 1\right)$$

and made two likelihood-based annealing runs, one beginning at $T_0 = 1$ and a "heated" version beginning at $T_0 = 30$. The annealing schedule again reduced T by 10% after each 10^5 iterations. With $T_0 = 30$, configurations are initially accepted with much higher probability than for the basic likelihood sampler ($T = 1$) and this case provided the best-estimated configuration among all of the likelihood-based samplers started at the all unrelated configuration, although this wasn't achieved until nearly 10^6 iterations had been completed. These preliminary results suggest that the performance of the likelihood-based sampler might be improved with further tuning, either using a fixed $T > 1$ or an annealing schedule with $T_0 > 1$.

There was great deal of run-to-run variability in the maximized value of the log likelihood, which may be due to inherent roughness of the likelihood surface and/or the starting point used and/or the relatively small number of iterations carried out. In one run the log-likelihood sampler was started at the true configuration and quickly (after only 362 iterations) found a configuration with much higher log likelihood after which there was no additional improvement through 10^6 iterations. In this case, the estimated configuration had one individual from family 6 incorrectly assigned to family 10, and one individual from family 8 was incorrectly assigned to family 6. In a study of parentage assignment on these same salmon using the parental genotype information (O'Reilly *et al.* 1998), one offspring (omitted from the present data set) could not be unambiguously assigned to a single family and could have come from either family 6 or 8 because of the similarity of the genotypes in these two families.

Figure 1 shows the estimated number of full-sib groups by iteration (plotted at multiples of 10^5 iterations). Much of the grouping is done in the first 10^5 iterations (from 759 to <30 groups in these examples). While in theory the Hastings-Metropolis algorithms are guaranteed to generate samples from the full likelihood or from $p_S(C)$ as the number of iterations increases, the figure indicates that several hundred thousand iterations are required before this might be the case. When sampling from $p_S(C)$ the parameter T governs the rate of acceptance of new configurations, with increasing acceptance rate (and therefore a more thorough sampling of the configuration space) as T increases. This is suggested by the run at $T = 30$, which found the best configuration over any of our examples, but only after a very large number of iterations. For the annealing likelihood run at $T_0 = 30$, the sampled configurations at $t = 10^5, 2 \times 10^5, \ldots, 10^6$ had numbers of groups equal to 252, 232, 234, 211, 206, 193, 149, 155, 86, and 21, respectively, which, except for the last point, are off scale on the plot.

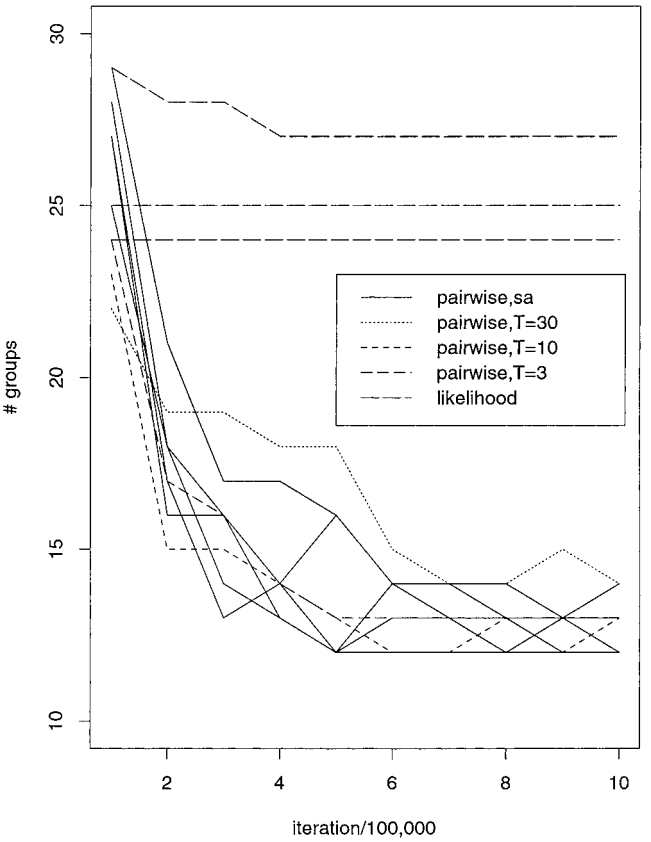

Figure 1.—Number of groups *vs.* iteration number for example 1. Pairwise, sampling from pairwise likelihood score; likelihood, sampling from full likelihood; sa, simulated annealing with annealing schedule decrementing T by 10% each 10^5 iterations.

Example 2: We sampled from both the likelihood and $p_S(C)$ (with $T = 10$) for the peregrine falcon data previously analyzed in Painter (1997), using empirical allele frequencies and running 100,000 iterations. The best 10 configurations found with each sampler are listed in Table 4. The best 5 ordered configurations are

TABLE 4

Ten best configurations found by pairwise likelihood and likelihood optimizations for the Falcon data of Example 2

Configuration	Log likelihood	Pairwise score S_C
(1 2)(3 4 7 8)(5)(6 9)	−150.1	430.3
(1 2)(3 4 7 8)(5)(6)(9)	−156.2	414.4
(1 2)(3 4 8)(5)(7)(6 9)	−156.5	411.5
(1 2)(3 4 7 8)(5 6)(9)	−156.6	401.1
(1 2)(3 4 7 8)(5 9)(6)	−157.0	397.5
(1 2)(3 4 7)(8)(5)(6)(9)	−167.4	
(1 2)(3 4 7)(8)(5 6)(9)	−167.8	
(1 2)(3 4 7)(8)(6)(5 9)	−168.1	
(1 2)(3 7)(4 8)(5)(6)(9)	−170.5	
(1 2)(3 7)(4 8)(5 6)(9)	−170.9	
(1)(2)(3 4 7 8)(5)(6 9)		397.0
(1 2)(3 4 8)(5)(6)(7)(9)		395.6
(1 2)(3 4 8)(5 7)(6 9)		383.8
(1 2)(3 4 8)(5 6)(7)(9)		382.3
(1)(2)(3 4 5 8)(6)(7)(9)		381.1

the same for each algorithm, and in both cases the best configuration is (1 2)(3 4 7 8)(5)(6 9), which means that birds 1 and 2 form a full-sib group, birds 3, 4, 7, and 8 form a second full-sib group, and so on. Painter (1997) found this to be the optimal configuration for a number of choices of allele frequencies sampled from a Dirichlet prior distribution. He found that the ordered posterior probability of configurations tended to be relatively insensitive to choice of prior. If this is typical for most data sets then there may be some justification in using empirical estimates of allele frequencies. Painter (1997) also carried out an analysis in which the joint distribution of configuration and allele frequencies was integrated over the prior distribution on the allele frequencies. The resulting marginal likelihood was sampled, and again the configuration (1 2)(3 4 7 8)(5)(6 9) was found to be optimal. This latter approach, while desirable when reasonable estimates of allele frequencies are unknown, adds the substantial computational burden of integrating over the prior distribution and restricts the method to relatively small problems.

Example 3: An initial simulation study was carried out to better assess the sensitivity of the procedures to the number of loci, the number of alleles, the allele distribution type at each locus, and the knowledge of independent population allelic frequencies. For each estimated configuration the number of moves to the true configuration was determined. The salient findings are summarized as follows (data not shown). As a general trend, with the pairwise likelihood score approach, configurations are more accurate when using independent population estimates of allelic frequencies than when using sample estimates. In contrast, accuracy is about equivalent with the full joint likelihood whether using independent population allele frequencies or sample estimates. The number of loci and the number of alleles per locus are the two factors that have the strongest impacts on the configuration accuracy. Using two loci is clearly insufficient in most cases. However, even with only two loci, but with eight alleles per locus, a few simulations with the pairwise likelihood score approach gave surprisingly good results, although there is considerable variability within each cell among the various replicated simulations. This indicates that the configuration accuracy based on such a low number of loci is quite unpredictable. At the other extreme, configurations based on eight loci are overall very good with both approaches, with little variability among the various cells. When each locus has 8 alleles, predicted configurations are almost always exact, independent of allele distribution, even when using sample allelic frequency estimates instead of independent population estimates with the pairwise likelihood score approach. Overall, there is a trade-off between the number of loci and the number of alleles per locus. We found that similar accuracy was attained with four loci each with 8 alleles per locus or with eight loci with 4 alleles per locus with the pairwise approach. This result is partly borne out by example 1, where excellent results were obtained with four loci having 8, 10, 11, and 14 alleles, respectively, with the pairwise approach. This is encouraging as it means that very good predictions can be attained with a reasonably low number of microsatellite loci, which typically have 6–12 alleles in most species. In contrast, when sampling from the full likelihood, better results were systematically obtained in simulations with four loci having 8 alleles each than with eight loci having 4 alleles. A small but fairly consistent trend was also observed concerning the effect of allelic distribution, better results being obtained with equifrequent allele distributions than with nonequifrequent, with the mixed distribution in between.

A second simulation was carried out to assess the effect of family distribution with the number of loci fixed at four and with equifrequent allelic distributions. The number of moves to correct configuration is displayed in Figure 2. The pairwise and full likelihood approaches have qualitatively different behavior. Not surprisingly, both approaches produce much better predictions with eight alleles per locus than with four. However, the pairwise likelihood score approach outperforms the full joint likelihood approach when only four alleles per locus are present, while the reverse is true when eight alleles per locus are available, particularly when using sample allelic frequencies. The pairwise likelihood-ratio approach performs better with independent (true) population allelic frequencies estimates than with sample estimates, with both four and eight alleles per locus. With increasingly skewed family distribution, the predictions worsen significantly when using sample estimates. In contrast, increasingly skewed family distributions produce a small improvement in the prediction accuracy when using independent population estimates with four alleles, while no changes were ob-

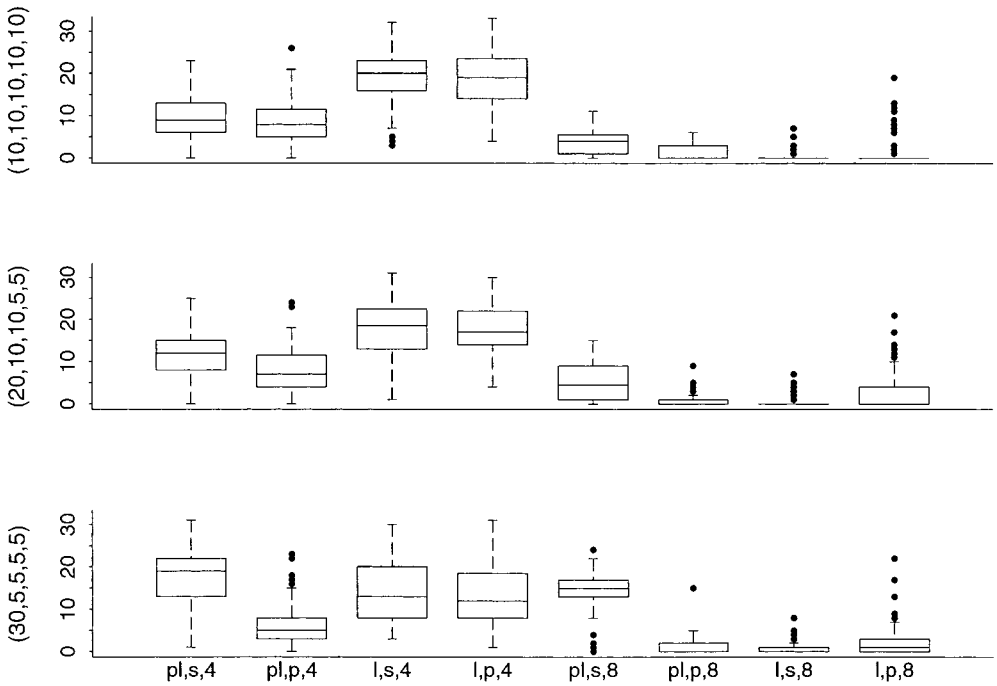

FIGURE 2.—Number of moves to correct configuration in the second simulation of example 2, with four loci in each case and three family distributions, (10, 10, 10, 10, 10) top, (20, 10, 10, 5, 5) middle, and (30, 5, 5, 5, 5) bottom. pl, pairwise likelihood; l, likelihood; s, sample estimate of allele frequencies, p, true population allele frequencies; 4, four equifrequent alleles per locus; 8, eight equifrequent alleles per locus.

served with eight alleles per locus, as predictions are then almost always exact.

The full joint likelihood approach performs about equally well when using independent population or sample estimates of allelic frequencies, which is consistent with PAINTER's (1997) finding for the falcon data that results were quite insensitive to estimates of allele frequencies. Results are overall fairly poor with loci having only four alleles, with a small improvement with increasingly skewed family distributions. On the other hand, results are almost always exact with eight alleles per locus, independent of family distribution.

It is apparent that both the pairwise likelihood-ratio approach when using the population allelic frequencies and the full likelihood method produce predictions that are fairly robust to the type of family distribution for the various combinations of number of loci and alleles per locus. On the other hand, with the pairwise likelihood approach when using allelic frequencies estimated on the target sample itself, the accuracy of the predictions worsens considerably as the family distribution departs from uniformity. This sensitivity of the configuration accuracy to family distribution when using sample allelic frequency estimates probably results from those estimates being increasingly biased and inaccurate as the family distribution departs from uniformity, particularly in the case of small samples. In the (30, 5, 5, 5, 5) family distribution, for example, 60% of the individuals come from one family, and at each locus this family makes a disproportionately large contribution to the estimate of allele frequencies, as all 30 individuals carry no more information than would the two unknown parents. For a dyad sharing the particular alleles from that family at most loci, the likelihood ratio of being full-sib *vs.* unrelated would be much lower if these alleles are very common (when frequencies are estimated in the target sample) than if the alleles are rarer (when independent population estimates are available). This obviously reduces the power to group together the various individuals from the common family.

It is also interesting to look at the average proportions of errors E_1 and E_2 that occurred in the various simulations (Table 5). The results are very consistent across every combination of factors, and so Table 5 presents only a portion of the cases to exemplify the findings. In this table every locus has eight equifrequent alleles and the variable factors are the number of loci (four or eight), the family distribution type [(10, 10, 10, 10, 10), (20, 10, 10, 5, 5) or (30, 5, 5, 5, 5)], and whether population or sample allelic frequencies were used. It is clear from Table 5 that the errors are predominantly of type E_1 rather than E_2; *i.e.*, true full-sib pairs are not always identified to belong to the same family group but truly unrelated pairs are very rarely construed to be full-sibs. Even when the classification does not perform very well, *e.g.*, when only sample allelic frequencies with nonuniform family distribution are available, the proportion of unrelated pairs falsely conjectured to be full-sibs remains quite low while the proportion of full-sibs construed to be unrelated is quite high. This shows that results from the proposed partitioning algorithm are essentially conservative. Individuals that are grouped together can safely be assumed to indeed be full-sibs, even though all sibs may not have been added to the group. These "other" sibs will very often be isolated in many single groups of size 1 or 2 or will be regrouped

TABLE 5

Average proportion of error types 1 and 2 in Example 3, when all loci have four or eight equifrequent alleles

		Pairwise likelihood				Likelihood			
		Sample		Population		Sample		Population	
na	Configuration	$p(E_1)$	$p(E_2)$	$p(E_1)$	$p(E_2)$	$p(E_1)$	$p(E_2)$	$p(E_1)$	$p(E_2)$
4	10, 10, 10, 10, 10	0.273	0.024	0.228	0.044	0.516	0.101	0.496	0.099
8	10, 10, 10, 10, 10	0.111	0.0002	0.044	0.0005	0.014	0.002	0.052	0.003
4	20, 10, 10, 5, 5	0.547	0.022	0.303	0.039	0.670	0.096	0.678	0.096
8	20, 10, 10, 5, 5	0.266	0.0006	0.032	0.0004	0.014	0.001	0.103	0.004
4	30, 5, 5, 5, 5	0.497	0.022	0.106	0.042	0.309	0.094	0.292	0.089
8	30, 5, 5, 5, 5	0.483	0.0006	0.012	0.002	0.032	0.002	0.062	0.004

$p(E_1)$ is the average proportion of errors of the first type; $p(E_2)$ is the average proportion of errors of the second type. na, number of alleles; sample, using sample allele frequencies; population, using population allele frequencies.

in another separate full-sib group as was the case in the three more pessimistic runs of example 1 using the pairwise likelihood score, where individuals from one true family were classified into two predicted groups.

Example 4: As shown in the previous example, the accuracy of the pairwise likelihood prediction worsens when allelic frequencies are estimated in samples with nonuniform family distribution. One solution might be to iterate the process, whereby sample allele frequencies are used to estimate a configuration on the basis of which new estimates of allele frequencies are made, and a new configuration is estimated. In example 4, a data set with a family configuration of (30, 5, 5, 5, 5) was created with individuals (1–30), (31–35), (36–40), (41–45), (46–50) belonging to the five families, respectively. The configuration

(1–30, 43) (36 37 38 39 40 48) (31 32 33 34 35)
(41 42 44 45) (46 47 49 50)

was estimated on the basis of known population allele frequencies with only individuals 43 and 48 misclassified. When using the sample allele frequencies, the following configuration was obtained:

(14 16 19 21 22 23 25 26 27 30) (2 7 9 10 12 37 38 39 40 48)
(5 6 15 17 18 20 24) (3 4 8 11 28 29) (31 32 33 34 35)
(1 46 47 49 50) (13 41) (42 44) (36) (43) (45).

It is clear that information on the population allele frequencies provides a much better partition in this case. Due to the imbalance in the true family configuration, the sample estimate of allele frequencies is heavily biased toward the alleles of the parents of family 1. One way to improve the estimated configuration is to improve the estimate of allele frequencies. Using the estimated sample configuration, new weighted estimates of allele frequencies were made and the algorithm was then rerun using these updated allele frequencies, again starting from the all unrelated configuration. The updated partition was given by

(2 7 9 10 12 13 14 16 19 22 23 25 26 27 30 40)
(3 4 6 8 11 15 17 18 20 24 28 29) (31 32 33 34 35 43)
(5 37 38 39 48) (1 46 47 49 50) (41 44 45) (36) (42).

The process was iterated a second time, leading to the partition

(2 5 7 9 10 12 13 14 16 19 21 22 23 25 26 27 30 40 48)
(1 3 4 6 8 11 15 17 18 20 24 28 29) (31 32 33 34 35)
(37 38 39) (41 44 45) (46 47 49 50) (36) (42) (43).

Each iteration resulted in an improved estimate, with 25, 19, and 18 errors, respectively, in the sample configuration and the first and second iterates. The first and second iterates had 12 and 13 errors, respectively, that were attributable to large single families not being joined to the largest family in the configuration. If group moves are counted as single errors, the error counts are reduced to 14, 8, and 6, which indicates that the estimates are often much better than the crude error count suggests. We expect that the iterative process will often improve sample estimates and that often one might wish to restart the estimation at the best current configuration to date.

In calculating the updated estimates, our weights are inversely proportional to estimated sibship size. An alternative approach was suggested by Thomas and Hill (2000), who used weighted least-squares estimates of allele frequencies, with weight matrix based on the current estimate of relationships among individuals.

CONCLUSIONS

We presented Monte Carlo algorithms to step through the space of family configurations and took as an optimal configuration that which maximized S_C or the full joint likelihood on the tour. As shown in the previous real and simulated examples, these approaches are quite successful at properly partitioning individuals into full-sib groups on the basis of the information pro-

vided by as few as four variable single-locus markers such as microsatellites, without any parental information. The full joint likelihood approach works as well with either independent (true) population estimates or sample estimates of allelic frequencies, but it performed rather poorly with the large data set (*e.g.*, 759 salmon) and when little information was available (*e.g.*, four loci with four alleles). The pairwise likelihood score approach is very fast and converged reliably in every situation that we tested. It also works surprisingly well when little information is available. However, it is particularly sensitive to the accuracy of the allelic frequency estimates. When using allelic frequencies estimated on the target sample itself, it can encounter problems if the sample is dominated by a few large families. In these cases, the pairwise approach produces conservative partitions where the large families are often split in smaller groups. A similar problem is noted by Thomas and Hill (2000) with their likelihood approach. We describe an iterative approach that partially improves accuracy when the family distribution seems heavily biased in the sample.

In the examples we considered we found the pairwise likelihood score approach to be very fast and quite accurate. The likelihood sampler is desirable from a theoretical standpoint but was sometimes slow to reach a neighborhood of the correct configuration. We are currently investigating a hybrid approach in which we use the pairwise likelihood method to generate the starting point for the likelihood sampler, and we anticipate that this will combine the merits of each approach.

Fortran versions of the pairwise likelihood and full likelihood programs are available for a Unix platform. PC versions are being developed and will be made available. Likelihood ratios for other sorts of relationships (such as half-sibs) can easily be calculated (Thompson 1991) and a similar partitioning approach is being developed to allow the reconstruction of more complex pedigrees. Other improvements being pursued include testing for the significance of the predicted partition (*i.e.*, could the predicted partition be simply the result of chance, particularly when uncovered family groups are small?) and testing its robustness to the presence of mutation or human errors in the DNA data.

We thank Patrick O'Reilly, University of Washington, Seattle, for permission to use his data and two referees whose suggestions greatly improved this article. This research was supported by a collaborative grant and a Network Centres of Excellence grant (MITACS) from the Natural Sciences and Engineering Research Council of Canada.

LITERATURE CITED

Aarts, E. H. L., and P. J. M. Van Laarhoven, 1993 Statistical cooling: a general approach to combinatorial optimization problems. Philips J. Res. **40:** 193–226.

Almudevar, A., and C. Field, 1999 Estimation of single-generation sibling relationships based on DNA markers. J. Agric. Biol. Environ. Stat. **4:** 136–165.

Apostol, B. L., W. C. Black IV, B. R. Miller, P. Reiter and B. J. Beaty, 1993 Estimation of the number of full sibling families at an oviposition site using RADP-PCR markers: applications to the mosquito *Aedes aegypti*. Theor. Appl. Genet. **86:** 991–1000.

Blouin, M. S., M. Parsons, V. Lacaille and S. Lotz, 1996 Use of microsatellite loci to classify individuals by relatedness. Mol. Ecol. **5:** 393–401.

Double, M. C., A. Cockburn, S. C. Barry and P. E. Smouse, 1997 Exclusion probabilities for single-locus paternity analysis when related males compete for matings. Mol. Ecol. **6:** 1155–1166.

Fitzsimmons, N. N., 1998 Single paternity of clutches and sperm storage in the promiscuous green turtle (*Chelonia mydas*). Mol. Ecol. **7:** 575–584.

Hastings, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57:** 97–109.

Herbinger, C. M., R. W. Doyle, E. R. Pitman, D. Paquet, K. A. Mesa *et al.*, 1995 DNA fingerprint based analysis of paternal and maternal effects on offspring growth and survival in communally reared rainbow trout. Aquaculture **137:** 245–256.

Herbinger, C. M., R. W. Doyle, C. T. Taggart, S. E. Lochmann, A. L. Brooker *et al.*, 1997 Family relationships and effective population size in a natural cohort of Atlantic cod (*Gadus morhua*) larvae. Can. J. Fish. Aquat. Sci. **54** (Suppl. 1): 11–18.

Kirkpatrick, S., C. D. Gelatt and M. P. Vecchi, 1983 Optimization by simulated annealing. Science **220:** 671–680.

Marshall, T. C., J. Slate, L. E. B. Kruuk and J. M. Pemberton, 1998 Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. **7:** 639–655.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, 1953 Equations of state calculations by fast computing machines. J. Chem. Phys. **21:** 1087–1091.

O'Reilly, P. T., L. C. Hamilton, S. K. McConnell and J. M. Wright, 1996 Rapid analysis of genetic variation in Atlantic salmon (*Salmo salar*) by PCR multiplexing of dinucleotide and tetranucleotide microsatellites. Can. J. Fish. Aquat. Sci. **53:** 2292–2298.

O'Reilly, P. T., C. M. Herbinger and J. M. Wright, 1998 Analysis of parentage determination in Atlantic salmon (*Salmo salar*) using microsatellites. Anim. Genet. **29:** 363–370.

Painter, I., 1997 Sibship reconstruction without parental information. J. Agric. Biol. Environ. Stat. **2:** 212–229.

Reeve, H. K., D. F. Westneat, W. A. Noon, P. W. Sherman and C. F. Aquadro, 1990 DNA "fingerprinting" reveals high levels of inbreeding in colonies of the eusocial naked mole-rat. Proc. Natl. Acad. Sci. USA **87:** 2496–2500.

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

Thompson, E. A., 1991 Estimation of relationships from genetic data, pp. 255–269 in *Handbook of Statistics*, edited by C. R. Rao and R. Chakraborty. Elsevier, Amsterdam/New York.

Communicating editor: M. W. Feldman