Parental assignment in fish using microsatellite genetic markers with finite numbers of parents and offspring

B. Villanueva*, E. Verspoor† and P. M. Visscher‡

*Scottish Agricultural College, West Mains Road, Edinburgh, UK. †FRS Marine Laboratory, Aberdeen, UK. ‡Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh, UK

Summary

Deterministic predictions for the proportion of offspring assigned to different numbers of parent-pairs are developed in order to investigate the power of microsatellite loci for parental assignment in fish species. Comparisons with stochastic simulation results show that predictions based on exclusion probabilities are accurate, provided that the number of parents involved in the crosses is large. Accounting for sampling of parents gave very accurate predictions for a small number of parents and a single biallelic locus. For large numbers of loci or large numbers of alleles per locus stochastic simulations are, however, the only available method to predict the power of assignment of a particular set of loci when the number of parents is small. Nine 5-allele loci or six 10-allele loci with equifrequent alleles, are sufficient for assigning, with certainty, parents to 99% of the fish resulting from either 100 or 400 crosses. Results simulating a set of highly polymorphic microsatellites developed for Atlantic salmon show that the four most informative loci are sufficient to assign at least 99% of the offspring to the correct pair with 100 crosses involving 100 males and 100 females. An additional locus is required for correctly assigning 99% of the offspring when the 100 crosses are produced with 10 males and 10 females.

Keywords DNA microsatellites, exclusion probability, fish, parental assignment, salmon.

Introduction

Selection programmes in animal breeding make use of information not only on the candidates for selection, but also on their relatives in order to increase the accuracy of selection and therefore selection responses. Ideally, individuals are identified uniquely when they are born and then the pedigrees of individual animals can be tracked across generations.

One of the most important impediments to applying effective selective breeding programmes for fish is that newborn individuals are too small to be tagged physically. Thus, selective programmes making use of family information have needed to keep families separated until the fish are large enough to be individually tagged (Doyle & Herbinger 1994). This is costly, limits the number of families available for selection and can induce environmental effects common to the members of the same family (Doyle & Herbinger 1994).

The problem of individual identification in fish species can be resolved by applying DNA-based genetic markers. Polymorphic markers have been used to assess parentage in many species (e.g. Avise 1994) and can be used to discriminate fish in mixed family groups (Doyle & Herbinger 1994). Fish from different families can be reared together in the same tank from hatching, or even at the egg stage. Subsequent genotyping of parents and offspring for particular loci allows assignment of offspring to parental pairs.

The most useful type of markers to assess genetic parentage are microsatellite DNA loci (e.g. O'Connell & Wright 1997), which have already been isolated and characterized in several fish species including salmon (Slettan *et al.* 1996; O'Reilly *et al.* 1998; Banks *et al.* 1999; Nelson & Beacham

Address for correspondence

B. Villanueva, Animal Breeding and Genetics Department, Animal Biology Division, SAC, Bush Estate, Penicuik, Midlothian EH26 0PH, UK. E-mail: b.villanueva@ed.sac.ac.uk

Accepted for publication 18 August 2001

1999), trout (Herbinger *et al.* 1995; Estoup *et al.* 1998; Khoo *et al.* 2000), carp (Crooijmans *et al.* 1997), turbot (Estoup *et al.* 1998), sea bass (Castilho & McAndrew 1998), sea bream (Perez-Enriquez *et al.* 1999) and tilapia (Lee & Kocher 1996). Furthermore, several studies have empirically used microsatellite loci to successfully reconstruct pedigrees in fish populations with families mixed from hatching (Herbinger *et al.* 1995; Estoup *et al.* 1998; O'Reilly *et al.* 1998; Herbinger *et al.* 1999; Perez-Enriquez *et al.* 1999; Norris *et al.* 2000). However, how many loci are needed and how informative they need to be for accurate parental identification is poorly understood. In this study, deterministic predictions for the power of microsatellites for parental assignment are developed and compared with stochastic simulation results. Predictions are developed for large (strictly infinite) numbers of parents and offspring and for finite numbers. The power of parental assignment is specifically investigated for a set of microsatellite loci developed in Atlantic salmon.

Materials and methods

In the situation considered here the matings, and therefore all possible parent-pairs, are known, and the problem is to assign parentage to individuals from the offspring generation. The question is how many loci are needed, and how informative (i.e. number of alleles/locus and allelic frequencies) they need to be, to obtain correct parentage assignments. Assumptions in the deterministic and stochastic models include the absence of mutation and measurement errors, unlinked marker loci and Hardy–Weinberg equilibrium.

Deterministic predictions

Assuming that there are $N + 1$ possible parent-pairs, we will calculate the expected proportion of progeny with one (the correct one), two (the correct one and one incorrect),..., $N + 1$ (the correct one and N incorrect) pairs assigned as parents.

Infinite numbers of parents and offspring

The expected proportion of offspring assigned to $n + 1$ parent-pairs in this case can be obtained as follows. First, all possible genotypes for parent-pairs and their probabilities are determined. Secondly, for each possible offspring genotype, the parent-pair genotypes that cannot be excluded (and the corresponding probability) are identified. Thirdly, the proportion of offspring and the number of parent-pairs assigned for each possible offspring genotype are computed given the allelic frequencies, the total number of offspring and the number of crosses. Finally, the proportions of offspring with the same number of parent-pairs assigned are added.

For example, for a single biallelic locus with allele frequencies p_1 and p_2, there are six possible parent-pair genotypes (shown in Table 1, with their probabilities) and three possible offspring genotypes (A_1A_1, A_1A_2 and A_2A_2, with probabilities p_1^2, $2p_1p_2$ and p_2^2, respectively). For progeny of genotype A_1A_1, three parent-pairs cannot be excluded ($A_1A_1 \times A_1A_1$, $A_1A_1 \times A_1A_2$ and $A_1A_2 \times A_1A_2$) with probability $p_1^4 + 4p_1^3p_2 + 4p_1^2p_2^2$. The non-excluded parent-pairs and corresponding probabilities for progeny of genotypes A_1A_2 and A_2A_2 are obtained in a similar way. Now, if the number of parent-pairs is $N + 1$ then a proportion p_1^2 of the offspring will be assigned to $(p_1^4 + 4p_1^3p_2 + 4p_1^2p_2^2)$ $(N + 1)$ parent-pairs, a proportion $2p_1p_2$ will be assigned to $(4p_1^3p_2 + 2p_1^2p_2^2 + 4p_1^2p_2^2 + 4p_1p_2^3)(N + 1)$ parent-pairs and a proportion p_2^2 will be assigned to $(4p_1^2p_2 + 4p^1p_2^3 + p_2^4)$ $(N + 1)$ parent-pairs. Thus for $p_1 = p_2 = 0.5$ and $(N + 1) = 100$, 50% of the progeny will be assigned to 56.25 parent-pairs and 50% of the progeny will be assigned to 87.50 parent-pairs.

With few loci, or with few alleles per locus, the number of possible parent-pair genotypes is small and therefore there will be considerable variation around these expected values, as sampling of parents is not accounted for. With a large number of loci, or with a large number of alleles per locus, the number of parent-pair genotypes can be enormous, making predictions following the approach described above difficult. For instance, for three loci, each with five alleles, the number of possible unique parent-pair genotypes is

Parent-pair genotypes	Probability	Number of parent-pairs	Number of offspring of genotype		
			A_1A_1	A_1A_2	A_2A_2
$A_1A_1 \times A_1A_1$	p_1^4	n_1	m		
$A_1A_1 \times A_1A_2$	$4p_1^3p_2$	n_2	m_{21}	m_{22}	
$A_1A_1 \times A_2A_2$	$2p_1^2p_2^2$	n_3		m	
$A_1A_2 \times A_1A_2$	$4p_1^2p_2^2$	n_4	m_{41}	m_{42}	m_{43}
$A_1A_2 \times A_2A_2$	$4p_1p_2^3$	n_5		m_{51}	m_{52}
$A_2A_2 \times A_2A_2$	p_2^4	n_6			m
	1.00	$N + 1$			

Table 1 Possible parent-pair genotypes, probabilities and numbers of parent-pairs for a single biallelic locus and numbers of offspring of different genotypes produce for each type of parent-pair.

5 697 000. In this situation, predictions for the power of assignment can be simplified by using the binomial distribution as it is described in the next section.

Infinite numbers of parents and offspring and many loci or many alleles per locus

For many loci, and/or many alleles per locus, the expected proportion of offspring assigned to different numbers of parent-pairs can be approximated from the binomial distribution. In this situation, the first step is to compute the exclusion probability, that here is the probability of a randomly chosen parent-pair being genetically excluded as parents of a randomly chosen offspring (and that parent-pair did not produced that offspring). The exclusion probability for a single locus l with k alleles (Q_l) is

$$Q_1 = 1 + 4\sum_{i=1}^{k} p_i^4 - 4\sum_{i=1}^{k} p_i^5 - 3\sum_{i=1}^{k} p_i^6 - 8\left[\sum_{i=1}^{k} p_i^2\right]^2 + 2\left[\sum_{i=1}^{k} p_i^3\right]^2 + 8\sum_{i=1}^{k} p_i^2 \sum_{i=1}^{k} p_i^3$$

where p_i is the frequency of the ith allele (A_i) (Jamieson 1965; Dodds *et al.* 1996). For multiple loci (L microsatellites), the combined (overall) exclusion probability can be calculated as

$$Q = 1 - \prod_{l=1}^{L} (1 - Q_l)$$

(Dodds *et al.* 1996).

Given Q, the next step is to calculate the expected number of offspring having $n + 1$ assigned (i.e. non-excluded) parent-pairs (the true pair and n incorrect pairs). For a sample of $N + 1$ possible parent-pairs, with one parent-pair being the correct one, the question is how many non-exclusions are expected from the remaining N ones. These expectations were obtained from the binomial distribution, i.e. the number of non-exclusions is assumed to follow a binomial distribution with parameters N and $(1 - Q)$. Thus the probability that n parent-pairs are not excluded is given by

$$\text{Prob}(n) = \binom{N}{n}(1 - Q)^n Q^{N-n}$$

for $0 \leq n \leq N$. The proportion of offspring assigned to n incorrect parent-pairs (i.e. offspring assigned to $n + 1$ parent-pairs) is Prob(n). For instance the proportion of offspring with a single assigned parent-pair (the correct one) is Prob(0), the proportion of offspring with two assigned parent-pairs (the correct one and one incorrect one) is Prob(1), and so on.

Finite number of parents and infinite number of offspring

The expectations derived above assume infinite population sizes. In practice, however, both the number of parent-pairs and the number of offspring are limited. There are thus two sampling processes to be considered: (i) sampling of parent-pairs from a large (strictly infinite) population, and (ii) sampling of progeny genotypes for a given distribution of parent-pairs. Both processes are assumed to be at random and independent. We will consider first only the sampling of parents and assume initially that the number of offspring is large enough to be considered infinite. This assumption is relaxed in the next section.

Without loss of generality, consider a single locus with two alleles with frequencies p_1 and p_2. There are six possible parent-pair genotypes (Table 1). Table 1 also shows the number of offspring of each genotype produced from each parent-pair type. When the sampling of offspring is ignored, then $m_{21} = ½$, $m_{22} = ½$, $m_{41} = ¼$, $m_{42} = ½$, $m_{43} = ¼$, $m_{51} = ½$, $m_{52} = ½$ and m is the number of offspring per parent-pair [$m = T/(N + 1)$, where T is the total number of offspring]. The distribution of the number of parent-pairs over the six groups is multinomial. The approach for finding the proportion of offspring assigned to $n + 1$ parent-pairs is as follows. For a given distribution of parent-pairs across the six groups (i.e. for a particular set of values for $n_1, n_2, \ldots, n_6$) first, the probability of that particular distribution is computed. This probability is obtained from the multinomial distribution:

$$\text{Prob}(n_1, n_2, \ldots, n_6) = \binom{N+1}{n_1, n_2, \ldots, n_6} \theta_1^{n_1} \theta_2^{n_2} \cdots \theta_6^{n_6}$$

where θ_i is the probability of parent-pair genotype i (Table 1). Secondly, the proportion of offspring of each genotype that is expected from that particular distribution is calculated. These expectations are $[(n_1 + ½n_2 + ¼n_4) \times \text{Prob}(n_1, n_2, \ldots, n_6)]/(N + 1)$ for genotype A_1A_1, $[(½n_2 + n_3 + ½n_4 + ½n_5) \times \text{Prob}(n_1, n_2, \ldots, n_6)]/(N + 1)$ for genotype A_1A_2, and $[(¼n_4 + ½n_5 + n_6) \times \text{Prob}(n_1, n_2, \ldots, n_6)]/(N + 1)$ for genotype A_2A_2. Thirdly, the number of parent-pairs assigned to each offspring genotype is obtained. These are $(n_1 + n_2 + n_4)$, $(n_2 + n_3 + n_4 + n_5)$ and $(n_4 + n_5 + n_6)$, for genotypes A_1A_1, A_1A_2 and A_2A_2, respectively. Finally, the proportions of offspring with the same number of parent-pairs assigned over all distributions of parents-pairs (i.e. sum over all combinations of values for $n_1, n_2, \ldots, n_6$) are added.

Finite numbers of parents and offspring

The previous section ignored the sampling of offspring given a particular distribution of parent-pairs. In order to account for the sampling of offspring, the terms m_{ij} shown in Table 1 (that represent Mendelian sampling) can be obtained from binomial (m_{21}, m_{22}, m_{51} and m_{52}) or multinomial (m_{41}, m_{42} and m_{43}) distributions. Thus, m_{21} and m_{22} can be obtained from a binomial (m, ½), m_{41}, m_{42} and m_{43} can be obtained

from a multinomial with probabilities 0.25, 0.5 and 0.25, and m_{51} and m_{52} can be also obtained from a binomial (m, ½).

Computer simulations

Stochastic computer simulations with finite numbers of parent-pairs and offspring were used to test the prediction models. They were also used to investigate the power of parental assignment in a practical situation for Atlantic salmon. Actual frequency data from a particular Scottish farm for seven characterized microsatellite loci (E. Verspoor, unpublished data) were simulated.

Genotypes of unrelated parents (N_m males and N_f females) were generated for L loci assuming Hardy–Weinberg equilibrium. The reproductive biology (high fecundity, external fertilization) of many fish species allows a very wide range of mating systems. Hierarchical mating designs (where each female is mated to a single male and each male is mated to d females) are common in fish breeding programmes, although for a fixed number of parents, factorial designs (where each female is mated to more than one male and each male is mated to more than one female) allow reduced inbreeding rates with no loss in selection responses (Woolliams 1989). Here both types of mating designs were considered: (i) hierarchical designs involving $N + 1 = dN_m = N_f$ parent-pairs and (ii) complete factorial designs where each male is mated to each female and each female is mated to each male, leading to $N + 1 = N_m N_f$ parent-pairs. Assignments of matings were at random and all matings produced the same number of offspring. The offspring genotype for each locus was obtained by sampling, at random, one allele from each parent.

For each offspring, the probability that each pair of parents has produced this offspring was computed as follows (i.e. $N + 1$ probabilities were computed for each offspring). Let A_kA_l and A_mA_n be the genotypes of a parent-pair at a single locus l. Following Sancristobal & Chevalet (1997) the probability that an offspring with heterozygote genotype A_iA_j has parents with genotypes A_kA_l and A_mA_n is

$$\begin{aligned} P_l &= \text{Prob}[(A_iA_j)|(A_kA_l),(A_mA_n)] \\ &= \text{Prob}[(A_i)|(A_kA_l)]\text{Prob}[(A_j)|(A_mA_n)] \\ &\quad + \text{Prob}[(A_j)|(A_kA_l)]\text{Prob}[(A_i)|(A_mA_n)] \end{aligned}$$

where $\text{Prob}[(A_i)|(A_kA_l)]$ is the probability that an individual with genotype A_kA_l has transmitted an allele A_i to its offspring. This probability is given by $(½\ \varepsilon_{ki} + ½\ \varepsilon_{li})$ where ε_{ki} is the probability that an allele A_k from a parent yields an allele A_i in the offspring. Assuming there are no genotyping errors nor mutations, $\varepsilon_{ki} = 0$ if $i \neq k$ and $\varepsilon_{ki} = 1$ if $i = k$. For an offspring with homozygote genotype A_iA_i, P_l is

$$\begin{aligned} P_l &= \text{Prob}[(A_iA_i)|(A_kA_l),(A_mA_n)] \\ &= \text{Prob}[(A_i)|(A_kA_l)]\text{Prob}[(A_i)|(A_mA_n)] \end{aligned}$$

For a set of L unlinked microsatellite loci, the probability of the offspring genotype conditional on the genotypes of the parental pair is calculated as

$$P = \prod_{l=1}^{L} P_l$$

The assigned parent-pairs for a particular offspring were those with non-zero probability. A minimum of 1000 replicates was run for each simulation.

Results

Exclusion probability

Exclusion probabilities increase with the number of loci used and with the number of alleles per locus. For the special case where all alleles at a given locus have the same frequency, six loci with more than two alleles (i.e. the case of microsatellite loci) or 15 biallelic loci would, in theory, be sufficient to exclude all incorrect pairs as parents of a given offspring ($Q \geq 0.99$). At a given locus, the exclusion probability reaches its maximum value when all alleles have the same frequency (e.g. Weir 1996; Jamieson & Taylor 1997) but this situation is rarely found in practice. With varying frequencies, the value of Q can be dramatically decreased and tends to zero as one allele frequency tends to zero. For instance, in an extreme situation where one allele is very common with a frequency >0.95, six 4-allele loci only allow 30% of the incorrect pairs to be excluded for a given offspring. In this situation, even 18 4-allele loci give a probability of exclusion of only 66%.

Parental assignment

Table 2 shows predicted and simulated percentages of offspring assigned to different numbers of parent-pairs when using three loci with five equiprobable alleles per locus. The assignment was for 800 offspring either from 100 or from 400 crosses. The crosses followed one of three mating designs considered, and they involved different numbers of parents. Predicted values in Table 2 are based on exclusion probabilities and binomial distributions (i.e. sampling of parents and offspring is ignored). In this case, predictions depend on the number of crosses but not on the number of parents and offspring involved.

The power of discrimination increased with the number of parents (for a given number of crosses) and it was substantially reduced when the number of crosses increased from 100 to 400. For a given number of crosses, the best

Table 2 Predicted and simulated percentage of offspring assigned to $n + 1$ parent-pairs when using three loci with five equifrequent alleles per locus. In the simulations, 800 offspring were generated from 100 or 400 parent-pairs resulting from three different mating designs: hierarchical with $N_m = N_f = 100$ or 400 ($H_{100\times100}$, $H_{400\times400}$); hierarchical with $N_m = 10$ or 40 and $N_f = 100$ or 400 ($H_{10\times100}$, $H_{40\times400}$); and factorial with $N_m = N_f = 10$ or 20 ($F_{10\times10}$, $F_{20\times20}$).

	100 crosses				400 crosses			
		Simulated				Simulated		
$n + 1$	Predicted	$H_{100\times100}$	$H_{10\times100}$	$F_{10\times10}$	Predicted	$H_{400\times400}$	$H_{40\times400}$	$F_{20\times20}$
1	30.64	34.05	25.55	21.23	0.85	3.54	3.74	3.21
2	36.46	33.40	28.68	26.83	4.08	8.40	7.84	7.53
3	21.48	19.59	20.83	17.36	9.76	12.04	10.71	7.98
4	8.35	8.71	12.59	13.92	15.52	13.63	12.21	10.13
5	2.40	3.05	6.67	7.39	18.46	13.49	12.30	8.53
6	0.55	0.93	3.24	5.79	17.53	12.47	11.51	9.77
7	0.10	0.21	1.45	3.00	13.84	10.74	10.18	7.86
8	0.02	0.05	0.61	1.88	9.34	8.71	8.58	7.69
9	0.00	0.01	0.24	1.13	5.50	6.57	6.78	6.64
>9	0.00	0.00	0.14	1.47	5.13	10.41	16.15	30.66

predictions were with the highest number of parents involved in the crosses ($H_{100\times100}$ and $H_{400\times400}$) as the sampling of parents is less important when more parents are used.

When the number of offspring was increased from 800 to 1600 the results from the simulations were practically the same as those shown in Table 2, indicating that the sampling of offspring was not important in these comparisons. The maximum difference between assignment percentages with 800 and 1600 offspring was 0.09 and it was found with 100 crosses and $F_{10\times10}$.

Table 3 shows the percentage of offspring assigned to a single parent-pair (the correct pair) when using different numbers of loci with different numbers of alleles with equal frequencies. Although the percentage of offspring with correctly assigned parents decreased with the number of crosses (particularly with small numbers of loci), the number of loci required to assign correctly most of the offspring was the same with 100 and 400 crosses. With equifrequent alleles, nine 5-allele loci or six 10-allele-loci were sufficient for assigning parents with certainty to 99% of the offspring.

Figure 1 shows the distributions of the percentage of offspring assigned to different numbers of parent-pairs obtained by computer simulation when using one, three or five biallelic loci. Predicted values are shown in Table 4 and they were obtained by determining all possible genotypes for parent-pairs and by identifying, for each possible offspring genotype, the parent-pair genotypes that cannot be excluded (as described in 'Infinite numbers of parents and offspring'). Although the expected values for the number of assigned parent-pairs (Table 4) coincide with the peaks in the simulations (Fig. 1), there was a considerable variation

Table 3 Predicted and simulated percentage of offspring with a single (correct) assigned parent-pair when using different numbers of loci (L) with different numbers of equifrequent alleles (k). In the simulations, 800 offspring were generated from 100 or 400 crosses resulting from different mating designs: hierarchical with $N_m = N_f = 100$ or 400 ($H_{100\times100}$, $H_{400\times400}$); hierarchical with $N_m = 10$ or 40 and $N_f = 100$ or 400 ($H_{10\times100}$, $H_{40\times400}$); and factorial with $N_m = N_f = 10$ or 20 ($F_{10\times10}$, $F_{20\times20}$).

		100 crosses				400 crosses			
			Simulated				Simulated		
k	L	Predicted	$H_{100\times100}$	$H_{10\times100}$	$F_{10\times10}$	Predicted	$H_{400\times400}$	$H_{40\times400}$	$F_{20\times20}$
5	3	30.64	34.05	25.55	21.23	0.85	3.54	3.74	3.21
	6	98.61	98.62	94.96	91.63	94.53	94.59	91.15	81.43
	9	99.98	99.98	99.72	99.47	99.93	99.93	99.68	98.84
	12	100.00	100.00	99.98	99.97	100.00	100.00	99.99	99.93
10	3	96.90	96.91	90.01	84.48	88.10	88.18	82.18	67.17
	6	100.00	100.00	99.94	99.87	100.00	100.00	99.93	99.72

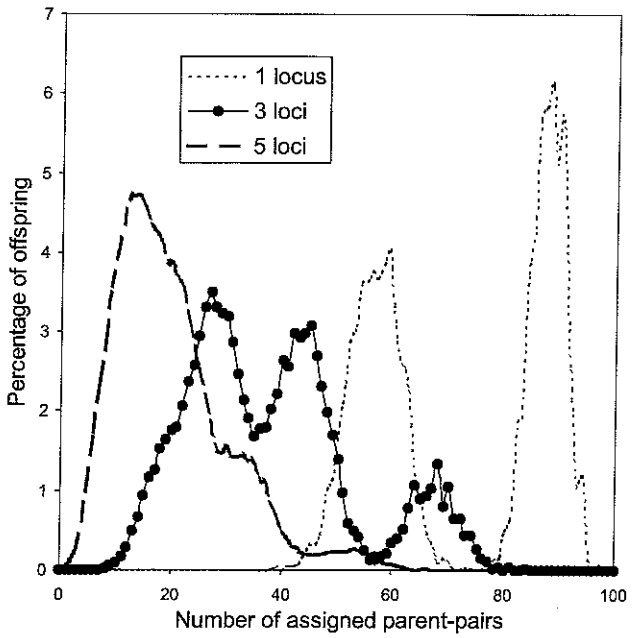

Figure 1 Percentage of offspring assigned to different numbers of parent-pairs for different number of biallelic loci. Both alleles had equal frequencies. Eight hundred offspring were generated from 100 crosses involving 100 sires and 100 dams.

Table 4 Predicted number of parent-pairs assigned to offspring ($n + 1$) and percentage of offspring (% offsp) when using different numbers of loci with two equifrequent alleles and 100 crosses.

One locus		Three loci		Five loci	
Offsp (%)	$n + 1$	Offsp (%)	$n + 1$	Offsp (%)	$n + 1$
50.00	56.25	12.50	17.80	3.125	5.63
50.00	87.50	37.50	27.69	15.625	8.76
		37.50	43.07	31.250	13.63
		12.50	66.99	31.250	21.20
				15.625	32.97
				3.125	51.29

around these values. With few loci, or with few alleles per locus, the number of possible parent-pair genotypes is small and accounting for sampling of parents is necessary.

Predictions accounting for sampling of parents

For a single biallelic locus, simulation results showed a wide distribution across expected values (Table 4 and Fig. 1). Also, predictions ignoring the sampling of parents deviated substantially from simulation results when the number of parents was small, even when considering a large number of loci and alleles per locus (Tables 2 and 3). Table 5 shows a comparison of predicted and simulated percentage of offspring assigned to different numbers of parent-pairs out of 10 possible pairs when sampling of parents is accounted for in the predictions. Accounting for sampling of parents leads to very accurate predictions (differences between prediction and simulation results were not significant). However, with a large number of parent-pairs or with a large number of loci or alleles per locus, the number of different distributions of parents across the different pair genotype types can be enormous. In this case, computer simulations would be the best option to predict the power of assignment.

Parental assignment in Atlantic salmon

The informativeness of the seven microsatellite loci (ranked according to their individual Q_i values) used in the analysis is shown in Table 6. All loci showed a high level of polymorphism. Combined probabilities of exclusion were calculated for different numbers of loci, adding loci from most to least informative. Based on exclusion probabilities, the joint use of the three most informative loci (1, 2 and 3) is expected to allow correct parental assignment for every offspring (i.e. 100% of non-parents will be excluded).

An example of the power of the seven loci for pedigree analysis is given in Table 7, which shows the percentages of offspring correctly assigned using different numbers of loci. Results are presented for the initial population and after five generations of random selection with constant numbers of parents across generations. Selection at random assumed that the microsatellite loci are unlinked to loci controlling production traits. For this specific example (allele frequencies found in a particular farm), the four loci with the highest exclusion probability were required in Scheme $H_{100\times100}$ to assign at least 99% of the offspring to the correct pair, both in the initial population and after selection. In the initial population, an additional locus was

Table 5 Predicted and simulated percentage of offspring assigned to $n + 1$ parent-pairs when using a single locus with two equifrequent alleles. Predictions account for sampling of parents. In the simulations, 1000 offspring was generated from 10 parent-pairs resulting from a hierarchical mating design with $N_m = N_f = 10$. Standard errors are given in parenthesis.

$n + 1$	Predicted	Simulated
1	0.03	0.03 (0.01)
2	0.34	0.32 (0.02)
3	1.75	1.66 (0.05)
4	5.25	5.38 (0.09)
5	10.22	10.20 (0.14)
6	13.79	13.94 (0.18)
7	14.82	14.90 (0.20)
8	17.18	16.75 (0.24)
9	21.30	21.34 (0.27)
10	15.32	15.47 (0.26)

Table 6 Number of alleles, allele frequencies and exclusion probabilities for seven microsatellite loci characterized in farmed Atlantic salmon. Exclusion probabilities for combined loci are shown at the bottom of the table.

Locus	Number of alleles	Allele frequencies	Exclusion probability
1	14	0.090, 0.167, 0.090, 0.064, 0.019, 0.109, 0.167, 0.032, 0.083, 0.032, 0.045, 0.006, 0.077, 0.019	0.929
2	21	0.279, 0.087, 0.007, 0.093, 0.007, 0.067, 0.007, 0.007, 0.007, 0.013, 0.020, 0.100, 0.053, 0.007, 0.040, 0.087, 0.033, 0.033, 0.013, 0.033, 0.007	0.919
3	15	0.080, 0.066, 0.015, 0.044, 0.102, 0.197, 0.204, 0.073, 0.044, 0.007, 0.015, 0.073, 0.007, 0.051, 0.022	0.915
4	13	0.006, 0.035, 0.018, 0.053, 0.006, 0.135, 0.141, 0.100, 0.212, 0.159, 0.024, 0.082, 0.029	0.896
5	12	0.012, 0.171, 0.012, 0.031, 0.006, 0.128, 0.110, 0.140, 0.226, 0.043, 0.116, 0.005	0.875
6	7	0.096, 0.096, 0.209, 0.136, 0.378, 0.079, 0.006	0.761
7	9	0.054, 0.524, 0.047, 0.067, 0.007, 0.080, 0.067, 0.141, 0.013	0.705
1 + 2			0.994
1 + 2 + 3			0.999
1 + 2 + 3 + 4			1.000

Table 7 Percentage of offspring with a single assigned parental pair when using different numbers of microsatellite loci resolved in Atlantic salmon. The assignment is based on the use of actual microsatellite frequency data from a farmstock (Table 6). Results are presented for the initial population ($t = 0$) and after five generations of random selection ($t = 5$). Eight hundred offspring was generated from 100 crosses resulting from different mating designs: hierarchical with $N_m = N_f = 100$ ($H_{100\times100}$); hierarchical with $N_m = 10$ and $N_f = 100$ ($H_{10\times100}$) and factorial with $N_m = N_f = 10$ ($F_{10\times10}$).

	$t = 0$			$t = 5$		
Loci used	$H_{100\times100}$	$H_{10\times100}$	$F_{10\times10}$	$H_{100\times100}$	$H_{10\times100}$	$F_{10\times10}$
1	5.0	4.2	3.9	4.4	2.3	0.5
1 + 2	67.9	55.5	50.7	60.6	26.5	10.5
1 + 2 + 3	96.0	88.8	87.1	93.2	61.9	35.1
1 + 2 + 3 + 4	99.6	97.2	96.7	99.0	83.4	59.8
1 + 2 + 3 + 4 + 5	100.0	99.3	99.2	99.8	92.7	77.1
1 + 2 + 3 + 4 + 5 + 6	100.0	99.7	99.7	99.9	96.0	85.4
1 + 2 + 3 + 4 + 5 + 6 + 7	100.0	99.8	99.8	100.0	97.5	90.2

needed in Schemes $H_{10\times100}$ and $F_{10\times10}$ to achieve the 99%. However, after five generations of selection, the seven loci were not enough for assigning correctly 99% of the offspring in Schemes $H_{10\times100}$ and $F_{10\times10}$.

Discussion

Deterministic predictions for the proportion of offspring assigned to a number of parent-pairs have been developed in order to investigate the number of marker loci required for correct parental assignment in fish breeding programmes. Predictions based on exclusion probabilities, that ignore the sampling of parents and offspring, were accurate provided that the number of parent-pairs involved in the crosses was large. These predictions assume binomial distributions for the number of non-exclusions. These type of distributions has been also suggested by Vankan & Faddy (1999) to estimate the reliability of paternity assignment in multiple-sire cattle herds.

In practical fish breeding programmes, with large numbers of offspring candidates for selection, the sampling of offspring can be ignored. Simulation results showed that the power of discrimination was the same with 800 and with 1600 offspring. However, the sampling of parents can be important as the high reproductive capacity of fish allows high selection intensities. With small numbers of parents, the exclusion techniques proved to be inaccurate (Tables 2 and 3). When sampling of parents was accounted for, then accurate predictions were obtained for a single biallelic locus and a small number of crosses (Table 5). These predictions would, however, be very difficult to derive analytically for loci with more than two alleles, or for a large number of crosses given the large number of marker genotypes. In these situations, Monte Carlo simulations are the only practical option to predict the power of parental assignment of particular sets of loci.

Predictions of the power of microsatellite markers for parental assignment rely on a number of simplifying assumptions. These include the absence of mutation and

measurement errors, unlinked loci, good estimates of allele frequencies and Hardy–Weinberg equilibrium (which implies that there is no selection and no inbreeding). The high mutation rates observed at microsatellite loci (from 10^{-4} to 10^{-2} per locus per gamete per generation; Weber & Wong 1993) can lead to frequent mismatches of offspring genotypes with those of their true parents. Scoring errors that can lead to incorrect determination of parentage, are also relatively common (2–3% per allele scored; O'Reilly *et al.* 1998). The assumption of unlinked loci may not represent a problem as no evidence of linkage for microsatellite loci isolated in fish species has been found (e.g. Banks *et al.* 1999; Verspoor, unpublished).

Of more concern is the assumption of Hardy–Weinberg equilibrium in real fish populations under artificial selection. Even if it is assumed that marker loci used in parental assignment are unlinked to loci controlling the traits under selection, a small effective population size would lead to disequilibrium for the loci considered and then the genotype probabilities cannot simply be deduced from allele frequencies. Analyses of microsatellites for Atlantic salmon in Scottish farms have shown that, while many loci do not have a significant departure from Hardy–Weinberg equilibrium, many other do (Verspoor, unpublished). However, unless departures are considerable, predictions are expected to behave well. For example, simulations were run with linkage disequilibrium in the parental generation (results not shown). Eight hundred offspring were obtained from 100 crosses involving 100 males and 100 females ($H_{100\times100}$) and assignments were made by using three loci with five equiprobable alleles per locus. When the number of homozygotes was reduced by 25% (relative to the number under Hardy–Weinberg equilibrium), the percentage of offspring assigned to 1, 2, 3, 4 and 5 parent-pairs was 29.37, 32.78, 21.66, 10.45 and 4.09%, respectively. The corresponding values obtained when the number of homozygotes was increased by 25% were 40.38, 34.44, 17.01, 6.07 and 1.68%. These results are not far from those obtained under Hardy–Weinberg equilibrium (Table 2).

In practice, in order to assign offspring to parent-pairs with exclusion techniques, the multilocus genotype of the offspring is compared with all possible parent-pair genotypes. Pairs that could not have produced the offspring's genotype are excluded, and one or more parent-pairs are assigned as possible parents. When *a priori* information on the potential parents is not available, then genetic likelihoods of parent–offspring relationships is a useful statistical method to reconstruct genealogies (e.g. Meagher & Thompson 1986). Log-likelihood ratios (LOD scores) are calculated for each offspring and the most likely parents are assigned. The situation in fish breeding programmes is different in that here the matings, and therefore all possible parent-pairs, are known. However, LOD analysis could be used to identify the genetically most likely parent-pair for each offspring, when the number of genetic exclusions is insufficient to narrow down parentage to one single parent-pair. Bernatchez & Duchesne (2000) have recently used this approach to predict parentage assignment success. Alternatively, the use of more loci could lead to unequivocal assignment of all the offspring. If there are genotyping errors, mutations or non-amplifying or 'null' alleles, then the exclusionary approach may lead to false exclusions of true parents. The likelihood approach has the advantage that it allows the inclusion of an error rate to account for errors in the genetic data (Marshall *et al.* 1998). In any case, however, the exclusion approach is a useful starting point to determine parentage.

The most informative four microsatellites developed for Atlantic salmon were sufficient to assign at least 99% of the offspring to the correct pair when 100 males and 100 females were used to produce 100 crosses (Table 7). An additional locus was required to correctly assign 99% of the offspring when the 100 crosses implied 10 males and 10 females. After five generations of random selection, the number of loci required for correct assignment remained unchanged in schemes using 100 male and 100 female parents (four loci). However, with small numbers of parents (10 males and 10 females) the use of all seven loci allowed correct assignment for only 90% of the offspring. The differences among different mating schemes in the power to discriminate among crosses was because of differences in the effective population sizes (N_e) which were 200, 36.4 and 20 for $H_{100\times100}$, $H_{10\times100}$ and $F_{10\times10}$, respectively. With small N_e there will be a higher chance of losing alleles (i.e. the frequency of heterozygotes will decrease more with small N_e) as a consequence of random drift.

The number of microsatellite loci and level of allelic diversity, indicated to be needed for assigning offspring to parents within breeding programmes, are well within the bounds of what is already available for commercial fish species such as Atlantic salmon (e.g. Table 6). While costs and time required for typing are declining rapidly with increased automation of screening and mass production of consumables, these are still sufficiently high that the use of the approach will be limited, particularly given the need to retype each time family assignment of a fish is necessary. However, this limitation can be overcome by using DNA analysis in conjunction with physical markers, such as passive integrated transponder (PIT) tags, applied once individuals are old enough for marking. This combined approach also allows individuals to be tracked to their family of origin when they are reared in test stations outside breeding facilities where destructive diseases challenge or slaughter quality tests are performed. Furthermore, DNA typing can also provide information to assist with other aspects of genetic

management such as monitoring levels of genetic variability (Verspoor 1998) or computing exact inbreeding coefficients (Visscher *et al.* 1998). However, the actual gains realized by using DNA typing are difficult to predict and will very much depend on breeding goals and breeding programme design (B. Villanueva, unpublished data).

Acknowledgements

We thank Prof. G. Simm and Dr B. McAndrew for useful comments on the manuscript. This work was funded by the Natural Environment Research Council (NERC) and Scottish Quality Salmon through the LINK Aquaculture Programme. SAC also receives financial support from the Scottish Executive Rural Affairs Department.

References

Avise J.C. (1994) *Molecular Markers, Natural History and Evolution.* Chapman & Hall, New York.

Banks M.A., Blouin M.S., Baldwin B.A., Rashbrook V.K., Fitzgerald H.A., Blankenship S.M. & Hedgecock D. (1999) Isolation and inheritance of novel microsatellites in chinook salmon (*Oncorhynchus tschawytscha*). *The Journal of Heredity* **90**, 281–8.

Bernatchez L. & Duchesne P. (2000) Individual-based genotype analysis in studies of parentage and population assignment: how many loci, how many alleles? *Canadian Journal of Fisheries and Aquatic Sciences* **57**, 1–12.

Castilho R. & McAndrew B. (1998) Two polymorphic microsatellite markers in the European seabass, *Dicentrarchus labrax* (L.). *Animal Genetics* **29**, 151–2.

Crooijmans R.P.M.A., Bierbooms V.A.F., Komen J., Van der Poel J.J. & Groenen M.A.M. (1997) Microsatellite markers in common carp (*Cyprinus carpio* L.). *Animal Genetics* **28**, 129–34.

Dodds K.G., Tate M.L., McEwan J.C. & Crawford A.M. (1996) Exclusion probabilities for pedigree testing farm animals. *Theoretical and Applied Genetics* **92**, 966–75.

Doyle R.W. & Herbinger C. (1994) *The Use of DNA Fingerprinting for High-intensity, Within-family Selection in Fish Breeding.* Proceedings of the 5th World Congress on Genetics Applied to Livestock Production, Guelph, Canada, Vol. 19, pp. 23–7.

Estoup A., Gharbi K., SanCristobal M., Chevalet C., Haffray P. & Guyomard R. (1998) Parentage assignment using microsatellites in turbot (*Scophthalmus maximus*) and rainbow trout (*Oncorhynchus mykiss*) hatchery populations. *Canadian Journal of Fisheries and Aquatic Sciences* **55**, 715–25.

Herbinger C.M., Doyle R.W., Pitman E.R., Paquet D., Mesa K.A., Morris D.B., Wright J.M. & Cook D. (1995) DNA fingerprint based analysis of paternal and maternal effects on offspring growth and survival in communally reared rainbow trout. *Aquaculture* **137**, 245–56.

Herbinger C.M., O'Reilly P.T., Doyle R.W., Wright J.M. & O'Flynn F. (1999) Early growth performance of Atlantic salmon full-sib families reared in single family tanks versus in mixed family tanks. *Aquaculture* **173**, 105–16.

Jamieson A. (1965) The genetics of transferrin in cattle. *Heredity* **20**, 419–41.

Jamieson A. & Taylor St.C.S. (1997) Comparisons of three probability formulae for parentage exclusion. *Animal Genetics* **28**, 397–400.

Khoo S.K., Ozaki A., Sakamoto T. & Okamoto N. (2000) Four highly polymorphic dinucleotide microsatellites in rainbow trout (*Oncorhynchus mykiss*). *Animal Genetics* **31**, 73–4.

Lee W.J. & Kocher T.D. (1996) Microsatellite DNA markers for genetic mapping in the tilapia, *Oreochromis niloticus*. *Journal of Fish Biology* **49**, 169–71. (no lo tengo pero citado en Kocher *et al.* 1998).

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**, 639–55.

Meagher T.R. & Thompson E. (1986) The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. *Theoretical Population Biology* **29**, 87–106.

Nelson R.J. & Beacham T.D. (1999) Isolation and cross species amplification of microsatellite loci useful for study of Pacific salmon. *Animal Genetics* **30**, 225–44.

Norris A.T., Bradley D.G. & Cunningham E.P. (2000) Parentage and relatedness determination in formed Atlantic Salmon (*Salmo salar*) using microsatellite markers. *Aquaculture* **182**, 73–83.

O'Connell M. & Wright J.M. (1997) Microsatellite DNA in fishes. *Reviews in Fish Biology and Fisheries* **7**, 331–63.

O'Reilly P.T., Herbinger C. & Wright J.M. (1998) Analysis of parentage determination in Atlantic salmon (*Salmo salar*) using microsatellites. *Animal Genetics* **29**, 363–70.

Perez-Enriquez R., Takagi M. & Taniguchi N. (1999) Genetic variability and pedigree tracing of a hatchery-reared stock of red sea bream (*Pagrus major*) used for stock enhancement, based on microsatellite DNA markers. *Aquaculture* **173**, 413–23.

Sancristobal M. & Chevalet C. (1997) Error tolerant parent identification from a finite set of individuals. *Genetical Research (Cambridge)* **70**, 53–62.

Slettan A., Olsaker I. & Lie Ø. (1996) Polymorphic Atlantic salmon, *Salmo salar* L., microsatellites at the SSOSL438, SSOSL439 and SSOSL444 loci. *Animal Genetics* **27**, 57–64.

Vankan D.M. & Faddy M.J. (1999) Estimations of the efficiency and reliability of paternity assignments from DNA microsatellite analysis of multiple-sire matings. *Animal Genetics* **30**, 355–61.

Verspoor E. (1998) Molecular markers and the genetic management of farmed fish. In: *Biology of Farmed Fish* (ed. by K.D. Black & A.D. Pickering), Chapter 11, pp. 355–82. Sheffield Academic Press, Sheffield.

Visscher P.M., Van der Beek S. & Haley C.S. (1998) Marker assisted selection. In: *Animal Breeding: Technology for the 21st Century* (ed. by A.J. Clark), pp. 119–36. Harwood Academic Publishers, Amsterdam.

Weber J.L. & Wong C. (1993) Mutation of human short tandem repeats. *Human Molecular Genetics* **2**, 1123–8.

Weir B.S. (1996) *Genetic Data Analysis II*, 2nd edn. Sinauer Associates, Inc, Sunderland.

Woolliams, J.A. (1989) Modifications to MOET nucleus breeding schemes to improve rates of genetic progress and decrease rates of inbreeding in dairy cattle. *Animal Production* **49**, 1–14