Relatedness between samples quantified and an optimal criterion for match detection approximated

THORVALDUR GUNNLAUGSSON[1]

Contact e-mail: thg@hafro.is

ABSTRACT

Data on relatedness of individuals between or within samples can be used to address population parameters in much the same way as conventional mark-recapture data and has some advantages, but also opens up new research areas. In such studies not only decisions on the sample size have to be made but also the number of genetic markers to be worked up, or even developed, and during analysis the criteria for accepting a match chosen. The likelihood of detecting a true match must be assessed and weighed against the likelihood of including a false positive. To aid with this, formulae are presented here for the probability of the number of relatives alive over periods of time and a process to approach the optimal criterion for match detection. To apply the process programs were developed that are made available, and an example is given.

KEYWORDS: ABUNDANCE ESTIMATE; MARK-RECAPTURE; SEX RATIO; BIOPSY SAMPLING; DNA FINGERPRINTING; MODELLING; AGE AT FIRST PARTURITION; SURVIVORSHIP; RECRUITMENT AGE

INTRODUCTION

Direct matches (recaptures) between collections of individual DNA-profiles (fingerprinting) have been used for mark-recapture analysis in natural populations. In addition these data can be used to infer kin relationships from likely related pairs (dyads). This has opened up many areas of research in behaviour, evolution and conservation (Blouin, 2003; Jones and Arden, 2004; Pemberton, 2008). There have been several studies where DNA-profiles have been used to detect instances of paternity in whale populations (Clapham and Palsbøll, 1997; Garrigue *et al.*, 2003; Nielsen *et al.*, 2001; Skaug and Øien, 2005). Skaug *et al*. (2005) screened the Norwegian minke whale (*Balaenoptera acutorostrata*) DNA-register for relatives and Skaug *et al*. (2006; 2009) screened samples from North Atlantic fin whales (*Balaenoptera physalus*) for relatedness. Advantages of relatedness are that matches of the potential types parent-offspring and half sibling/grand parentage can equal that of direct matches (recaptures), so in effect tripling the number of matches from given samples and over time this advantage becomes even greater as demonstrated below. Relatedness matches can also be found to existing samples that came from dead animals such as from strandings, by-catches and direct catches. Sampling biases are of much less concern than for direct matches (see under Methods).

Drawbacks in using relatedness matches are that estimates are needed of the probabilities of true detection, which may be significantly less than 1 due to type 2 errors, and of false positive inclusion, or type 1 errors, which may be unavoidable. For this, the frequency of marker alleles in the population is needed but never completely known. Certain assumptions about the genetic markers also have to be met to guarantee unbiased results (Skaug, 2001), which are not addressed here. The detection and false positive probability estimates are therefore inaccurate. Estimates are also needed of the number of relatives occurring (over time) of each type per individual. For parentage and grand parentage this is simply 2 and 4 (assuming inbreeding is negligible), but the number of half siblings may be species specific (see below). Full siblings are a small proportion (of the order 1/N) in large randomly mating populations and are not addressed here. Also needed is the probability that these relatives are alive (available) between samples, which is more complicated than with direct matching, but formulae for this are presented below.

This research was initiated by the need for a research programme to address stock structure hypothesis of North Atlantic fin whales. Numerous samples exist from catches 1983–1989. Research vessel time to obtain biopsies is expensive and therefore optimal use must be made of each new sample. Of particular importance are estimates of dispersal rates and abundance. The coefficient of variation (CV) of the estimated true number of matches (recaptures) will translate directly into the CV of these estimates (and most others). An approximate way of minimising this CV was therefore developed and is demonstrated below. General programs that were written to assist in this process are available on the web page [*http://www.iwcoffice.org/publications/additions.htm#add*].

A different approach to achieve the best precision in estimates is to use simulations (Økland *et al.*, 2009). This will require programming in each case, but may be the only option in complex scenarios.

MATERIAL

Pampoulie *et al.* (2008) found a lack of genetic divergence among samples of North Atlantic fin whales so the allele frequency table of North Atlantic fin whales combined (n = 469) at 15 micro-satellite loci (Skaug *et al.*, 2009) was used to calculate the probability of detection of relatedness and false positive inclusion. The observed matches as reported in Skaug *et al*. (2009) using the same micro-satellite DNA loci in samples of fin whales caught off West Iceland 1983–1989, are used to exemplify the method.

[1] *Marine Research Institute, Skúlagata 4, PO Box 1390, 121 Reykjavík, Iceland.*

METHODS

What counts as an observation

A sample consists of distinct DNA profiles. An animal recaptured within a sample will only provide a single profile (so without replacement). The number of distinct time ordered profile pairs is here denoted by n_p and is $n(n-1)/2$ within a sample of size n and between samples of size n_1 and n_2 is $n_1 n_2$ (minus the number of direct matches to the same individual which is assumed negligible in comparison). For convenience relatedness is here considered only to an older (earlier born) half sibling (HS) to make that test directional with time, as are the tests to parent (PO) and to grandparent (GP). Each ordered pair therefore provides two tests of relatedness of each type, such as given the first profile, test the second as parent and given the second profile, test the first as parent. The number of observations (matching tests) within a sample, and when summed both ways between samples, is therefore $2n_p$. Rather than multiply by 2, the probability of the relatives both being alive at the times of sampling is here added up both ways. Without auxiliary data, the direction of relatedness can (need) not be discerned. In general HS and GP relatedness also cannot be discerned and must be grouped under type-two relatedness (T2). Misassignment between type-one (PO) and T2 relatedness needs consideration in each case, but is not believed a significant problem in reasonable situations (Skaug, 2001), and not addressed here.

Issues of the sampling and survival

Mark-recapture analysis need random or balanced samples (at least one of the two samples) for unbiased results. If for instance males are more easily individually recognised and in the extreme only males are recaptured, an abundance estimate based on recaptures will refer only to the male component. Such a problem would have to be quantified and the data treated appropriately. In a relatedness study, however, the matching probabilities are unaffected and the data could be treated as representing the whole population.

Spatially one of the two samples needs to be random or balanced for unbiasedness, unless a random redistribution of the marked animals between samples can be assumed. When the marked animals are not the identified or tagged animals themselves but their relatives, this assumption is more plausible.

When direct recapture sampling is sufficiently random with respect to age (or size) above a certain limit, results will be unbiased referring to that component of the population. Recruitment age to the sampling (R) and the age at first parturition (B) are here assumed knife-edge and $R \leq B$. A year (time step) starts with birth. An average annual survival rate (S) of animals recruited to the sampling (above the limit) can be used to calculate recapture probabilities over time. If S can be assumed constant over age, then it is sufficient that age distribution of just the recapture (later) sample is representative of the population (if recruitment age of the samples differs by less than their timing $d \geq R_2 - R_1$). It is not sufficient that just the earlier sample is random. If for instance the later sample is biased towards younger animals there will be too few recoveries with increasing distance in time. However in such a case kinship may be slightly more likely one way, but then less likely the other way, so the total number of matches in a relatedness study may be little affected.

Number of half siblings

Siblings only count if they reach the age of recruitment to the sampling. To estimate the number of older half siblings born per individual an assumption such as of equilibrium or a stable population is needed such that each animal born will on average produce exactly two offspring, each recruit will produce two recruits and each mature animal produce two mature offspring. The expected number of half siblings over time is then two (one maternal and one paternal) and one of them expected to be older, but this holds only if each animal has exactly two offspring. If reproduction and survival of offspring is assumed random with on average two offspring surviving to maturity then this is a *Poisson* process with mean 2 (λ) over an average lifespan. Given one offspring of an average lifespan parent the expected number of other offspring is then independent and still 2 (λ). The λ_i of the *Poisson* distribution for parents with a different lifespan i is proportionately different. With the assumption of same survival at all mature ages the mean number of half siblings weighted with parent lifespan probability and offspring number over all i is then 4. If full siblings are rare the total number of half siblings is then double and half of them expected older, so that total is also 4. If animals intrinsically differ in their ability to produce mature offspring, such as some being sterile, this number will be larger, but to the opposite effect the assumption of a *Poisson* process and of no senescence may be too extreme, in particular for the females. When animals are recruited prior to maturity the number is higher by $1/S^{B-R}$, but the number 4 is used in demonstrations here.

Probability of a direct recapture

The probability of a direct recapture of the same animal is presented here for comparison. Given an animal from the earlier sample the probability of the animal being alive in the later sample d years later will have decreased by S^d. If the conditions discussed above are assumed met, the expected number of matches between samples with n_p pairs, that are d years apart, where N is the size of the recruited population, is

$$S^d n_p / N \qquad (1)$$

Probabilities of relatives alive

The summation identities $\sum x^k = 1/(1-x)$ and $\sum k\, x^{k-1} = 1/(1-x)^2$ over all $k \geq 0$ are used in the following. The given animal is in all cases the younger (later born) animal and is a randomly sampled animal in the population (and so alive at that time). When paternal half siblings are born in the same year the one in the later sample is here assumed older. The probability that an animal is sampled at age $R+i$ is $S^i(1-S)$ (sum over $i \geq 0$ is 1). The probability that this animal was born to a parent that is at age $B+k$ is $S^k(1-S)$ (sum over $k \geq 0$ is 1). The recruitment age of the older (earlier born) animal matters in the case of siblings but in case of other relatedness only when it came from a nonlethal sample. In the sibling case their survival before age R has no effect here and both the samples can be thought of as having taken place R years earlier and then directed at all ages, so R balances out and is

omitted in the following notation, but if R_1 differs in the earlier sample from R_2 in the latter sample then d should be replaced by $d-R_1+R_2$. The probabilities here refer to the stock recruited to the sample of the earlier born animal so when the recruitment ages differ, before adding the probabilities, they must be synchronised to refer both to for instance the total stock, by dividing with S^R where R is the recruitment age of the earlier born animal in each case.

Probability that a single older half sibling is alive within sample/year

Consider an age difference $k>0$ so the age of the older sibling would be $i+k$ at the time of the sampling if it were alive. The older sibling has to have survived the $i+k$ period so the probability of it being alive is S^{i+k}. The average probability of the older sibling being alive is obtained by summing this over all i and k weighted by the likelihood of each instance. The joint probability that the older sibling is alive and the younger sibling is at age i is $S^{i+k}S^i(1-S)$. Summing over all $i\geq 0$ gives the probability that the older sibling is alive given an age difference k as $S^k/(1+S)$. The probability of an age difference k decreases by S^k as the parent is less likely to be alive as time passes. This probability must sum to 1 so for single maternal siblings where $k\geq 1$ is $S^{k-1}(1-S)$. Paternal siblings can be born in the same year so the probability is for $k\geq 0$ and is $S^k(1-S)$. Applying these weights and summing over all k gives the older half sibling probability alive as

$$S/(1+S)^2 \text{ maternal}, \qquad 1/(1+S)^2 \text{ paternal}. \tag{2}$$

If recruitment is gradual over a limited age span the maternal probability is slightly higher, but over a longer span the effect is negative.

Probability single older half sibling is alive between samples spaced d years apart

Given an animal from the earlier sample, the probability of its older half sibling being alive in the later sample decreases by S^d. Given the expression above the older half sibling probability alive in the later sample becomes

$$S^{d+1}/(1+S)^2 \text{ maternal}, \qquad S^d/(1+S)^2 \text{ paternal}. \tag{3}$$

Given an animal in the later sample, the probability that its older half sibling is alive in the earlier sample is higher by S^{-d}, that is S^{i+k-d} while $k\geq d$. The joint probability of the younger sibling being at age i (as above) and the older sibling alive is $S^{i+k-d}S^i(1-S)$. When the age difference is $k\geq d$ summed over all $i\geq 0$ this is $S^{k-d}/(1+S)$.

When the age difference is $k<d$ the older sibling will not have been born or recruited while $i+k<d$ and the sum is taken only over all $i\geq d-k$ when the older sibling is available, this *is* $S^{d-k}/(1+S)$. Weighting both cases with the probability that the age difference is k (is as above) and summing $S^{k-1}(1-S)S^{k-d}/(1+S)$ over $k\geq d$ and $S^{k-1}(1-S)S^{d-k}/(1+S)$ over $k<d$ gives $S^{d-1}/(1+S)^2+S^{d-1}(d-1)(1-S)/(1+S)$ which gives the probability that older half sibling is alive in the earlier sample as

$$S^{d-1}(S^2+d(1-S^2))/(1+S)^2 \tag{4}$$

The combined probability, the sum both ways of the older half sibling being alive is then

$$\begin{aligned} &S^{d-1}(2S^2+d(1-S^2))/(1+S)^2 \text{ maternal} \\ &S^{d-1}(S+S^2+d(1-S^2))/(1+S)^2 \text{ paternal} \end{aligned} \tag{3}+(4)$$

If the females only give birth every second year then S must be replaced by S^2 and d by $d/2$ in the maternal probabilities above.

Probability that the parent is alive between samples spaced d years apart

When d is 0 the parent has to have survived at least to the recruitment age of the offspring, to be matched, so probability of survival is S^{R+i}. Summed and weighted with the age distribution of the offspring (see above) gives $S^R/(1+S)$. For larger d the probability decreases by S^d so the probability of its parent alive in a later sample is

$$S^{R+d}/(1+S) \tag{5}$$

Given an animal from a later sample the probability of the parent being alive in the earlier sample increases to

$$S^{R-d}/(1+S) \text{ while } d\leq R \tag{6}$$

When $d>R$ and the earlier sample came from dead animals, an animal in that sample may be excluded as the parent of young animals in the later sample so the probability of a parent possible and alive is

$$S^{d-R}/(1+S) \text{ when } d\geq R \tag{7}$$

When the earlier sampling was non-lethal and $d>R$ parents may be sampled before the birth of the offspring and the probability parent alive in nonlethal earlier sample increases to

$$S^{d-R}/(1+S)+1-S^{d-R} \text{ when } B\geq d\geq R \tag{8}$$

When $d>B$ some parents will however not be born or recruited to the sampling and the probability parent is alive in nonlethal earlier sample is

$$S^{d-R}/(1+S)+(1-S^{B-R}+(d-B)(1-S))S^{d-B} \text{ when } d\geq B\geq R \tag{9}$$

If the recruitment age of the parent (R_p) differs from the recruitment age of the offspring then B must be replaced with $B-R_P+R$ in the above. If recruitment is assumed 20, 40, 60, 80% by year the probability can be calculated by averaging these 4 cases and within sample is higher by 3% compared to a knife-edge recruitment (at the age of 50% recruitment +0.5), but this difference is less if samples are spaced R years apart.

The probability that the grandparent is alive between samples spaced d years apart

The probability that the offspring is at age $R+i$ at the time of the its sampling is $S^i(1-S)$ and that it was born when its parent was at the age $B+k-i$ ($k\geq i$) is $S^{k-i}(1-S)$. Joining the probabilities and summing over all $i(0\leq i\leq k)$ gives $(1-S)^2S^k(k+1)$ for the probability of the parent having been born exactly $B+R+k$ prior to the time of the sampling of the offspring. The grandparent must have been alive (probability 1) at the birth of the parent of the sampled animal. Given an animal in the earlier sample its grandparent must then survive the $B+R+k+d$ years since it gave birth to the parent to be alive in the later sample, so the probability is $S^{B+R+k+d}$. Summing these joined probabilities over all $k\geq 0$ gives the probability that grandparent is alive in the later sample as

$$S^{B+R+d}/(1+S)^2 \tag{10}$$

Given an animal from the later sample the probability that the grandparent is alive in the earlier sample is the same, but d has reversed sign

$$S^{B+R-d}/(1+S)^2 \text{ when } d \le B+R \quad (11)$$

The maximum probability is 1 when $d = B+R+k$ and is then 0 for larger d when the grandparent sample came from animals dead before the birth of the parent, so the combined probability of the grandparent present in an earlier sample from dead animals is

$$S^{d-B-R}(1+(1-S^2)(d-B-R))/(1+S)^2 \text{ when } d \ge B+R \quad (12)$$

In a non-lethal sample the probability that the grandparent is present is also 1 while $d > B+R+k$ and the grandparent is recruited although not mature, however for $d > 2B+k$ the grandparent is not recruited so in a non-lethal earlier sample the probability that is added to (12) is

$$+S^{v}(1+v(1-S)+(1-S)^2v(v+1)/2)-S^{d-B-R}(1+(d-B-R)(1-S)) \quad (13)$$

Where v = $d-2B$ and is replaced with zero if negative. Additional details such as for $R_1 \neq R_2$ are given on the web page.

Optimal matching criterion

When n_R is the expected number of relatives alive (available) in the population per individual (summed both ways) of the type of relatedness tested (see Table 1) the total number of potential positives in a population of size N is $n_R N$ and negatives (or false positives) is $N^2 - n_R N$. In a sample of n_p pairs expected positives are $n_p n_R/N$ and the negatives are then n_p minus the positives, which are supposedly negligible in comparison so that term is ignored. In most instances one will want to minimise the CV in the expected number (m_T) of detected true matches, obtained by subtracting the expected number (m_F) of false positives from the observed total number (m) of matches:

$$m_T = m - m_F \quad (14)$$

These numbers are a function of the critical value (c) chosen as a criterion for accepting matches, and their expected values are:

$$m_T = T(c)\, n_p\, n_R\, /N, \text{ and } m_F = F(c) n_p \quad (15)$$

where the probabilities of the detection of a true match $T(c)$ and the inclusion of a false positive match $F(c)$ depend on the allele frequencies and the measure of gene profile similarity used.

Here the interest was on a situation where expected returns are few, but the allele frequencies relatively well determined, therefore the variation in the observed number of matches m (approximate variance m) would dominate and the variation in the estimated probability of detection (F) and false positives (T) from the allele frequencies was ignored and the CV^2 approximated by m/m_T^2, that is $(m_T + m_F)/m_T^2$. Minimum is attained when:

$$T'(c)\, T(c)\, n_R\, /N + 2T'(c)\, F(c) - F'(c)\, T(c) = 0 \quad (16)$$

Note how the sample size n_p has conveniently cancelled out. When the allele frequencies are known T', T, F' and F can be computed and the equation can be solved for N/n_R and this tabulated as a function of c so an optimal critical value can be looked up for any population size (programs on web page). From an initially anticipated population size a critical value c is found and the corresponding observed matches $m(c)$ are then used to calculate a new Petersen population estimate $N = T(c)\, n_p\, n_R/(m(c) - F(c) n_p)$. This process is then repeated until N is stable and the minimum CV has been attained.

An efficient measure of gene profile similarity is the LOD (logarithm of odds) score (Meagher, 1986), where the odds are defined as P(match | related)/P(match | unrelated). The formulas for different kinds of relatedness at a single locus are given in Skaug (2001) and were used in programs (on web page) that can be used to calculate parent-offspring (PO) as well as half sibling/grand parentage (HS) LOD score mass density and cumulative distributions by c with assumptions of no relatedness (F' and F), and of parent-offspring as well as half sibling/grand parentage relatedness (T' and T) from allele frequencies, and the convolution of many loci. These are then combined into a look-up table for the optimal matching criterion.

RESULTS

Table 1a shows the probabilities of relatives alive and expected number of relatives alive over a range of years for $S = 0.9$ (used below) and $S = 0.95$ where the earlier sample came from dead animals. Table 1b gives a comparison to a non-lethal sample. A program is available on the web page (http://www.iwcoffice.org/publications/additions.htm#add) to generate tables for any value of S, R and B. Skaug *et al.* (2009) detected 11 potential parent-offspring pairs in 358 samples from North-Atlantic fin whales caught off West Iceland 1983−1989 using 15 micro-satellite DNA loci. The 63,903 distinct pairs (n_p, $\sqrt{n_p} = 253$) had been screened with a HS-LOD score of above 6.7. Table 2a shows by population size (per related individual) the optimal HS-LOD score and the mass density and cumulative probability of detection of true PO relatedness and false inclusion of unrelated pairs. The table shows that this HS-LOD score criterion should have detected 80% of the true PO matches and included 2.7 false positives. Based on these and assuming S = 0.9 and $n_R = 1.05$ parents alive (see Table 1) the Petersen population estimate is $0.8 \cdot 1.05 \cdot 63{,}903/(11-2.7) = 6{,}500$. Table 2a gives 7.1 as the optimal critical value for a stock of that size. Skaug *et al*. (2008) list the top HS-LOD score matches and there it is seen that there is in fact no PO match excluded by using this more stringent criterion. The PO detection ratio with this criterion is 0.73 and false positives 1.5 resulting in an estimate of 5,115, which by table 2a implies an optimal critical value of 6.9 which results in an estimate of 5,600 (CV 0.37) and so on, but here this process was terminated. This process is quickly applied if the pairs have been ordered by their LOD score. A PO-LOD score is more powerful for PO relatedness. Table 2b shows that a PO-LOD score of 8.4 would be optimal for screening when N/n_R is 5,417 and then with expected detection ratio of 95% and 1.1 false positives. This should result in a total of 12.8 observed matches and a CV of 0.29. The PO-LOD score for the top 11 HS-LOD score matched pairs is also given in (Skaug *et al.*, 2010) and the lowest PO-LOD score is 11.3 so it is likely that some pairs with a HS-LOD score less than 6.7 had a PO-LOD score higher than 8.4 and were missed. In table 2b it is seen that over 20% of all

Table 1a

Probabilities of relatives (PO = parent-offspring, HS = maternal half-sibling, GP = grandparentage, T2 = HS + GP) alive and expected number of relatives alive between years (d) when earlier sample came from dead animals for an equilibrium random mating population with on average 2 older half-siblings, knife-edge recruitment age (R) 7 and first parturition (B) 10 years. Left part with survival (S) 0.9 and right part 0.95. Column All/S^d gives total PO+T2 relatives alive divided by the direct recapture availability S^d.

S =	0.90	Probability alive			Relatives alive				0.95	Probability alive			Relatives alive			
D	S^d	PO	HS	GP	PO	T2	All	All/S^d	S^d	PO	HS	GP	PO	T2	All	All/S^d
0	1.000	0.503	0.498	0.092	1.006	2.360	3.366	3.37	1.000	0.716	0.499	0.219	1.432	2.872	4.304	4.30
1	0.900	0.506	0.501	0.092	1.012	2.372	3.384	3.76	0.950	0.717	0.500	0.220	1.434	2.880	4.314	4.54
2	0.810	0.514	0.498	0.094	1.029	2.368	3.397	4.19	0.902	0.720	0.499	0.221	1.440	2.880	4.320	4.79
3	0.729	0.528	0.491	0.097	1.057	2.352	3.409	4.68	0.857	0.724	0.497	0.222	1.449	2.876	4.325	5.04
4	0.656	0.548	0.480	0.100	1.097	2.320	3.417	5.21	0.814	0.731	0.494	0.224	1.462	2.872	4.334	5.32
5	0.590	0.574	0.467	0.105	1.149	2.288	3.437	5.82	0.773	0.739	0.491	0.227	1.479	2.872	4.351	5.62
6	0.531	0.607	0.451	0.111	1.214	2.248	3.462	6.51	0.735	0.750	0.486	0.230	1.500	2.864	4.364	5.94
7	0.478	0.647	0.434	0.118	1.293	2.208	3.501	7.32	0.698	0.763	0.480	0.234	1.526	2.856	4.382	6.27
8	0.430	0.582	0.416	0.127	1.164	2.172	3.336	7.75	0.663	0.725	0.474	0.238	1.450	2.848	4.298	6.48
9	0.387	0.524	0.397	0.137	1.058	2.136	3.194	8.24	0.630	0.689	0.468	0.243	1.377	2.844	4.221	6.70
10	0.348	0.471	0.377	0.148	0.943	2.100	3.043	8.73	0.598	0.654	0.460	0.249	1.308	2.836	4.144	6.92
15	0.205	0.278	0.283	0.233	0.567	2.064	2.631	12.78	0.463	0.506	0.419	0.288	1.012	2.828	3.840	8.29
20	0.121	0.164	0.202	0.322	0.339	2.096	2.435	20.03	0.358	0.392	0.372	0.330	0.783	2.808	3.591	10.02
25	0.071	0.097	0.140	0.303	0.194	1.772	1.966	27.39	0.277	0.303	0.325	0.341	0.606	2.664	3.270	11.79

Table 1b

Comparison of probabilities when earlier sample came from dead animals (left columns) and non-lethal (right columns). Probabilities are the same for PO when $d \leq R$ and for GP when $d \leq B+R$ and always for HS.

S =	90	Probability alive				0.95	Probability alive			
d	S^d	PO	PO	GP	GP	S^d	PO	PO	GP	GP
8	0.430	0.582	0.682			0.663	0.725	0.774		
9	0.387	0.524	0.713			0.630	0.689	0.786		
10	0.348	0.471	0.742			0.598	0.654	0.796		
15	0.205	0.278	0.610			0.463	0.506	0.767		
20	0.121	0.164	0.487	0.322	0.375	0.358	0.392	0.722	0.330	0.345
25	0.071	0.097	0.363	0.303	0.503	0.277	0.303	0.658	0.341	0.409
30	0.042	0.057	0.259	0.246	0.550	0.214	0.234	0.586	0.329	0.463
40	0.015	0.020	0.121	0.133	0.460	0.129	0.140	0.442	0.276	0.521
50	0.005	0.007	0.053	0.062	0.296	0.077	0.084	0.320	0.212	0.511

Table 2a

HS-LOD score for PO relatedness of North-Atlantic fin whales using 15 micro-satellite DNA loci. Optimal LOD critical value for a given population size per parent alive (N/n_R). True detection (T) and false positive detection (F). Divide column $CV\sqrt{n_p}$ by $\sqrt{n_p}$ to obtain approximate expected CV.

LOD	T'	T	F'	F	N/n_R	$CV\sqrt{n_p}$
9.0	0.150	0.379	1.31E-06	8.88E-07	247,112	1,013
8.0	0.189	0.563	7.99E-06	5.11E-06	41,697	319
7.5	0.204	0.657	1.96E-05	1.18E-05	16,723	181
7.4	0.165	0.674	1.89E-05	1.37E-05	13,622	160
7.3	0.181	0.692	2.44E-05	1.61E-05	11,392	144
7.2	0.196	0.711	3.17E-05	1.93E-05	9,327	128
7.1	0.157	0.727	3.03E-05	2.23E-05	7,630	113
7.0	0.188	0.746	4.34E-05	2.67E-05	6,287	101
6.9	0.150	0.761	4.14E-05	3.08E-05	5,130	90
6.8	0.178	0.779	5.91E-05	3.67E-05	4,211	80
6.7	0.156	0.795	6.31E-05	4.30E-05	3,393	71
6.6	0.136	0.808	6.56E-05	4.96E-05	2,788	63
6.5	0.160	0.824	9.33E-05	5.89E-05	2,276	56
6.0	0.125	0.887	1.95E-04	1.27E-04	785	31

Table 2b

PO-LOD score for PO relatedness of North-Atlantic fin whales.

LOD	T'	T	F'	F	N/n_R	$CV\sqrt{n_p}$
13.0	0.122	0.442	2.75E-7	2.73E-7	980,526	1,886
12.0	0.126	0.581	7.73E-7	8.40E-7	308,305	875
11.0	0.114	0.716	1.90E-6	2.30E-6	98,154	425
10.0	0.090	0.829	4.06E-6	5.62E-6	31,549	215
9.0	0.060	0.911	7.37E-6	1.20E-5	10,338	114
8.9	0.063	0.917	8.59E-6	1.29E-5	9,326	107
8.8	0.066	0.923	9.94E-6	1.39E-5	8,317	101
8.7	0.057	0.929	9.49E-6	1.48E-5	7,398	94
8.6	0.054	0.934	9.92E-6	1.58E-5	6,633	89
8.5	0.046	0.939	9.30E-6	1.67E-5	5,973	84
8.4	0.048	0.944	1.07E-5	1.78E-5	5,417	80
8.3	0.045	0.948	1.11E-5	1.89E-5	4,872	75
8.2	0.047	0.953	1.27E-5	2.02E-5	4,349	71
8.1	0.040	0.957	1.20E-5	2.14E-5	3,872	66
8.0	0.037	0.961	1.24E-5	2.27E-5	3,479	63
7.0	0.019	0.986	1.71E-5	3.70E-5	1,200	36
6.0	0.007	0.997	1.73E-5	5.30E-5	423	21

PO pairs should have a PO-LOD score in the range 8.4 to 11.3.

In Table 3 it is assumed that the 15 available micro-satellite loci could be effectively tripled for T2 (HS+GP) relatedness ($n_R \approx 2.3$ from table 1a). It is seen that the optimal HS-LOD score is 7.5 for a stock of the size implied by the PO matches (2.3*2.570≈5.600) and would give a detection ratio of about 61%, or 16 true matches and 3.2 false positives and a smaller CV of 0.28 (71/253) than from the now available PO pair profiles.

Table 3

HS-LOD score for North-Atlantic fin whales assuming tripled the presently available 15 micro-satellite DNA loci. Optimal value for a given population size per available older half-sibling (N/n_R). True detection (T) and false positive detection (F). Divide column $CV\sqrt{n_p}$ by $\sqrt{n_p}$ to obtain approximate CV.

LOD	T'	T	F'	F	N/n_R	$CV\sqrt{n_p}$
13.0	0.052	0.179	1.17E-8	1.06E-7	928,881	2,834
12.0	0.062	0.241	3.81E-8	3.61E-7	319,239	1,400
11.0	0.071	0.313	1.18E-7	1.17E-6	109,791	703
10.0	0.078	0.395	3.53E-7	3.62E-6	37,057	355
9.0	0.081	0.482	1.00E-6	1.07E-5	12,753	184
8.8	0.096	0.500	1.44E-6	1.34E-5	10,336	162
8.6	0.089	0.518	1.65E-6	1.65E-5	8,289	142
8.5	0.082	0.526	1.66E-6	1.81E-5	7,463	134
8.4	0.089	0.535	1.99E-6	2.01E-5	6,764	126
8.3	0.087	0.544	2.15E-6	2.23E-5	6,071	118
8.2	0.096	0.553	2.63E-6	2.49E-5	5,429	111
8.1	0.089	0.562	2.70E-6	2.76E-5	4,849	103
8.0	0.086	0.571	2.88E-6	3.05E-5	4,363	97
7.8	0.094	0.588	3.85E-6	3.75E-5	3,543	86
7.5	0.088	0.614	4.88E-6	5.04E-5	2,579	71
7.0	0.086	0.656	7.80E-6	8.23E-5	1,516	52
6.0	0.077	0.735	1.91E-5	2.11E-4	527	29

DISCUSSION

The objective method proposed here for choosing an optimal screening criterion in relatedness analyses has ignored the uncertainty in the calculations of the false positives $F(c)$ and detection ratios $T(c)$. These might affect the choice of the optimal c and would probably best be studied with simulations or resampling of the allele frequency data. However, choosing a higher detection ratio (close to 1) and therefore relatively more precise, will lead to a larger number of false positives and less precision there, and vice versa, so these factors may balance out over some range.

For the PO pairs that are false positives, too little age difference would be expected in about 50% of the cases. Age readings from earplugs were available for the sample used in the example here, but based on these none of the pairs could be excluded (Gunnlaugsson *et al.*, 2010). In such a case the derivation of the optimal criterion could be modified to take advantage of this by using $m_F = 0.5n_p\ F(c)$. In situations where mtDNA is available it can be used to exclude some false positives in maternal-offspring relatedness so a different LOD score criterion might be used for such pairs. In cases where one parent has been identified or is known (such as mother foetus pairs) there is also more power to identify the other parent and its relatives.

The Petersen population estimate was used here for simplicity, but the refined less biased version with m replaced by m+1 is generally recommended. The CV^2 is then approximated by $(m+1)/((m_T+1)(m_T+2))$ and the optimal critical value c is no longer independent of the sample size. The effect is larger when n_p is small (and therefore m small). Then a somewhat higher critical value should be investigated, in particular when the emphasis is on getting a safe lower bound for the population rather than a point estimate.

An alternate approach is to include (sum over) all pairs, but weight down poorer matches, and thereby avoid the need to choose a critical value. For a pair with LOD score x the likelihood of a positive is $T'(x)\ n_R\ n_p/N$ and a negative $F'(c)n_p$. A natural weight is the probability that this pair is positive which therefore is $T'(x)/[T'(x)+F'(x)\ n_RN]$. This requires access to and processing of all the data and the problem is here again that N needs to be known, but supposedly N is to be estimated from the obtained m_T so iteration is also needed here. The variance should be smaller and could be obtained through resampling.

ACKNOWLEDGEMENTS

The author wishes to thank Hans Skaug (University of Bergen) and Gunnar Stefánsson (University of Iceland) for helpful comments during preparation of the manuscript as well as the anonymous referees.

REFERENCES

Blouin, M.S. 2003. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends Ecol. Evol.* 18: 503–11.

Clapham, P.J. and Palsbøll, P.J. 1997. Molecular analysis of paternity shows promiscuous mating in female humpback whales (*Megaptera novaeangliae*, Borowski). *Proc. R. Soc. Lond. Ser. B.* 264(1378): 95–98.

Garrigue, C., Dodemont, R., Steel, D. and Baker, C.S. 2003. Organismal and 'gametic' capture-recapture using microsatellite genotyping confirm demographic closure and reproductive autonomy of a humpback whales wintering ground. *Marine Ecology. Progress Series* 274: 251–62.

Gunnlaugsson, T., Víkingsson, G.A., Pampoulie, C. and Elvarsson, B.T. 2010. Research programme on North Atlantic fin whales in relation to RMP Variant 2 and stock structure hypothesis IV. Paper SC/62/RMP1 presented to the IWC Scientific Committee, June 2010, Agadir, Morocco (unpublished). 22pp. [Paper available from the Office of this Journal].

Jones, A.G. and Arden, W.R. 2004. Methods of parentage analysis in natural populations. *Mol. Ecol.* 12(10): 2511–23.

Meagher, T.R. 1986. Analysis of paternity within a natural population of *Chamaelirium luteum*. 1. Identification of most-likely male parents. *Am. Nat.* 128: 199–215.

Nielsen, R., Mattila, D.K., Clapham, P.J. and Palsbøll, P.J. 2001. Statistical approaches to paternity analysis in natural populations and applications to the North Atlantic humpback whale. *Genetics* 157(4): 1673–82.

Økland, J.M., Haaland, Ø.A. and Skaug, H.J. 2009. A method for defining management units based on genetically determined close relatives. *ICES J. Mar. Sci* 67: 551–58.

Pampoulie, C., Daníelsdottir, A.K., Bérubé, M., Palsbøll, P.J., Árnason, A., Gunnlaugsson, T., Ólafsdóttir, D., Øien, N., Witting, L. and Víkingsson, G.A. 2008. Lack of genetic divergence among samples of the North Atlantic fin whale collected at feeding grounds: congruence among microsatellite loci and mtDNA in the new Icelandic dataset. Paper SC/60/PFI11 presented to the IWC Scientific Committee, June 2008, Santiago, Chile (unpublished). 17pp. [Paper available from the Office of this Journal].

Pemberton, J.M. 2008. Wild pedigrees: the way forward. *Proceedings of the Royal Society B – Biological Sciences* 275: 613–21.

Skaug, H. and Daníelsdóttir, A.K. 2006. Relatedness of North Atlantic fin whales. Paper SC/58/PFI9 presented to the IWC Scientific Committee, May 2006, St. Kitts and Nevis, West Indies (unpublished). 8pp. [Paper available from the Office of this Journal].

Skaug, H., Palsbøll, P., Bérubé, M. and Rew, M.B. 2005. Application of high resolution DNA profiling to management of northeastern Atlantic minke whales. Paper SC/57/SD2 presented to the IWC Scientific Committee, June 2005, Ulsan, Korea (unpublished). 8pp. [Paper available from the Office of this Journal].

Skaug, H., Pampoulie, C., Danielsdóttir, A. and Vikingsson, G. 2009. Report of the First Intersessional RMP Workshop on North Atlantic Fin Whales, 31 March to 4 April 2008, Greenland Representation, Copenhagen. Annex D. Relatedness of North Atlantic fin whales: an update. *Journal of Cetacean Research and Management (Suppl.)* 11: 439.

Skaug, H.J. 2001. Allele-sharing methods for estimation of population size. *Biometrics* 57: 750–56.

Skaug, H.J., Berube, M. and Palsboll, P.J. 2010. Detecting dyads of related individuals in large collections of DNA-profiles by controlling the false discovery rate. *Mol. Ecol. Resources*. 10: 693–700. [doi: 10.1111/j. 1755-0998.2010. 02833. x].

Skaug, H.J. and Øien, N. 2005. Genetic tagging of male North Atlantic minke whales through comparison of maternal and foetal DNA-profiles. *J. Cetacean Res. Manage.* 7(2): 113–18.

Date received: September 2011
Date accepted: September 2011