Finding relatives among North Atlantic common minke whales (*Balaenoptera acutorostrata*) based on microsatellite data: the relationship between false discovery rate (FDR) and detection power

Ralph Tiedemann[1], Magnús R. Tiedemann[2], Þorvaldur Gunnlaugsson[3], Christophe Pampoulie[3] and Gísli A. Víkingsson[3]

[1]Unit of Evolutionary Biology/Systematic Zoology, Institute of Biochemistry and Biology, University of Potsdam, D-14476 Potsdam, Germany

[2]Faculty of Computer Science, University of Magdeburg, D-39016 Magdeburg, Germany

[3]Marine Research Institute, IS-101 Reykjavík, Iceland

ABSTRACT

We use a data set of 244 samples of Icelandic minke whales completely typed (=no missing data) for 16 microsatellites to look for related individuals with a likelihood-based approach based on probabilities to share a certain number of alleles identical-by-descent at a given locus. We use simulated data sets both to establish statistical significance (controlling for the false-discovery rate, FDR) and to estimate the power of detection. We also investigate the impact of typing error on relatedness inference.

For duplicate samples and without typing error, the power of detection is 100% under all applied FDRs. For detection of parent-offspring pairs and full siblings, the power is still acceptable, while it is poor for pairs with lower level of relatedness (half siblings, first cousins).

Having 15 mother-fetus pairs in the data set allowed us to compare the estimated detection power to the observed probability of detection. These measures are closely correlated, pointing towards the validity of our power estimation.

Two duplicate samples were identified. Except for mother-fetus pairs, one additional parent-offspring pair was inferred. For three further pairs it is reasonable to assume that they might constitute also parent-offspring pairs, which were not unambiguously identified as such due to a single mistyped locus.

Having up to four mother-offspring pairs identified from a rather small data set (229 specimens, if fetuses are not counted) of a restricted area may indicate some non-random spatial aggregation of kin. It would translate into a conservative abundance estimate of 7849 individuals for West Iceland, a number in line with sighting surveys.

INTRODUCTION

The common minke whale (*Balaenoptera acutorostrata*) is distributed throughout the North Atlantic. The species is migratory spending the summer at high latitude feeding grounds and the winters at lower latitudes where breeding takes place although the winter distribution is poorly known (Víkingsson & Heide-Jørgensen 2014). The traditional IWC stock delineation is based on limited evidence (Donovan 1991) and a major review of stock structure is underway. For its feeding grounds, which reach from the East of Canada and Western Greenland in the West over Iceland, East Greenland and Jan Mayen to Spitsbergen, the Barent Sea, the Coast of Norway, and the North Sea in the East, it has been debated whether there is some population structure, pointing towards a deviation from panmixia. Possible hypotheses include mixing of specimens from several breeding sites as well as isolation-by-distance. However, given the vagility of minke whales throughout their life time, population structure – if any – may be only subtle. Under such scenarios and given the large and increasing body of genetic information available for North Atlantic minke whales, it is of interest to know, (1) whether these data sets contain related individuals and (2) how identified related pairs of individuals are distributed relative to each other, both in space and time.

Skaug et al. (2010) have suggested a statistical approach to such relatedness analysis which has been subsequently applied to several cetacean species, including fin whales (Pampoulie et al. 2013) and minke whales (Benónísdóttir et al. 2013).

Here, we apply this method to a new microsatellite data set of Icelandic common minke whales. With this example, we also evaluate how the detection power is affected by choice of the false discovery rate (FDR) for various relatedness categories of interest. We also investigate the impact of typing errors typically associated with microsatellite analysis on relatedness inference.

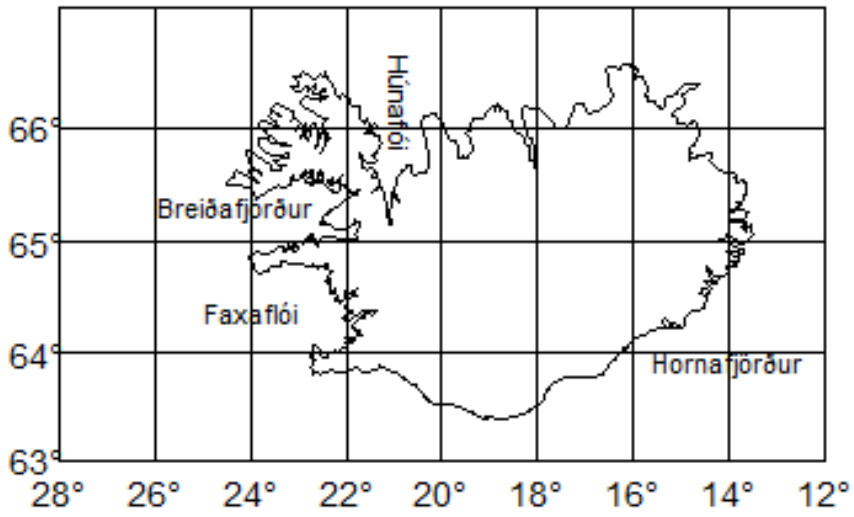

Figure 1: Sampling locations around Iceland

Table 1: Sampling locations within Icelandic waters and year of sampling of minke whale individuals scored at 16 microsatellites. Brei: Breiðafjörður Bay, West Iceland, Fax: Faxaflói Bay, SW-Iceland, Horn: Hornafjörður, SE-Iceland; Hún: Húnaflói Bay, N-Iceland (cf. Fig. 1).

Sampling-acronym	Number of samples
Brei2007	1
Brei2008	35
Brei2009	16
Brei2010	1
Brei2012	8
Fax2009	60
Fax2010	47
Fax2011	60
Fax2012	41
Horn1992	1
Horn2004	1
Horn2007	1
Hun2007	1
Hun2008	2
Hun2009	4
Hun2011	3
Iceland - others	3
Total	**285**

MATERIALS AND METHODS

Samples were obtained in Icelandic coastal waters from commercial and special permit catches. The Icelandic continental shelf area has a high density of common minke whales although the summer abundance has fluctuated widely (between 10,000 and 44,000 animals) in recent years, probably reflecting changes in prey availability (Víkingsson et al. 2014).

For this paper, we used data for 16 microsatellites (EV001, EV=037, EV094, EV096, GATA028, GATA053, GATA098, GATA417, GT011, GT023, GT195, GT211, GT310, GT509, GT575, and Sam25) from n=285 minke whales sampled around Iceland and typed at the University of Potsdam (table 1; details on microsatellite typing and results are published elsewhere). Typing error rates from re-typing of 12 randomly chosen samples and from inspecting genotype patterns in 15 known mother-offspring pairs were estimated as 0.01 and 0.004, respectively, such that the true typing error may be assumed to be at or below 1%.

We performed a relationship analysis as follows:

1) We omitted 41 samples because of missing data from further analysis, rendering a data set of 244 specimens completely typed at all 16 loci.
2) We calculated locus-wise relative allele frequencies (for mother-fetus pairs, the fetus genotype was excluded from this calculation).
3) Among all pairs i,j of individuals, we calculated pairwise LOD scores (=logarithmic probabilities to be related in a specific manner, e.g., parent-offspring, full-siblings, half-siblings) according to

$$LOD_{i,j} = \log\left\| \prod_{s=1}^{m} \left[\begin{array}{l} k_o + k_1\left(\dfrac{I(a_{i,s}^{(1)} = a_{j,s}^{(1)}) + I(a_{i,s}^{(1)} = a_{j,s}^{(2)})}{4p(a_{i,s}^{(1)})} + \dfrac{I(a_{i,s}^{(2)} = a_{j,s}^{(1)}) + I(a_{i,s}^{(2)} = a_{j,s}^{(2)})}{4p(a_{i,s}^{(2)})} \right) \\ + k_2\left(\dfrac{I(a_{i,s}^{(1)} = a_{j,s}^{(1)})I(a_{i,s}^{(1)} = a_{j,s}^{(2)}) + I(a_{i,s}^{(2)} = a_{j,s}^{(1)})I(a_{i,s}^{(2)} = a_{j,s}^{(2)})}{2p(a_{i,s}^{(1)})p(a_{i,s}^{(2)})} \right) \end{array} \right] \right\| \quad (1)$$

where i and j are the specimens to be compared, m is the number of microsatellites, k_0, k_1, and k_2 are the expected probabilities of having 0, 1, and 2 alleles identical-by-descent for a given category of relationship (table 2), I is the identity function (i.e., I=1 for identity, I=0 for non-identity), $a^{(1)}_{i,s}$ and $a^{(2)}_{i,s}$ are the first and second allele of specimen i at locus s, and p(a) is the relative frequency of allele a in the population (see Skaug et al. 2010, Benónísdóttir 2012, and Pampoulie et al. 2013 for derivation).

4) p-values for each LOD score of the original data set were estimated by comparing them to LOD scores obtained in a random data set of unrelated specimens. Therefor, a dataset 4 times the size of the original data set (976 specimen, corresponding to 475,800 pairwise comparisons among any two specimens) was simulated with the locus-specific allele frequencies (fetuses excluded; see (2)).
5) Significance was established controlling for the false discovery rate (FDR; Benjamini & Hochberg 1995), as outlined in Pampoulie et al. (2013). Briefly, LOD score-specific p-values of the original data set were ordered in increasing order and re-numbered such that $p(m) \leq p(n)$ for all $m<n$. For a given false discovery rate q, the threshold LOD_q score was determined as the LOD score with the probability p(r) where r is the largest value fulfilling the equation:

$$p(r) \leq \frac{r}{n} q \qquad \text{(Pampoulie et al. 2013)} \qquad (2)$$

6) We simulated a data set of 10,000 specimens with the relatedness of interest, taking into account the applicable probabilities of locus-specific numbers of alleles identical-by-descent (table 2). The detection power of a given FDR and relatedness category was calculated as the relative proportion of the simulated related specimens with a LOD score equal to or exceeding the FDR-specific LOD_q score threshold.

We repeated this analysis for various combinations of FDR (0.001, 0.00178, 0.00316, 0.00562, 0.01, 0.0178, 0.0316, 0.0562, 0.1, 0.178, 0.316, 0.562; values were chosen to be approximately equidistant on a logarithmic scale) and relatedness categories to estimate the detection power. Power estimates were always based on 6 replicate simulations with identical parameters.

We also used the method to detect pairs of related specimens in our sample, using FDR rates of 0.001, 0.01, 0.05, and 0.1. To further increase precision in p-value estimation for this analysis, we increased the number of simulated unrelated specimens to 5x the original data set (1220 specimens, corresponding to 743,590 pairwise LOD scores).

In this analysis, we investigated also the impact of a potential typing error e by using adjusted k values as follows:

$$k_0^* = k_0 + e\,k_1 + e^2 k_2 \;; \quad k_1^* = (1-e)\,k_1 + 2e\,k_2 \;; \quad k_2^* = (1-2e-e^2)\,k_2 \qquad (3)$$

To investigate the effect of potential typing error, we performed these analyses for e=0.00, e=0.01, and e=0.02. For e>0, adjusted k-values were used both for LOD score calculation and for the simulation of related specimens of a specified relatedness category.

Table 2: Probability of number of alleles (k) to be identical-by-descent at any locus, given a specified relatedness.

Relatedness of interest	k_0	k_1	k_2
Identity*	0	0	1
Parent-offspring	0	1	0
Full siblings	0.25	0.5	0.25
Half siblings**	0.5	0.5	0
First cousins	0.75	0.25	0

*cannot be distinguished from monozygotic twins; **cannot be distinguished from grandparent-grandchild or uncle/aunt-nephew/nice.

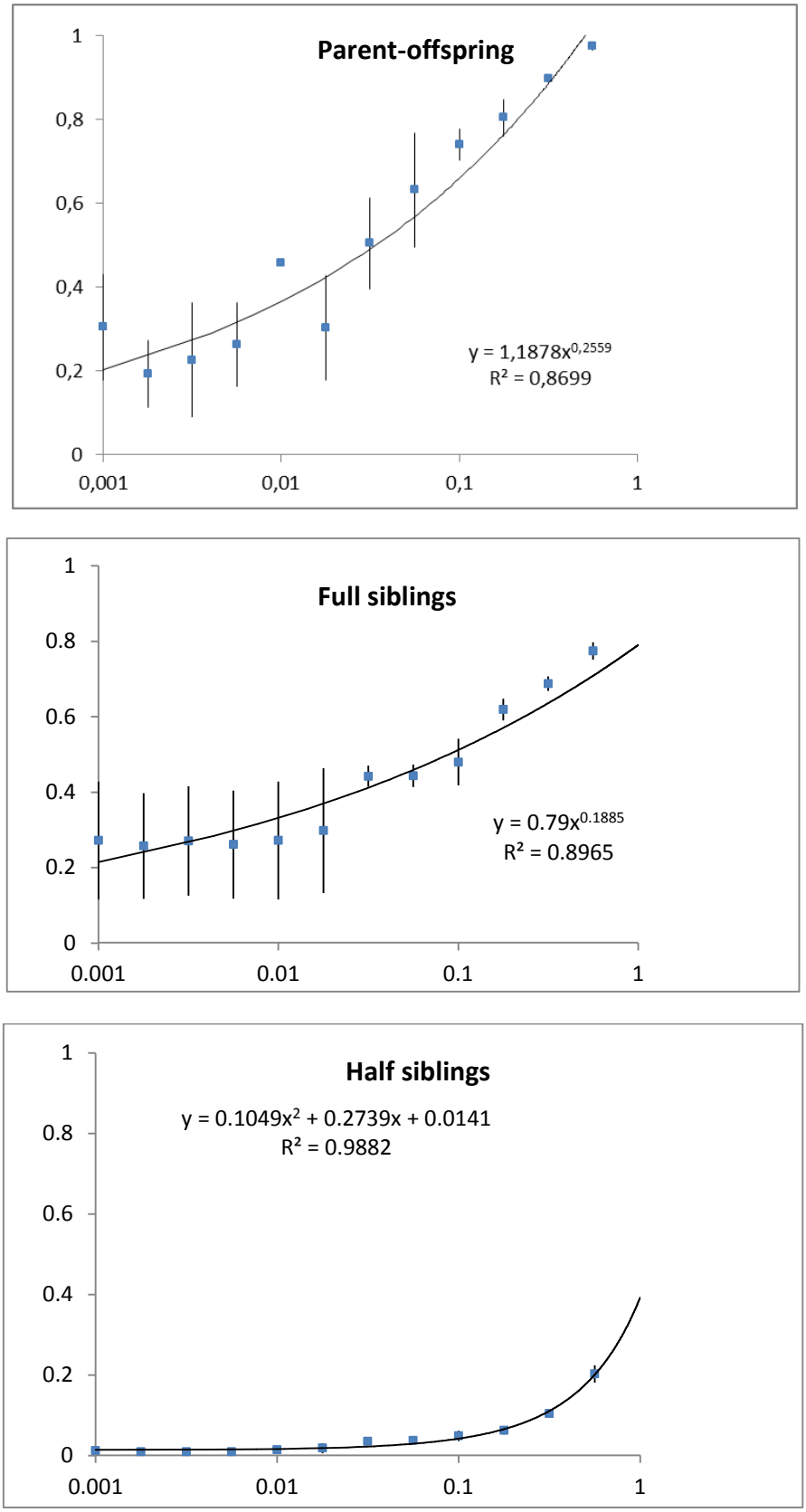

Figure 2: Correlation between false-discovery rate (x-axis) and detection power (y-axis) for identification of parent-offspring pairs (upper), full siblings (mid), and half siblings (lower). Dots represent arithmetic means of 6 simulations and are provided with 95% confidence limits.

Table 3: Inferred relatives among Icelandic common minke whales (in parentheses location and year of catch; B=Breiðafjörður, F=Faxaflói; F indicates fetus of female with same sample number; LOD scores are provided for typing error 0.00, 0.01, and 0.02).

Pair		Typing error	Identical	Parent-Offspring	Full siblings	Half siblings	First cousins	Interpretation
E09-009	E09-010	0.00	17.08***	8.53***	11.78***	5.68***	3.55***	Identical
(B2009)	(B2009)	0.01	17.00***	8.48***	11.70***	5.65***	3.53***	
		0.02	16.93***	8.43***	11.63***	5.61***	3.50***	
E09-012	E09-013	0.00	19.35***	9.76***	13.54***	6.75***	4.43***	Identical
(B2009)	(B2009)	0.01	19.26***	9.71***	13.46***	6.71***	4.40***	
		0.02	19.17***	9.66***	13.38***	6.68***	4.37***	
E09-011	E09-011F	0.00	-∞	6.07***	4.37(*)	4.47***	3.22***	Parent-offspring
(B2009)	(B2009)	0.01	-11.46	6.05***	4.38*	4.45***	3.21***	
		0.02	-8.16	6.02***	4.39*	4.43***	3.19***	
E09-020	E09-020F	0.00	-∞	3.62	3.59(*)	2.37	1.42	Parent-offspring
(F2009)	(F2009)	0.01	-8.18	3.61	3.59	2.36	1.41	
		0.02	-5.49	3.59	3.58(*)	2.34	1.39	
E09-021	E09-021F	0.00	-∞	6.78***	4.54*	4.78***	3.21***	Parent-offspring
(F2009)	(F2009)	0.01	-12.90	6.75***	4.55*	4.73***	3.19***	
		0.02	-9.30	6.72***	4.56**	4.71***	3.17***	
E09-062	E09-062F	0.00	-∞	6.29***	4.20(*)	4.12***	2.55**	Parent-offspring
(F2009)	(F2009)	0.01	-11.45	6.25***	4.21*	4.09**	2.53***	
		0.02	-8.15	6.22***	4.21(*)	4.07***	2.52**	
I10-024	I10-024F	0.00	-∞	5.05*	3.34	3.34*	2.12	Parent-offspring
(F2010)	(F2010)	0.01	-14.36	5.02*	3.36	3.32*	2.10(*)	
		0.02	-10.75	4.99***	3.38	3.30*	2.09(*)	
I11-004	I11-019	0.00	-∞	4.26(*)	3.20	2.70	1.59	Parent-offspring
(F2011)	(F2011)	0.01	-11.95	4.24(*)	3.20	2.68	1.58	
		0.02	-8.66	4.21*	3.21	2.66	1.57	
I11-018	I11-018F	0.00	-∞	7.28***	5.02*	5.28***	3.72***	Parent-offspring
(F2011)	(F2011)	0.01	-10.84	7.25***	5.03***	5.26***	3.70***	
		0.02	-7.55	7.22***	5.04***	5.23***	3.68***	
I11-020	I11-020F	0.00	-∞	4.38*	3.47	2.75	1.60	Parent-offspring
(F2011)	(F2011)	0.01	-8.69	4.35(*)	3.47	2.73	1.59	
		0.02	-6.00	4.32*	3.47	2.71	1.58	
I11-036	I11-036F	0.00	-∞	4.27(*)	2.26	2.65	1.54	Parent-offspring
(F2011)	(F2011)	0.01	-15.35	4.24(*)	2.28	2.62	1.53	
		0.02	-11,74	4.21*	2.29	2.61	1.52	
I11-045	I11-045F	0.00	-∞	5.85***	5.22***	3.79*	2.29*	Parent-offspring
(F2011)	(F2011)	0.01	-5.04	5.82***	5.21***	3.76*	2.28*	
		0.02	-2.66*	5.78***	5.20***	3.74**	2.26*	
B08-01	I12-022	0.00	-∞	-∞	3.71(*)	3.18*	2.03(*)	Parent-offspring#
(B2008)	(F2012)	0.01	-11.42	3.22	3.72(*)	3.17*	2.02	
		0.02	-8.13	3.50	3.73(*)	3.15*	2.01	
B08-04	E09-041	0.00	-∞	-∞	3.73(*)	2.44	1.50	Parent-offspring#
(B2008)	(F2009)	0.01	-9.30	2.29	3.53(*)	2.43	1.49	
		0.02	-6.31	2.57	3.72(*)	2.41	1.48	
I10-015	I10-032	0.00	-∞	-∞	2.71	2.96(*)	1.96	Parent-offspring#
(F2010)	(F2010)	0.01	-14.64	2.90	2.73	2.94(*)	1.95	
		0.02	-11.03	3.17	2.75	2.92	1.93	

Detected at FDR of: ***≤0.001; **≤0.01; *≤0.05;(*)≤0.1 #likely parent-offspring with a single mismatch due to mistyping/mutation; see text for details.

RESULTS

Estimation of detection power

For the detection of duplicate specimens in the data set and assuming no typing error, the detection power was always 1, even with the lowest FDR tested (0.001). Figure 2 shows the detection power, conditional on the false discovery rate, for parent-offspring pairs, full siblings and half siblings. For parent-offspring pairs, the detection power for the lowest tested FDR=0.001 was estimated 30±13% (confidence interval, CI), for FDR=0.01 46±1%, and for FDR=0.1 74±4%. For full siblings, the detection power both for FDR=0.001 and FDR=0.01 was estimated 30±13% CI and for FDR=0.1 48±6%.

Table 4: Observed and estimated detection probability of parent-offspring pairs. Observed probability was estimated as the detected percentage among the 14 known mother-fetus pairs with genotypes compatible at all loci with a parent-offspring relationship (i.e., excluding the 15th pair with a single mismatch at one locus).

False discovery rate	Observed detection probability	Estimated detection probability
0.3	0.86	0.90
0.2	0.79	0.88
0.1	0.57	0.72
0.05	0.50	0.68
0.01	0.36	0.23

For these relatedness categories, detection power is hence reasonable, but estimates vary considerably, at least with our relatively small original data set of 244 specimens under the parameters chosen for simulation (see above).

For the inference of half siblings, the detection power was considerably lower, i.e., only 1.2% for FDR=0.001, 1.5% for FDR=0.01, and 5.0% for FDR=0.1. Power for inference of first cousins was even lower (data not shown).

Inferred relatives in Icelandic minke whales

Looking for identical genotypes revealed two pairs of samples, each of which originating from the same source and assigned to adjacent sample numbers (table 3). It is hence reasonable to assume that these samples are duplicates.

If allowing for a false discovery rate of 0.1 and assuming no typing error, 10 parent-offspring pairs are discovered, 9 of which are mother-fetus pairs (table 3). The inferred parent-offspring pair I11-004 and I11-019 are a 7.94m long female caught in May 2011 and a 7.61m long male caught in June 2011, both from Faxaflói Bay/Southwest Iceland.

When no typing error was assumed, there were two inferred pairs of full siblings (B08-01/I12-22 and B08-04/E09-41) and one inferred pair of half-siblings (I10-015/I10-32). B08-01 is a 7.40m male caught in May 2008 in Breiðafjörður, I12-22 is a 7.38m male caught in June 2012 in Faxaflói. B08-04 is a 7.40m female caught in June 2008 in Breiðafjörður, E09-41 is a 7.65m male caught in August 2009 in Faxaflói. I10-15 and I10-32 are two 7.78m and 7.90m males, both caught in Faxaflói in June resp. July 2010. However, when taking a potential typing error of 0.01 resp. 0.02 into account, all these pairs yielded positive LOD scores for a parent-offspring relationship. Moreover, for the pair I10-015/I10-32 and a typing error of 0.02, a parent-offspring relationship yielded the highest LOD score (table 3). Close inspection of the genotypes of these ambiguous pairs revealed that in all three pairs only a single locus (a different one in each pair) is incompatible with a parent-offspring relationship. We hence tentatively also classify them as potential parent-offspring pairs; re-typing of these pairs is underway.

No first cousin was inferred.

The data set contained 15 mother-fetus pairs. In 14 of these pairs, genotypes were at all loci compatible with a parent-offspring relationship (the 15th had a single mismatch at on locus). Comparing observed detection probability (=percentage of detected mother-fetus pairs) to the detection probability estimated from the simulated set of related individuals showed a high correlation (r=0.903) and a reasonably good fit (table 4), in particular when considering the low number of known positives (i.e., 14) and the stochastic variation in power estimation (cf. figure 2).

DISCUSSION

The probability to correctly identify relatives hidden in a large genetic data set is dependent on

1) the frequency of the occurence of such relatives in the data set,
2) the ability to keep the identification of false-positives low (controlled by the FDR), and
3) the power of the applied method, i.e., the probability to detect true positives.

While 1) is the biological relevant parameter to be studied, 2) and 3) are highly dependent on the applied method. With regard to the utilized genetic marker and given a particular FDR allowed for, the number of loci and the frequency distribution of their alleles dictate the power of the analysis.

The example of the Icelandic minke whale typed for 16 microsatellites shows both potential and limitation of the inference of relatedness. The power to detect identical individuals (duplicates) is 100%, even if FDR is kept low. For parent-offspring and full sibling detection, the power of the applied method is still reasonable, while it is poor for less related pairings (half-sibs, first cousins).

Previous studies on relatedness inference of baleen whales have often not looked for full siblings, as the likelihood of full siblings may be considered very low, i.e., there is so far no evidence for pair bonds to be maintained throughout consecutive breeding seasons. However, our pairs inferred to be siblings when assuming no typing error have pointed us towards potential parent-offspring pairs with a single mismatch due to mistyping or mutation. Indeed, taking into account the effect of typing error in the relatedness analysis yielded positive LOD scores for a parent-offspring relationship of these pairs.

Assuming that the 4 potential parent-offspring pairs in this data set (229 specimens, when fetuses are not counted, or 229*228/2 = 26106 tests for matches) are factual at FDR 0.1 then the expected number of true detections is 4*0.9 or 3.6. The detection power at FDR 0.1 was estimated 0.7216 which gives 3.6/0.7216 = 4.9889, i.e., approximately 5 expected parent-offspring pairs in total in the sample. From table 1 in Gunnlaugsson (2012), expected parents alive in the first years when survival is taken to be 0.95 is around 1.5. This gives the estimate N = 26106*1.5/4.9889 = 7849 for the West Iceland grounds assuming there a fully mixed isolated population, which is a conservative assumption. This is in line with sighting survey estimates in this restricted area and may hint at some non-random spatial aggregation of kin. This has to be further investigated in a larger data set covering a larger area and ideally comprising further informative genetic markers.

References

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. B 57, 289-300.

Benónísdóttir S (2012) Statistical approach to relatedness analysis in large collections of genetic profiles – an application to a DNA registry of Fin whales. M.Sc. thesis, University of Bergen.

Benónísdóttir S, Skaug HJ, Glover K, Elvarsson BÞ, Víkingsson GA, Pampoulie C (2013) Genetic study on close relatedness of common minke whale Balaenoptera acutorostrata in the Central and Northeast Atlantic. IWC SC/F13/SP20.

Donovan GP (1991) A review of IWC stock boundaries. Rep Int Whal Commn Spec Issue 13, 39–68.

Gunnlaugsson Þ (2012) Relatedness between samples quantified and an optimal criterion for match detection approximated. Journal of Cetacean Research Management. 12, 335-340.

Pampoulie C, Benónísdóttir S, Skaug HJ, Víkingsson G (2013) Genetic relatedness of North Atlantic fin whale *Balaenoptera physalis* in Icelandic waters. IWC SC/65a/RMP01

Skaug HJ, Bérubé M, Palsbøll P (2010) Detecting dyads of related individuals in large collections of DNA-profiles by controlling the false discovery rate. Molecular Ecology Resources 10, 693-700.

Víkingsson GA, Elvarsson BT, Ólafsdottir D, Sigurjonsson J, Chosson V, Galan A (2014) Recent changes in the diet composition of common minke whales (Balaenoptera acutorostrata) in Icelandic waters. – A consequence of climate change? Mar Biol Res 10, 138–152.

Víkingsson GA, Heide-Jørgensen MP (2014) First indications of autumn migration routes and destination of common minke whales tracked by satellite in the North Atlantic during 2001 - 2011. Mar Mammal Sci, in press.