Preliminary estimates of error rates in the Norwegian minke whale DNA-register based on mother-fetus pairs

Øystein Haaland and Hans J. Skaug

Address: Department of Mathematics, University of Bergen, Norway

April 25, 2007

Abstract

We estimate genotyping error rates in the Norwegian minke whale DNA-register using DNA-profiles from 589 mother-fetus pairs. The basic idea is that mother and offspring must share at least one allele per locus. It is found that the laboratory currently used for the DNA-register has a much lower error rate than the laboratory used until 2002. This conclusion is supported by auxiliary data consisting of a repeated scoring of 25 individuals for which the true genotype is believed to be known. The error rates for the period 2002- are comparable to those found in the published literature.

1 Introduction

The Norwegian DNA-register for common minke whales was established in 1997 and contains DNA-profiles of almost every individual whale caught by Norway since (Olaisen 1997; Anon. 2007). The register has been challenged at one occasion, and it was concluded that twenty minke whale tissue samples collected at markets in Norway matched DNA-profiles in the register (Palsbøll et al. 2006). Knowledge about genotyping error rates is essential for the operation of the DNA-register, and the goal of the present study is to provide such estimates.

Since year 2000 tissue samples from the fetuses of pregnant females have been collected as part of the sampling procedure for the DNA-register. For a subset of 291 fetuses, DNA-profiles have been established using the same laboratory as used for the corresponding mother profiles in the DNA-register (Skaug and Øien 2005). The laboratory was 'blinded', i.e. they analyzed the fetus profiles without knowing the mother profiles. Because mother and offspring necessarily must share at least one allele per locus these data provide an opportunity to estimate the error rates of the DNA-register. Most studies aiming at estimating genotyping error rates involve scoring the same individual more than once (Bonin et al. 2004; Paetkau 2003; Pompanon et al. 2005; Waits and Paetkau 2005; Broquet and Petit 2004; Hoffman and Amos 2005). Except for a small part of the DNA-register, which has been scored by independent laboratories, this more direct (and powerful) approach is not available in the present study. Instead, we device new statistical techniques for estimating error rates from mother-offspring pairs of DNA-profiles. Recently, DNA-profiles for a second batch of 298 fetuses have established, using a different genetic laboratory

than was used for Batch 1. The increased sample size provides more accurate estimates, as well as an opportunity to estimate variation in error rates between laboratories.

In the general scientific literature, error rate estimates have been published for genotypes based on human DNA (Ewen et al. 2000) and for genotypes used in ecological studies (Bonin et al. 2004; Hoffman and Amos 2005). The metrics used in literature varies, ranging from error rate per gene copy (allele), locus and PCR reaction to error rate per multilocus genotype (Hoffman and Amos 2005).

2 Material and methods

2.1 Origin of samples and genetic analysis

The establishment of the Norwegian minke whale DNA-register ensures that samples (muscle tissues) are taken of each animal caught under the Norwegian catch quota, and that a DNA-profile is established and stored in a database from each individual whale (Olaisen 1997). The DNA-profile consists of 10 microsatellites (Table 3), mtDNA and a sex-marker (Dupuy and Olaisen 1999). The present study addresses only error rates in the microsatellites. For the period 1997-2002 the genetic analysis have been conducted by the Canadian company Vitatech, while for the period 2003-2006 the analysis are conducted by the Marine Research Institute (MRI) at Island.

Starting from year 2000, tissue samples have also been taken from the fetuses of pregnant females. The present study utilizes DNA profiles from two disjoint batches of fetuses. The first batch was analyzed at Vitatech, and the second batch at MRI, using the same protocol as used for the mothers in the DNA-register (Dupuy and Olaisen 1999). Details about sample sizes are given in Table 3. The design of the study was that both laboratories should be blinded, i.e. they would not know the mother's profile when establishing the fetus profile. By mistake, the fetuses sent to Vitatech were labeled so that the mother could be identified, but apparently this information was not used.

As part of the process of changing genetic laboratory for the DNA-register, 25 individuals previously analyzed by Vitatech was re-analyzed by MRI. Fifteen of these samples had in addition previously been analyzed by two independent laboratories (Cellmark and Norwegian Forensic Institute[1]). These replica arose out of a quality control conducted in 2000 not reported in full detail here. Mismatches between Vitatech and Cellmark were obtained for 19 of 81 individuals. The laboratories were given the chance to revise their genotypes, but the discrepancies were never fully resolved. The Norwegian Forensic Institute was then brought in as an independent party. The relevance of this to the current study is that for 15 of the 25 individuals reanalyzed by MRI we can be quite certain that we know the correct genotype. The genotypes for the 25 individuals produced by MRI all matched the consensus dataset described above at all 10 loci.

2.2 Match criteria and off-ladder alleles

A mother-fetus pair is said to be (Mendelian) consistent at a given locus if they share at least one allele. We introduce the error indicator d, defined as $d = 0$ in case of consistency, and $d = 1$ otherwise. Since one can not tell which of the two fetus gene copies are inherited from the mother, other than by inspecting the actual allelic values, determination of d involves all of the four gene copies that constitute the joint mother-fetus genotype. In absence of typing errors we will always have $d = 0$ for mother-fetus pairs. On the other

[1] Only 8 of the 15 samples

hand, $d = 0$ does not guaranty that no errors has occurred. Clearly, we can only detect errors occurring on the allele shared by decent, but even for that pair of gene copies, an error may be masked by an accidental match with the other gene copies in the genotype. These principles are illustrated in Table 2.

Certain alleles occurring in the DNA-register are not described in the protocol of Dupuy and Olaisen (1999), and are referred to as 'off-ladder' alleles (Table 5). It is not clear how off-ladder alleles should be handled in the practical operation of the DNA-register. One can argue that they should be 'rounded' to the closest allele specified by the allelic ladder (Table 5). We present error rate estimates both with and without applying rounding rules.

A related issue is that of on-ladder alleles differing in length only by one base pair. Since such alleles are difficult to distinguish from each other, one can expect that they have an increased error rate. Thus, we also consider a matching criterion that merges alleles differing only by one base pair (Table 5).

2.3 Statistical methods

At a given locus, let (M_A, M_B) be the mother's genotype, and similarly let (F_A, F_B) be the fetus genotype. By definition, M_A and F_A are the alleles shared by decent, so that $M_A = F_A$. While these are the true genotypes, we denote by $(\tilde{M}_A, \tilde{M}_B)$ and $(\tilde{F}_A, \tilde{F}_B)$ the observed genotypes, i.e. the result of the genetic analysis. In absence of typing errors we must have $M_A = \tilde{M}_A$, and similarly for M_A, F_A and F_B.

The per-allele error rate γ is defined by

$$\gamma = P\left(\tilde{M}_A = M_A\right), \tag{1}$$

and is assumed to be the same for all alleles at a locus, but potentially different across loci. Our goal is to estimate γ based on observations of the error indicator d, defined above. As described in Appendix the probability $P(d = 1)$ depends on γ, as well as the population allele frequencies, and hence we define

$$\tau(\gamma) = P\left(d = 1|\gamma\right). \tag{2}$$

A plot of $\tau(\gamma)$ is shown in Figure (1). Somewhat surprisingly the approximation $\tau(\gamma) \approx \gamma$ holds reasonably well uniformly in γ, although it should be noted that it is only the range $\gamma \in [0, 0.01]$ that is of relevance in the present context.

To estimate γ we solve $\tau(\gamma) = \hat{\tau}$ with respect to γ, where

$$\hat{\tau} = \frac{1}{10N} \sum_{i=1}^{N} \sum_{l=1}^{10} d_{il},$$

with d_{il} being the error indicator for locus l of individual i, and N is the number of individuals. Alternatively, the error rate may be estimated by locus:

$$\hat{\tau}^{(l)} = \frac{1}{N} \sum_{i=1}^{N} d_{il}, \qquad l = 1, \ldots, 10.$$

2.3.1 A crude estimator

A simpler estimate of γ may be obtained under the assumptions that a typing error in M_A or F_A can not be masked by M_B or F_B, and that if both M_A or F_A are erroneously

typed, then we have $\tilde{M}_A \neq \tilde{M}_B$ so that an inconsistency will be detected.[2] In this case $d = 0$ means that no typing error has occurred in either of M_A or F_A, so that $P(d = 1|\gamma) = 1 - (1 - \gamma)^2$. By inverting this relationship we get the 'crude' estimator

$$\hat{\gamma}_c = 1 - \sqrt{1 - \hat{\tau}}. \tag{3}$$

3 Results and discussion

The error rate estimates varied across loci, and were much higher for Batch 1 of fetuses than for Batch 2 (Table 6). Since Batch 1 consisted entirely of individuals (mothers and fetuses) analysed by Vitatech, this leads to the conclusion that the MRI has a much lower error rate than Vitatech. This hypothesis was investigated further by excluding from Batch 2 those mother-fetus pairs where the mother has been analyzed by Vitatech. This left us with a set of 126 mother-fetus pairs. Among these there were only 2 mismatching alleles across all loci, leading to the per-allele error rate estimate $100\hat{\gamma} = 0.16$.

We also derived an independent estimate of γ from the 25 consensus samples described earlier. The DNA-profiles established by MRI matched what is believed to be the true 10-locus genotypes for all 25 individuals. This represents a binomial experiment with $25 \times 2 \times 10 = 500$ trials and success probability γ, leading to the 95% confidence interval $[0.0, 0.74]$ for 100γ.

For the purpose of estimating the error rate, re-typing of the same individual is much more efficient than considering mother-fetus pairs. The mathematics becomes simpler, and the information content in a re-typing is much higher than for a mother-fetus pair. However, as a large number of mother-pairs already are available for North Atlantic minke whales (Skaug and Øien 2005), it was natural for us to develop the methodology described in the Appendix.

It is noted that the crude estimator is severely downwards biased (Table 6), showing that the rather complicated calculations undertaken in the appendix are required.

The analysis is based on several assumptions. First, it is assumed that the error rate does not vary with years. Temporal variations is likely to occur as personnel and equipment in the laboratories may change over time. Hence, the error rates presented in Table 6 should be interpreted as time averages. For similar reasons the error rates in the fetus samples could differ from those in the DNA-register (the mothers).

For the operation of the DNA-register, it is the multilocus error rate which is relevant. As seen from Table 4 the 10-locus error rate ranges from 0.03 to 0.30, depending on dataset and the assumptions made.

Appendix

The purpose of this appendix is to derive the mathematical expression for the function $\tau(\gamma) = P(d = 1|\gamma)$. We shall need the following definitions:

- ϵ is the number of errors occuring at joint mother-fetus locus (M_A, M_B, F_A, F_B), i.e. $\epsilon \in 0, \ldots, 4$.
- D is short hand notation for the event that $d = 1$, where d is the error indicator (see Section 2.3).

[2] These assumptions will approximately be fullfilled if the number of alleles is large.

- $Er(\tilde{M}_A)$ denotes the event that an error occurs in M_A. Essentially, this means that $\tilde{M}_A \neq M_A$, but we include the possibility that the substitued allele is the same as the original (see assumption 3 below). Similarly, $Er(\tilde{M}_B)$, $Er(\tilde{F}_A)$ and $Er(\tilde{F}_B)$ are the events that errors occur in M_B, F_A and F_B, respectively.
- Note that when $\epsilon = 1$, an error in one gene copy implies no error in the other three gene copies.
- When $\epsilon = 2$ we use the shorthand notation $Er(\tilde{M}_A, \tilde{M}_B)$ for the event $Er(\tilde{M}_A) \cap Er(\tilde{M}_B)$, and similarly for the other possibilities.
- When $\epsilon = 3$, $\overline{Er}(\tilde{M}_A)$ denotes the event that M_A is the only gene copy without error.
- At a given locus, $a_1, \ldots, a_n$ is the set of alleles.
- The allele configuration indicator Ψ_{ijk} denotes the event $(M_A = a_i, M_B = a_j, F_B = a_k)$, where we recall that $M_A = F_A$ by descent.
- Population allele frequencies: $f_i = P(a_i)$, $i = 1, \ldots, n$.

In this section we will assume the following:

1. The population is in Hardy-Weinberg equilibrium: $P(M_A = a_i, M_B = a_j) = f_i f_j$.
2. Errors are independently distributed across gene copies.
3. Uniform substitution rates:

$$P\left(\tilde{M}_A = a_i | Er(\tilde{M}_A)\right) = \frac{1}{n}, \qquad i = 1, \ldots, n.$$

In this notation $\tau(\gamma) = P(D;\gamma) = P(D|\epsilon > 0;\gamma)\, P(\epsilon > 0;\gamma)$, where $P(\epsilon > 0;\gamma) = 1 - (1-\gamma)^4$. The rest of this appendix is concerned with finding an expression for $P(D|\epsilon > 0;\gamma)$.

We define $P(q;\gamma) = P(\epsilon = q | \epsilon > 0)$, and since ϵ is binomially distributed with success probability γ, we have

$$P(q;\gamma) = \frac{\frac{4!}{(4-q)!q!}\gamma^q(1-\gamma)^{4-q}}{1-(1-\gamma)^4} \ .$$

It follows that

$$P(D|\epsilon > 0;\gamma) = \sum_{q=1}^{4} P(D, \epsilon = q | \epsilon > 0) = \sum_{q=1}^{4} P(D|\epsilon = q) P(q;\gamma) \ . \tag{4}$$

By conditioning on the allele configuration Ψ we get

$$\begin{aligned} P(D|\epsilon = q) = &\sum_i P(D|\Psi_{iii}, \epsilon = q) P(\Psi_{iii}|\epsilon = q) \\ &+ \sum_{i \neq j} P(D|\Psi_{iij}, \epsilon = q) P(\Psi_{iij}|\epsilon = q) \\ &+ \sum_{i \neq j} P(D|\Psi_{iji}, \epsilon = q) P(\Psi_{iji}|\epsilon = q) \\ &+ \sum_{i \neq j} P(D|\Psi_{jii}, \epsilon = q) P(\Psi_{jii}|\epsilon = q) \\ &+ \sum_{i \neq j \neq k} P(D|\Psi_{ijk}, \epsilon = q) P(\Psi_{ijk}|\epsilon = q) \ , \end{aligned} \tag{5}$$

where $i \neq j \neq k$ means that all three indices are different. By assumption ϵ and Ψ are independent random quantities, and therefore $P(\Psi|\epsilon = q) = P(\Psi)\ \forall\ q$. By assumption 1) we have

$$\begin{aligned}
P(\Psi_{iii}) &= f_i^3 \ , \\
P(\Psi_{iji}) &= f_i^2 f_j \ , \\
P(\Psi_{iij}) &= f_i^2 f_j \ , \\
P(\Psi_{jii}) &= f_j f_i^2 \ , \\
P(\Psi_{ijk}) &= f_i f_j f_k \ .
\end{aligned}$$

Further, by assumption 2)

$$P\left(Er(\tilde{M}_A)|\epsilon = 1\right) = \cdots = P\left(Er(\tilde{F}_B)|\epsilon = 1\right) = \frac{1}{4} \ ,$$

$$\begin{aligned}
P(D|\Psi_{iii}, \epsilon = 1) &= P(D|\Psi_{jii}, \epsilon = 1) \\
&= 0 \\
P(D|\Psi_{iij}, \epsilon = 1) &= P(D|\Psi_{iji}, \epsilon = 1) \\
&= P\left(Er(\tilde{F}_A)|\epsilon = 1\right)(1 - \frac{1}{n}) \\
&= \frac{n-1}{4n} \\
P(D|\Psi_{ijk}, \epsilon = 1) &= P\left(Er(\tilde{M}_A)|\epsilon = 1\right)(1 - \frac{2}{n}) \\
&\quad + P\left(Er(\tilde{F}_A)|\epsilon = 1\right)(1 - \frac{2}{n}) \\
&= \frac{n-2}{2n} \ .
\end{aligned}$$

Thus (5) becomes

$$P(D|\epsilon = 1) = \frac{n-1}{2n} \sum_{i \neq j} f_i^2 f_j + \frac{n-2}{2n} \sum_{i \neq j \neq k} f_i f_j f_k \ . \tag{6}$$

Using the same approach for $\epsilon = 2$, we get

$$\begin{aligned}
P(D|\epsilon = 2) = &\sum_i P(D|\Psi_{iii}, \epsilon = 2) P(\Psi_{iii}) + \sum_{i \neq j} P(D|\Psi_{iij}, \epsilon = 2) P(\Psi_{iij}) \\
&+ \sum_{i \neq j} P(D|\Psi_{iji}, \epsilon = 2) P(\Psi_{iji}) + \sum_{i \neq j} P(D|\Psi_{jii}, \epsilon = 2) P(\Psi_{jii}) \\
&+ \sum_{i \neq j \neq k} P(D|\Psi_{ijk}, \epsilon = 2) P(\Psi_{ijk}) \ .
\end{aligned} \tag{7}$$

Let us look at

$$
\begin{aligned}
& P(D|\Psi_{iii},\epsilon=2) \\
&= P\left(D,Er(\tilde{M}_A,\tilde{M}_B)|\Psi_{iii},\epsilon=2\right)+P\left(D,Er(\tilde{M}_A,\tilde{F}_A)|\Psi_{iii},\epsilon=2\right) \\
&\quad +P\left(D,Er(\tilde{M}_A,\tilde{F}_B)|\Psi_{iii},\epsilon=2\right)+P\left(D,Er(\tilde{F}_A,\tilde{M}_B)|\Psi_{iii},\epsilon=2\right) \\
&\quad +P\left(D,Er(\tilde{F}_A,\tilde{F}_B)|\Psi_{iii},\epsilon=2\right)+P\left(D,Er(\tilde{F}_B,\tilde{M}_B)|\Psi_{iii},\epsilon=2\right) \\
&= P\left(D|Er(\tilde{M}_A,\tilde{M}_B),\Psi_{iii},\epsilon=2\right)P\left(Er(\tilde{M}_A,\tilde{M}_B)|\Psi_{iii},\epsilon=2\right) \\
&\quad +P\left(D|Er(\tilde{M}_A,\tilde{F}_A),\Psi_{iii},\epsilon=2\right)P\left(Er(\tilde{M}_A,\tilde{F}_A)|\Psi_{iii},\epsilon=2\right) \\
&\quad +P\left(D|Er(\tilde{M}_A,\tilde{F}_B),\Psi_{iii},\epsilon=2\right)P\left(Er(\tilde{M}_A,\tilde{F}_B)|\Psi_{iii},\epsilon=2\right) \\
&\quad +P\left(D|Er(\tilde{F}_A,\tilde{M}_B),\Psi_{iii},\epsilon=2\right)P\left(Er(\tilde{F}_A,\tilde{M}_B)|\Psi_{iii},\epsilon=2\right) \\
&\quad +P\left(D|Er(\tilde{F}_A,\tilde{F}_B),\Psi_{iii},\epsilon=2\right)P\left(Er(\tilde{F}_A,\tilde{F}_B)|\Psi_{iii},\epsilon=2\right) \\
&\quad +P\left(D|Er(\tilde{F}_B,\tilde{M}_B),\Psi_{iii},\epsilon=2\right)P\left(Er(\tilde{F}_B,\tilde{M}_B)|\Psi_{iii},\epsilon=2\right)\ .
\end{aligned}
\tag{8}
$$

Because $P\left(Er(\tilde{M}_A,\tilde{M}_B)|\Psi,\epsilon=2\right)=\cdots=P\left(Er(\tilde{F}_A,\tilde{F}_B)|\Psi,\epsilon=2\right)=\frac{1}{6}$, (8) becomes

$$
\begin{aligned}
P(D|\Psi_{iii},\epsilon=2) &= [P\left(D|Er(\tilde{M}_A,\tilde{M}_B),\Psi_{iii},\epsilon=2\right)+P\left(D|Er(\tilde{M}_A,\tilde{F}_A),\Psi_{iii},\epsilon=2\right) \\
&\quad +P\left(D|Er(\tilde{M}_A,\tilde{F}_B),\Psi_{iii},\epsilon=2\right)+P\left(D|Er(\tilde{F}_A,\tilde{M}_B),\Psi_{iii},\epsilon=2\right) \\
&\quad +P\left(D|Er(\tilde{F}_A,\tilde{F}_B),\Psi_{iii},\epsilon=2\right)+P\left(D|Er(\tilde{F}_B,\tilde{M}_B),\Psi_{iii},\epsilon=2\right)]\frac{1}{6} \\
&= \frac{(n-1)^2}{3n^2}\ .
\end{aligned}
\tag{9}
$$

Summing probabilities as above, gives

$$
\begin{aligned}
P(D|\Psi_{iij},\epsilon=2) &= P(D|\Psi_{iji},\epsilon=2) \\
&= [P\left(D|\Psi_{iij},\epsilon=2,Er(\tilde{M}_A,\tilde{M}_B)\right)+P\left(D|\Psi_{iij},\epsilon=2,Er(\tilde{F}_A,\tilde{F}_B)\right) \\
&\quad +P\left(D|\Psi_{iij},\epsilon=2,Er(\tilde{M}_A,\tilde{F}_A)\right)+P\left(D|\Psi_{iij},\epsilon=2,Er(\tilde{M}_A,\tilde{F}_B)\right) \\
&\quad +P\left(D|\Psi_{iij},\epsilon=2,Er(\tilde{F}_A,\tilde{M}_B)\right)+P\left(D|\Psi_{iij},\epsilon=2,Er(\tilde{M}_B,\tilde{F}_B)\right)]\frac{1}{6} \\
&= \frac{4n^2-12n+11}{6n^2}\ .
\end{aligned}
\tag{10}
$$

Further,

$$
\begin{aligned}
P(D|\Psi_{jii},\epsilon=2) &= [P\left(D|\Psi_{jii},\epsilon=2,Er(\tilde{M}_A,\tilde{M}_B)\right)+P\left(D|\Psi_{jii},\epsilon=2,Er(\tilde{F}_A,\tilde{F}_B)\right) \\
&\quad +P\left(D|\Psi_{jii},\epsilon=2,Er(\tilde{M}_A,\tilde{F}_A)\right)+P\left(D|\Psi_{jii},\epsilon=2,Er(\tilde{M}_A,\tilde{F}_B)\right) \\
&\quad +P\left(D|\Psi_{jii},\epsilon=2,Er(\tilde{F}_A,\tilde{M}_B)\right)+P\left(D|\Psi_{jii},\epsilon=2,Er(\tilde{M}_B,\tilde{F}_B)\right)]\frac{1}{6} \\
&= \frac{2n^2-7n+7}{3n^2}\ ,
\end{aligned}
\tag{11}
$$

and

$$\begin{aligned}P(D|\Psi_{ijk},\epsilon=2) &= [P\left(D|\Psi_{ijk},\epsilon=2,Er(\tilde{M}_A,\tilde{M}_B)\right)+P\left(D|\Psi_{ijk},\epsilon=2,Er(\tilde{F}_A,\tilde{F}_B)\right)\\ &\quad+P\left(D|\Psi_{ijk},\epsilon=2,Er(\tilde{M}_A,\tilde{F}_A)\right)+P\left(D|\Psi_{ijk},\epsilon=2,Er(\tilde{M}_A,\tilde{F}_B)\right)\\ &\quad+P\left(D|\Psi_{ijk},\epsilon=2,Er(\tilde{F}_A,\tilde{M}_B)\right)+P\left(D|\Psi_{ijk},\epsilon=2,Er(\tilde{M}_B,\tilde{F}_B)\right)]\frac{1}{6}\\ &= \frac{5n^2-17n+17}{6n^2}\,.\end{aligned} \tag{12}$$

Finally, using the results from (9) - (12), (7) becomes

$$P(D|\epsilon=2)=\frac{(n-1)^2}{3n^2}\sum_i f_i^3+\frac{6n^2-19n+18}{3n^2}\sum_{i\neq j}f_i^2f_j+\frac{5n^2-17n+17}{6n^2}\sum_{i\neq j\neq k}f_if_jf_k\,. \tag{13}$$

For $\epsilon=3$

$$\begin{aligned}P(D|\epsilon=3) &= P(D|\Psi_{iii},\epsilon=3)P(\Psi_{iii})+2P(D|\Psi_{iij},\epsilon=3)P(\Psi_{iij})\\ &\quad+P(D|\Psi_{jii},\epsilon=3)P(\Psi_{jii})+P(D|\Psi_{ijk},\epsilon=3)P(\Psi_{ijk})\,.\end{aligned} \tag{14}$$

Because $P\left(\overline{Er}(\tilde{M}_A)|\Psi_{iij},\epsilon=3\right)=\cdots=P\left(\overline{Er}(\tilde{F}_B)|\Psi_{iij},\epsilon=3\right)=\frac{1}{4}$,

$$P(D|\Psi_{iii},\epsilon=3)=\frac{(n-1)(n-2)^2+(n-1)^2}{n^3}\,, \tag{15}$$

and

$$\begin{aligned}P(D|\Psi_{iji},\epsilon=3) &= P\left(D,\overline{Er}(\tilde{M}_A)|\Psi_{iij},\epsilon=3\right)+P\left(D,\overline{Er}(\tilde{M}_B)|\Psi_{iij},\epsilon=3\right)\\ &\quad+P\left(D,\overline{Er}(\tilde{F}_A)|\Psi_{iij},\epsilon=3\right)+P\left(D,\overline{Er}(\tilde{M}_A)|\Psi_{iij},\epsilon=3\right)\\ &= [3P\left(D|\overline{Er}(\tilde{M}_A),\Psi_{iij},\epsilon=3\right)+P\left(D|\overline{Er}(\tilde{F}_B),\Psi_{iij},\epsilon=3\right)]\frac{1}{4}\\ &= \frac{(n-1)(n-2)^2+(n-1)^2}{n^3}\,.\end{aligned} \tag{16}$$

Further

$$\begin{aligned}P(D|\Psi_{jii},\epsilon=3) &= \left(2P\left(D|\overline{Er}(\tilde{M}_A),\Psi_{jii},\epsilon=3\right)+2P\left(D|\overline{Er}(\tilde{M}_B),\Psi_{jii},\epsilon=3\right)\right)\frac{1}{4}\\ &= \frac{(n-1)(n-2)^2+(n-1)^2}{n^3}\,,\end{aligned} \tag{17}$$

and finally

$$\begin{aligned}P(D|\Psi_{ijk},\epsilon=3) &= [2P\left(D|\overline{Er}(\tilde{M}_A),\Psi_{ijk},\epsilon=3\right)+P\left(D|\overline{Er}(\tilde{M}_B),\Psi_{ijk},\epsilon=3\right)\\ &\quad+P\left(D|\overline{Er}(\tilde{F}_B),\Psi_{ijk},\epsilon=3\right)]\frac{1}{4}\\ &= \frac{(n-1)(n-2)^2+(n-1)^2}{n^3}\,.\end{aligned} \tag{18}$$

Summing (15), (16), (17) and (18), (14) becomes

$$P(D|\Psi,\epsilon=3)=\frac{(n-1)(n-2)^2+(n-1)^2}{n^3}(\sum_i f_i^3+2\sum_{i\neq j}f_i^2f_j+\sum_{i\neq j\neq k}f_if_jf_k)\ . \quad (19)$$

For $\epsilon=4$

$$P(D|\Psi,\epsilon=4)=\frac{(n-1)(n-2)^2+(n-1)^2}{n^3}\sum_{i,j,k}f_if_jf_k\ . \quad (20)$$

Note that in (20), i,j and k may be equal to each other. Substituting (6), (13), (19) and (20) into (4) we get

$$\begin{aligned}P\left(D|\epsilon>0;\gamma\right)=&[\frac{n-1}{2n}\sum_{i\neq j}f_i^2f_j+\frac{n-2}{2n}\sum_{i\neq j\neq k}f_if_jf_k]P(1;\gamma)\\&+[\sum_i\frac{(n-1)^2}{3n^2}f_i^3+\frac{6n^2-19n+18}{3n^2}\sum_{i\neq j}f_i^2f_j\\&+\frac{5n^2-17n+17}{6n^2}\sum_{i\neq j\neq k}f_if_jf_k]P(2;\gamma)\\&+\frac{(n-1)(n-2)^2+(n-1)^2}{n^3}[\sum_i f_i^3+2\sum_{i\neq j}f_i^2f_j\\&+\sum_{i\neq j\neq k}f_if_jf_k]P(3;\gamma)\\&+\frac{(n-1)(n-2)^2+(n-1)^2}{n^3}\sum_{i,j,k}f_if_jf_kP(4;\gamma)\ ,\end{aligned} \quad (21)$$

and our goal is achieved.

References

Anon. Progress report from Norway to be presented to the Scientific Committee of the International Whaling Commission, May, 2007. 2007.

A. Bonin, E. Bellemain, P. Bronken Eidesen, F. Pompanon, C. Brochmann, and P. Taberlet. How to track and assess genotyping errors in population genetics studies. *Molecular Ecology*, 13(11):3261–3273, Nov 2004.

Thomas Broquet and Eric Petit. Quantifying genotyping errors in noninvasive population genetics. *Molecular Ecolology*, 13(11):3601–3608, Nov 2004.

B. Dupuy and B. Olaisen. Typing procedure for the Norwegian minke whale DNA register. 1999.

K. R. Ewen, M. Bahlo, S. A. Treloar, D. F. Levinson, B. Mowry, J. W. Barlow, and S. J. Foote. Identification and analysis of error types in high-throughput genotyping. *American Journal of Human Genetics*, 67(3):727–736, 2000.

J. I. Hoffman and W. Amos. Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology*, 14(2): 599–612, Feb 2005.

B. Olaisen. Proposed specification for a Norwegian DNA database register for minke whales. Paper SC/49/NA1 presented to the Scientific Committee of the International Whaling Commission, 1997.

D. Paetkau. An empirical exploration of data quality in DNA-based population inventories. *Molecular Ecology*, 12(6):1375–1387, Jun 2003.

P. Palsbøll, M. Bérubé, H.J. Skaug, and C. Raymakers. DNA Registers of Legally Obtained Wildlife and Derived Products as Means to Identify Illegal Takes. *Conservation Biology*, 20(4):1284–1293, 2006.

François Pompanon, Aurélie Bonin, Eva Bellemain, and Pierre Taberlet. Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, 6(11):847–859, Nov 2005.

Hans J. Skaug and Nils Øien. Genetic tagging of male North Atlantic minke whales through comparison of maternal foetal DNA-profiles. *Journal of Cetecean Research and management*, 7(2):113–117, 2005.

Lisette P. Waits and David Paetkau. Noninvasive genetic sampling tools for wildlife biologists: A review of applications and recommendations for accurate data collection. *Journal of Wildlife Managemend*, 69(4):1419–1433, 2005.

Tables and figures

Locus	Mother		Fetus		d_l
GATA098	**83**	**91**	**83**	**91**	0
GT509	199	**207**	193	**207**	0
EV001	155	**157**	149	**157**	0
EV037	205	205	203	203	1
GT310	**113**	**113**	**113**	117	0
GT211	**104**	**104**	**104**	110	0
GT575	148	**154**	**154**	156	0
GT023	105	109	97	103	1
GATA028	**211**	**211**	161	**211**	0
GATA417	217	**240**	209	**240**	0

Table 1: Example of mother-fetus pair (#0003073) with inconsistent loci. Matching alleles are shown in boldface. Loci EV037 and GT023 had no matching alleles, so that the total number of mismatches is $\sum_{l=1}^{10} d_l = 2$.

Mother:	$M_A = a_i$	$M_B = a_j$	$\overset{observed}{\longrightarrow}$	$\tilde{M}_A = a_i$	$\tilde{M}_B = a_j$
Fetus:	$F_A = a_i$	$F_B = a_k$		$\tilde{F}_A = a_i$	$\tilde{F}_B = a_m$

Table 2: Illustration of mother-fetus allele configuration at a locus: (M_A, M_B) is the mother's true genotype and (F_A, F_B) the fetus' true genotype. Because M_A and F_A are shared by descent one must have $M_A = F_A$. The alleles are $a_1, ..., a_n$. The symbol '$\sim$' denotes the observed genotypes, and an error has occurred in F_B.

Batch	Period	Mothers		Fetuses	
		Vitatech	MRI	Vitatech	MRI
1	2000-2002	291	0	291	0
2	2002-2003	172	126	0	298

Table 3: Number of samples by genetic laboratory used in this study.

Study	Samples	Error rates (%)			
		GC	React	Sing	Multi
Bonin et al.	Tissue			0.8	17.6
	Feces			2.0	2.1
Ewen et al.	Set 1, concordance			0.16	
	Set 1, Mendelian			0.25	
	Set 2, Mendelian			1.37	
	Set 2, between gels			2.38	
	Set 2, within gels			0.76	
Hoffman and Amos	Repeat-genotype	0.22	0.38		
	Del. res. females	0.14	0.28		
	Del. res. males	0.21	0.42		
	Unint. res. males	0.37	0.74		
DNA-register	Batch1-Original	1.77		3.52	0.30
	Batch1-Off-ladder	1.54		3.05	0.27
	Batch1-Rounded	1.38		2.74	0.24
	Batch2-Original	0.38		0.78	0.07
	Batch2-Off-ladder	0.25		0.47	0.05
	Batch2-Rounded	0.16		0.31	0.03

Table 4: Comparison of error rate estimates for the minke whale DNA-register based on mother-fetus pairs, with estimates of error rates from other studies (Bonin et al. 2004; Ewen et al. 2000; Hoffman and Amos 2005). The metrics are error rate per allele/gene copy (GC), per reaction (React), per single locus (Sing) and across multilocus genotype (Multi). For the DNA-register the column Multi refers to the 10-locus genotype. The samples in Hoffman and Amos' study are the ones from repeated-genotyping, deliberately resampled females, deliberately resampled males and unintentionally resampled males, respectively.

Locus	Allele	Rounded value
GATA417	221*	220
GATA417	233*	232
GATA417	241	240
GATA417	245	244
GATA417	248	249
EV001	172	171 or 173
GT509	206	205 or 207

Table 5: Rounding scheme employed when matching mother and fetus alleles. For instance, at locus EV001 the allele 172 will be taken to match both 171 and 173. Off-ladder alleles are denoted by a "*".

Batch 1 ($N = 291$)

		Original			Off-ladder			Rounded		
	H	$\sum d$	$100\hat{\gamma}$	Crude	$\sum d$	$100\hat{\gamma}$	Crude	$\sum d$	$100\hat{\gamma}$	Crude
GATA098	0.28	5	1.95	0.86	5	1.95	0.86	5	1.95	0.86
GT509	0.19	6	1.71	1.04	6	1.71	1.04	6	1.71	1.04
EV001	0.17	1	0.27	0.17	1	0.27	0.17	1	0.27	0.17
EV037	0.34	6	2.38	1.04	6	2.38	1.04	6	2.38	1.04
GT310	0.32	1	0.42	0.17	1	0.42	0.17	1	0.42	0.17
GT211	0.22	6	1.95	1.04	6	1.95	1.04	6	1.95	1.04
GT575	0.23	4	1.30	0.68	4	1.30	0.68	4	1.30	0.68
GT023	0.22	5	1.63	0.86	5	1.63	0.86	5	1.63	0.86
GATA028	0.18	7	1.95	1.21	7	1.95	1.21	5	1.39	0.86
GATA417	0.15	16	4.18	2.79	7	1.82	1.21	3	0.78	0.52
Mean	0.23	5.7	1.77	0.99	4.8	1.54	0.83	4.2	1.38	0.72
Lower			1.33	0.17	-	1.11	0.08	-	0.97	0.02
Upper	-	-	2.22	1.81	-	1.97	1.58	-	1.78	1.43

Batch 2 ($N = 298$)

		Original			Off-ladder			Rounded		
	H	$\sum d$	$100\hat{\gamma}$	Crude	$\sum d$	$100\hat{\gamma}$	Crude	$\sum d$	$100\hat{\gamma}$	Crude
GATA098	0.28	0	0	0	0	0	0	0	0	0
GT509	0.19	0	0	0	0	0	0	0	0	0
EV001	0.17	0	0	0	0	0	0	0	0	0
EV037	0.34	1	0.39	0.17	1	0.39	0.17	1	0.39	0.17
GT310	0.32	0	0	0	0	0	0	0	0	0
GT211	0.22	2	0.64	0.34	2	0.64	0.34	2	0.64	0.34
GT575	0.23	0	0	0	0	0	0	0	0	0
GT023	0.22	0	0	0	0	0	0	0	0	0
GATA028	0.18	4	1.09	0.67	4	1.09	0.67	2	0.54	0.34
GATA417	0.15	7	1.78	1.18	1	0.25	0	0	0	0
Mean	0.23	1.4	0.39	0.24	0.8	0.24	0.12	0.5	0.16	0.09
Lower	-	-	0.26	0.00	-	0.13	0.00		0.08	0.00
Upper	-	-	0.52	0.64	-	0.34	0.44		0.23	0.32

Table 6: Number of inconsistent mother-fetus pairs ($\sum_{i=1}^{N} d_{il}$) by locus, and the corresponding estimated error rates in % ($100\hat{\gamma}$). Also shown is the crude estimate (3). In the column "Original" all data are included, in "Off-ladder" only the off-ladder alleles are corrected, while in "Rounded" all corrections of Table 5 are applied. The column 'H' shows the homozygosity of the loci. Bottom row "Mean" gives the average of the table. 'Lower' and 'Upper' are the averages of by-locus approximate lower and upper 95% confidence limits.

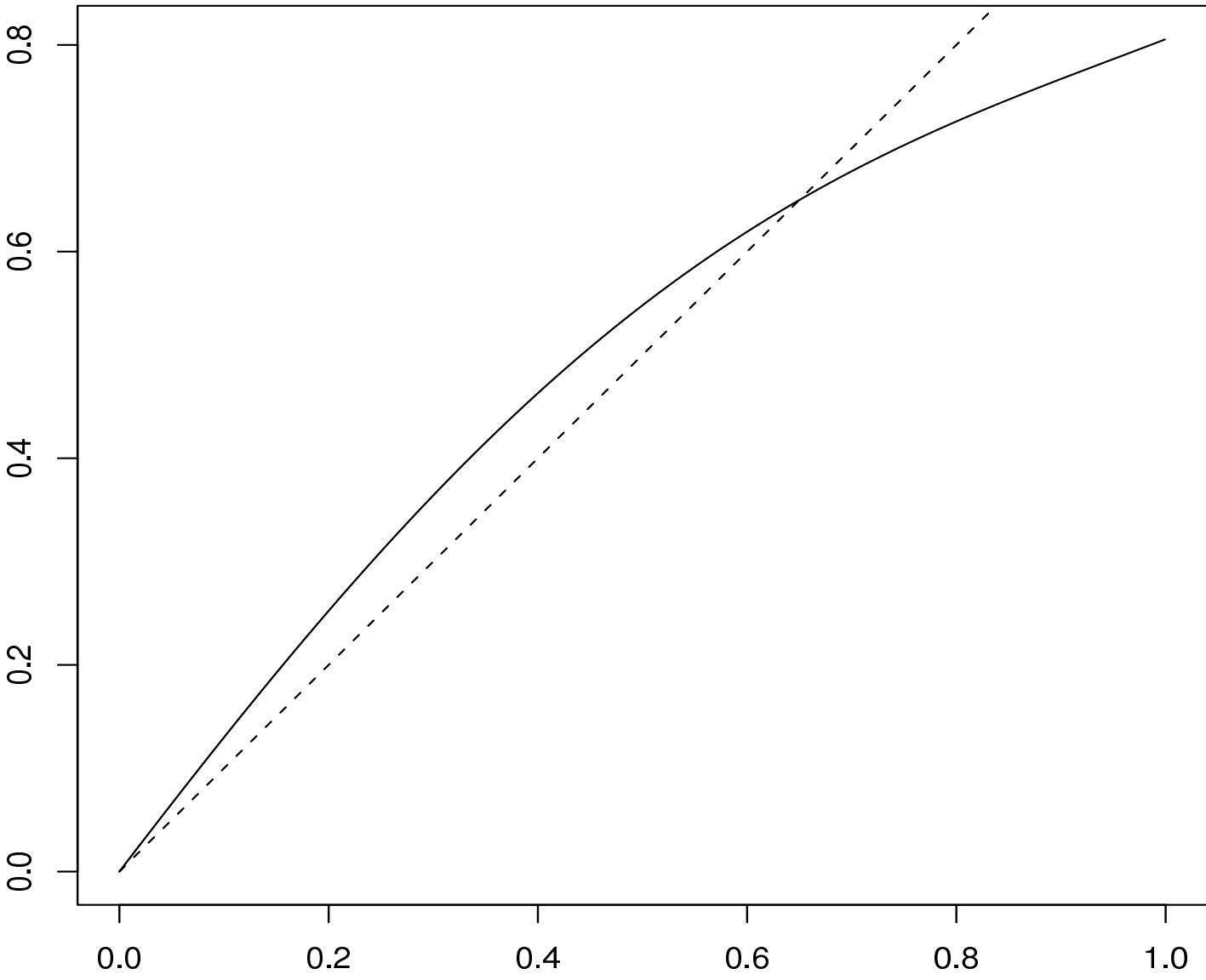

Figure 1: The function $\tau(\gamma)$ (solid line) as defined in (2), and the line '$x = y$' (slope 1 and going through the origin).