Statistics and Probability Letters 83 (2013) 812–819

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Estimating genotyping error rates from parent–offspring dyads

Øystein A. Haaland[a], Hans J. Skaug[a,b,*]

[a] Department of Mathematics, University of Bergen, Johannes Brunsgate 12, 5008 Bergen, Norway

[b] Institute of Marine Research. P.O. Box 1870, Nordnes. N-5817 Bergen, Norway

ARTICLE INFO

Article history:
Received 11 April 2011
Received in revised form 2 November 2012
Accepted 11 November 2012
Available online 3 December 2012

Keywords:
DNA register
Parent–offspring dyad
Genotyping error
Pedigree
Bias

ABSTRACT

A common approach when estimating the error rate of a DNA register is to genotype some of the individuals twice. This may be both expensive and time consuming. As an alternative, we present a new method for estimating genotyping errors based on parent–offspring dyads. The basic idea is that parent and offspring must share at least one allele per locus. Others have previously devised similar techniques, but depended on the assumption that at most one error may occur per dyad. In this paper we examine the bias caused by this simplification. Further, we apply our method on a data set from the Norwegian minke whale DNA register, and find that the error rates are in the range of 0–0.0418, which is comparable to those in the published literature.

1. Introduction

DNA registers are most widely known for their use in forensics, but such data bases are also playing an increasingly important role in other fields, such as ecology (Blouin, 2003) and medicine (Ewen et al., 2000). Typing errors, i.e., the occurrence of errors in the genetic analysis causing the inferred and true genotypes to differ, may severely affect the conclusions drawn from DNA profile data if not accounted for Pompanon et al. (2005). Most studies aiming at estimating such rates involve scoring the same individual more than once (Paetkau, 2003; Bonin et al., 2004; Hoffman and Amos, 2005; Tintle et al., 2007; Haaland et al., 2011). However, this powerful and direct approach is not always available, e.g., due to financial reasons, or if the organism is small (Steinauer et al., 2008). In wildlife populations, mother–offspring dyads are often easy to identify (Hoffman and Amos, 2005), and in medical research it is common to collect DNA profiles on pedigrees (Ewen et al., 2000; Sobel et al., 2002). Error rate estimation from such data, and parent–offspring data in particular, is the theme of this paper.

At a locus an individual has two gene copies, one inherited from each parent, that are selected from a set of n possible alleles $(a_1, \dots, a_n)$, so a parent and an offspring must necessarily share at least one allele per locus. This motivated the development of statistical techniques enabling us to estimate genotyping error rates from parent–offspring dyads of DNA profiles. Similar methods have been presented by others (Douglas et al., 2002; Saunders et al., 2007), but rest on the assumption that no more than one error occurs in each family under consideration (in our case this translates to one error per parent–offspring dyad). We also consider microsatellites (more than two alleles per locus) instead of SNPs (two alleles per locus) (Douglas et al., 2002).

[*] Corresponding author at: Department of Mathematics, University of Bergen, Johannes Brunsgate 12, 5008 Bergen, Norway.
E-mail address: hans.skaug@math.uib.no (H.J. Skaug).

0167-7152/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2012.11.009

Table 1
Illustration of mother-fetus allele configuration at a locus. (P_A, P_B) is the mother's true genotype and (O_A, O_B) is the fetus' true genotype, while $(\tilde{P}_A, \tilde{P}_B)$ and $(\tilde{O}_A, \tilde{O}_B)$ are the corresponding observed genotypes. Because P_A and O_A are shared by descent one must have $P_A = O_A$. The different alleles are represented by a_i, a_j, a_k and a_m. Note that an error has occurred in O_B.

P_A	P_B	$\tilde{P}_A$	$\tilde{P}_B$
O_A	O_B	$\tilde{O}_A$	$\tilde{O}_B$
↓		↓	
a_i	a_j	a_i	a_j
a_i	a_k	a_i	a_m

A parent–offspring dyad is said to be (Mendelian) consistent at a given locus if they share at least one allele. We introduce the error indicator

$$e = \begin{cases} 1, & \text{dyad inconsistent} \\ 0, & \text{otherwise.} \end{cases}$$

Since one can not tell which of the offspring's two gene copies are inherited from which parent, other than by inspecting the actual allelic values, determination of e involves all of the four gene copies that constitute the joint parent–offspring genotype. In the absence of typing errors we will always have $e = 0$ for parent–offspring dyads. On the other hand, $e = 0$ does not guarantee that no errors have 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 accidentally matching the remainder of the genotype. These principles are illustrated in Table 1. The main result in this paper is the derivation of (14) below, describing the relationship between the observed error rate and the actual error rate.

We denote by γ the error rate per gene copy, which 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 e. Letting $g \in \{0, \ldots, 4\}$ denote the number of errors occurring in a dyad at a locus, and E be shorthand notation for $e = 1$, we define the observed error rate

$$\tau_1(\gamma) \overset{def}{=} P(E; \gamma). \tag{1}$$

The subscript '1' indicates that we have $g \leq 4$. We estimate τ_1 by calculating the proportion of observed errors in the sample, $\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} e_i$, where N is the number of sampled parent–offspring dyads. A moment estimator, $\hat{\gamma}$, is obtained by solving $\tau_1(\gamma) = \hat{\tau}$ with respect to γ. Assuming $g \leq 1$, estimates of γ will be biased.

The Norwegian minke whale DNA register (NMDR) contains DNA profiles of almost every individual whale caught by Norway since 1997 (Olaisen, 1997; Anon, 0000), and from the year 2000 tissue samples from the fetuses of pregnant females were also included. In addition to providing error rate estimates for the NMDR by applying parent–offspring techniques, we study the bias introduced by the one-error assumption of Douglas et al. (2002).

The paper is organized as follows. Section 2 addresses the calculation of error rates and the underlying assumptions. Section 2.3 makes the assumption that there can be maximum of one error per parent–offspring dyad, and finds and quantifies the resulting bias. In Section 3 we apply the error rate estimators to the NMDR.

2. Error probabilities

Conditioning on $g > 0$ in (1), we get

$$\tau_1(\gamma) = P(E|g > 0; \gamma)P(g > 0; \gamma), \tag{2}$$

where

$$P(g > 0; \gamma) = 1 - (1 - \gamma)^4. \tag{3}$$

As mentioned earlier, Douglas et al. (2002) and Saunders et al. (2007) simplify their calculations by ruling out the possibility of more than one error occurring per parent–offspring dyad, i.e., they assume $g = 0$ or $g = 1$ (from now on called the Douglas assumption). In our setting, this assumption yields the following approximation of τ_1:

$$\tau_2(\gamma) \overset{def}{=} P(E|g = 1; \gamma)P(g > 0; \gamma), \tag{4}$$

where the subscript '2' signifies $g \leq 1$. It will be shown later that both (2) and (4) are polynomials of degree 4 with no constant term, and we may therefore write

$$\tau_i(\gamma) = \sum_{k=1}^{4} c_{ik}\gamma^k, \tag{5}$$

where $i \in \{1, 2\}$ is an index for the method.