Conserv Genet (2010) 11:1567–1571
DOI 10.1007/s10592-009-9950-9

SHORT COMMUNICATION

A new method for quantifying genotyping errors for noninvasive genetic studies

Xiangjiang Zhan · Xiudeng Zheng · Michael W. Bruford · Fuwen Wei · Yi Tao

Received: 22 March 2009 / Accepted: 29 May 2009 / Published online: 18 June 2009

Abstract More and more noninvasive genetic data are being produced but a general methodology to quantify genotyping error rates from non-pilot data remains lacking. Here we propose a mathematical approach to estimate genotyping error rates by exploring the relationship between errors and PCR replicates. This method can be used to quantify the error rates for either the multi-tubes approach designed by Taberlet et al. (Nucleic Acids Res 24: 3189–3194, 1996) or the pilot method by Prugh et al. (Mol Ecol 14: 1585–1596, 2005).

Keywords Genotyping errors · Noninvasive genetics · Quantification · Multi-tubes approach

Xiangjiang Zhan and Xiudeng Zheng contribute equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s10592-009-9950-9) contains supplementary material, which is available to authorized users.

X. Zhan · X. Zheng · F. Wei (✉) · Y. Tao (✉)
Key Lab of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, No 1, Beichenxilu, 100101 Beijing, People's Republic of China
e-mail: weifw@ioz.ac.cn

Y. Tao
e-mail: yitao@ioz.ac.cn

X. Zhan
e-mail: zhanxj@cf.ac.uk

X. Zhan · M. W. Bruford
Biodiversity and Ecological Processes Group, Cardiff School of Biosciences, Cardiff University, Cardiff CF10 3AX, UK

X. Zheng
Graduate School of Chinese Academy of Sciences, Beijing 100039, People's Republic of China

Introduction

In recent years, noninvasive genetic sampling has been increasingly applied to study the population biology of wild animals, especially elusive species. It offers the advantage that it can provide DNA profiles without causing undue disturbance using material such as feces and hair, enabling the genetic tagging of individuals. These approaches have led to such applications as the identification of individuals (Taberlet et al. 1997), estimation of population size (Solberg et al. 2006; Zhan et al. 2006), and analysis of dispersal patterns (Flagstad et al. 2004; Zhan et al. 2007).

While many researchers welcome the application of noninvasive genetic sampling, they caution that a major problem for the technology, particularly for microsatellite DNA profiling, is inherent genotyping error, where the observed genotype of an individual does not correspond to the true genotype. A bibliographic survey has indicated that many genetics studies in which errors were checked report non-negligible rates (Pompanon et al. 2005), and relatively higher rates can be found in studies using noninvasive genetic samples (Broquet et al. 2007). So far, two main sources of DNA profiling errors have been recognized: one is allelic dropout, where one or more allele of an individual's locus is not detected by PCR leading to incorrect scoring as a homozygote or failed sample, thought to be resulting from low DNA copy number in the sample (Taberlet et al. 1996; Valière 2002); the other is false allele generation, namely, one or more extra 'alleles' can be produced by PCR as a result of experimental artifacts, sporadic contamination, as well as human error (Taberlet et al. 1996; Pompanon et al. 2005).

Although genotyping errors occur in all but the smallest datasets that are generated in non-invasive genetic studies,

they can greatly bias the final conclusions. For example, Creel et al. (2003) reported that a 5% error rate per locus when using seven to ten microsatellite loci for genotype identification could cause a 200% overestimate of population size. Various methods, thus, have been proposed to limit genotyping errors in noninvasive studies and the most accepted is the "multi-tubes" method (Taberlet et al. 1996), which includes two basic steps: (1) three PCR repeats are carried out and a heterozygote is determined if both alleles are repeated at least twice among three positive PCRs, and (2) otherwise, four additional repeats are needed before homo-/heterozygotes are confirmed. However, there has been some debate on the best method to use and one criticism of the multi-tubes method is that it cannot quantify genotyping error rates if there is no reference genotype obtained using other sources of 'non-erroneous' DNA (Creel et al. 2003; Mckelvey and Schwartz 2004). Some studies have instead estimated genotyping error rates directly by using a pilot study designed to reveal and quantify error (e.g., Prugh et al. 2005). Different from the multi-tubes method, the pilot study normally carries out a pre-experiment that amplifies each sample at a locus with the same number of PCR repeats to construct the reference genotype. The researcher then utilizes the error rates inferred from the pre-experiment to estimate the genotyping errors of the final dataset. Recently, the use of pilot studies has become more popular due to its direct method of inference. However, the accuracy of this method depends on two factors: firstly, that these rates are correctly estimated, and secondly, the samples chosen for the pilot study accurately reflect all samples in the population. Both problems are difficult to entirely obviate under most circumstances, not only because amplification rates may vary among individuals as a result of dietary preference, age, sex and location; but also because, until now, there is no evidence that the PCR replicates adopted in pilot studies were sufficient to accurately estimate the intrinsic error rates of whole samples.

In the present study, we aimed to develop a new general method for noninvasive genetics studies, (1) to estimate the genotyping error rates by using the full data instead of a proportion of the genotypes; (2) to directly quantify the genotyping error rates for studies using the multi-tubes method regardless of the number of PCR replicates or when adopting the pilot method with or without replicates.

Materials and methods

For noninvasive genetics studies, because most genotyping errors have a measurable probability to occur in PCR, we can expect that there is a relationship between genotyping errors and the number of replicates of PCR, i.e., the more PCR replicates the closer experiment approximates to the true data. In our analysis, we divide the genotyping errors into two classes, allelic dropout and false allele generation, and conservatively assume each class independently happens within each allele with an equal error probability. Then for a certain locus amplified by PCR replicates, we calculate the probability of genotyping error for a heterozygous and homozygous sample, respectively. Finally, we use the mean error probability to measure the genotyping error in the population.

Definitions and assumptions

1. Consider only the identification of heterozygotes and homozygotes for each locus.
2. Let μ_i and v_i denote the per-allele dropout rate and false allele generation rate, respectively, at locus i ($i = 1, 2, \ldots, N$, where N is the total number of loci).
3. In the absence of a reference genotype, to construct a consensus genotype for each sample and for each locus, general protocols to date use a criterion that replicate PCRs produce at least n apparent heterozygous genotypes (for example two under the Taberlet approach) or at least m apparent homozygote genotypes (for example six under the Taberlet approach). For convenience, this protocol is denoted by $\Lambda_{n,m}$. In general, we take $n < m$ since $\mu_i > v_i$ ($i = 1, 2, \ldots, N$; Broquet and Petit 2004). Table 1 shows an example of how to use the protocol $\Lambda_{2,3}$ to construct a consensus genotype.
4. For the protocol $\Lambda_{n,m}$, the number of PCR replicates, denoted by L, should be theoretically equal to $n + m - 1$ since (a) if $L < n+m-1$, the existence of the consensus genotype cannot be always guaranteed, for example, for the PCR replicate data 'AB/AA/AA', we cannot use the protocol $\Lambda_{2,3}$ to construct consensus genotype since it is not possible to tell whether the genotype is heterozygote or homozygote; (b) however, if $L > n+m-1$, a consensus genotype

Table 1 The protocol for constructing consensus genotype with $n = 2$ and $m = 3$

Number of replications	Replication results	Consensus genotype
2	AB/AB	Heterozygote
3	AB/AA/AB	Heterozygote
4	AB/AA/AB/AA	Heterozygote
4	AB/AA/AA/AA	Homozygote
4	AA/AB/AA/AA	Homozygote
4	AA/AA/AB/AA	Homozygote
3	AA/AA/AA	Homozygote

may be difficult to produce, for example, for the PCR replication data 'AB/AA/AA/AB/AA', we cannot also use the protocol $\Lambda_{2,3}$ to determine the genotype.

Theoretical analyses of genotyping errors

At locus i the probability of genotyping error for a heterozygous sample is

$$\tilde{\phi}_i = \sum_{k=0}^{n-1} C_{n+m-1}^{k} (1-\mu_i)^k \mu_i^{n+m-1-k}. \quad (1)$$

Similarly, the probability of genotyping error for a homozygous sample is

$$\tilde{\vartheta}_i = \sum_{k=0}^{m-1} C_{n+m-1}^{k} (1-\upsilon_i)^k \upsilon_i^{n+m-1-k}. \quad (2)$$

Let q_i and $1 - q_i$ denote the proportions of heterozygous and homozygous samples in the sample set, respectively, at locus i ($i = 1, 2, \ldots, N$). The mean probability of genotyping error at locus i is given by

$$P_i = q_i \tilde{\phi}_i + (1-q_i)\tilde{\vartheta}_i. \quad (3)$$

Some related equations are shown in the Supplemental Material. For a multi-locus genotype, we assume that for errors, loci are independent of each other. Thus, the mean genotyping error of the multi-locus profile can be measured by

$$P = \frac{1}{N}\sum_{i=1}^{N} P_i, \quad (4)$$

i.e., for all N loci, the mean probability of genotyping error is P.

Results and discussions

Effects of allelic dropout and false allele rates on genotyping errors

Theoretically, if both allelic dropout and false allele rates are small enough, then the genotyping error rate should also be small enough using reasonable protocols. In order to show this, using the protocols $\Lambda_{2,3}$ and $\Lambda_{2,6}$, for different values of the allelic dropout rate (μ_i), false allele rate (v_i) and proportion of heterozygous samples in the sample set (q_i), the mean probability of genotyping error at a single locus (P_i) is plotted in Fig. 1.

In Fig. 1a, b, It is obvious that P_i is an increasing function of μ_i and v_i, namely, the increase of genotyping error is related to an increase in allelic dropout and false allele rates. However, when m is comparatively large, e.g., 6, P_i will only be sensitive to an increase of v_i (compare Fig. 1a, b with c, d), which results from an indexed increase in the equations on which P_i is based. It also indicates that the protocol of $\Lambda_{2,6}$ effectively minimizes the

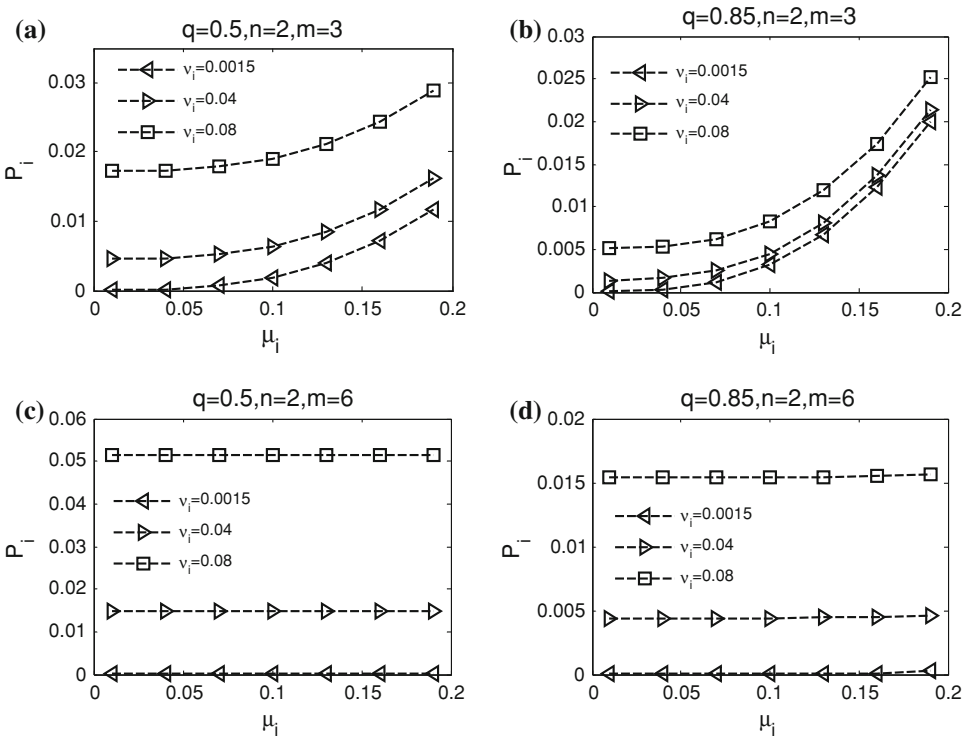

Fig. 1 Effects of allelic dropout (μ_i) and false allele rate (v_i) on the probability of genotyping error at single locus (see Eq. 4). The heterozygosity (q) is set to two levels: 0.5 and 0.85 as Johnson and Haydon (2006). PCR replicates for heterozygous samples (n) are 2 and the replicates for homozygous samples (m) are set to 3 and 6, respectively

Table 2 Genotyping error rate estimates for giant panda fecal samples in Wanglang Nature Reserve

Locus	Number of heterozygote	Number of homozygote	Proportion of heterozygote	$\tilde{\phi}_i^{\min}$	$\tilde{\phi}_i^{\max}$	$\tilde{\vartheta}_i$	$P_i^{\min}$	$P_i^{\max}$
Ame-μ5	239	80	0.7492	7.87E−09	1.32E−03	6.76E−05	1.12E−06	1.01E−03
Ame-μ10	238	103	0.6980	3.03E−08	1.34E−03	9.20E−05	1.35E−06	9.65E−04
Ame-μ13	208	125	0.6246	2.32E−09	1.34E−03	1.10E−04	1.68E−06	8.80E−04
Ame-μ15	175	175	0.5000	1.50E−07	1.22E−03	3.48E−04	2.24E−06	7.85E−04
Ame-μ19	192	144	0.5714	4.09E−08	1.36E−03	4.70E−05	1.92E−06	7.98E−04
Ame-μ22	161	181	0.4708	5.16E−08	1.32E−03	8.34E−05	2.37E−06	6.66E−04
Ame-μ24	160	31	0.8377	5.41E−10	1.32E−03	4.38E−04	7.26E−07	1.18E−03
Ame-μ26	211	111	0.6553	1.45E−08	1.32E−03	0.00E+00	1.54E−06	8.64E−04
Ame-μ27	230	85	0.7302	2.95E−10	1.33E−03	0.00E+00	1.21E−06	9.69E−04
Average			0.6486	3.31E−08	1.32E−03	1.32E−04	1.57E−06	9.01E−04

Base on the allelic dropout rate μ_i and false allele rate v_i inferred from analyzing the data, the probability of genotyping error for heterozygote $\tilde{\phi}_i$, of homozygote $\tilde{\vartheta}_i$ and mean probability of genotyping error P_i are given. Science Counting: 1E−05 denotes 10^{-5}

genotyping errors from allele dropout, which corroborates the original intentions of Taberlet et al. (1996).

Case analysis

We used non-invasive data from the giant panda (*Ailuropoda melanoleuca*, Zhan et al. 2006) to examine the above equations and estimate genotyping errors for those studies using the standard multi-tubes method. In the above study, we investigated the abundance of giant panda in Wanglang Nature Reserve, China (Table 2). With nine loci, we adopted a modified multi-tubes approach (Taberlet et al. 1996) to determine the genotypes. Loci that gave rise to the same heterozygous genotype twice were accepted as heterozygotes. Otherwise, a third repeat was conducted and heterozygotes were accepted if both alleles appeared twice or more among three repeats. Then four more PCR repeats were carried out for all homozygotes and appropriate heterozygotes, using the same criteria as those designed by Taberlet et al. (1996).

In the study, nine microsatellite loci successfully amplified 375 samples collected in Wanglang and its neighboring areas. Heterozygotes were determined after two, three or seven PCR repeats and all homozygotes were repeated seven times. The protocol here is $\Lambda_{2,6}$. As stated before, the technical bottleneck for the protocol is to estimate the allelic dropout rate (μ_i) and false allele rate (v_i), which is the reason why many similar studies adopting the multi-tubes method, without a reference genotype, could not provide actual error rates. However, for v_i, we can in principle use the observed error rates calculated from homozygotes with seven PCR repeats, assuming the rate represents both kinds of zygotes. As for μ_i, we can consider (1) its lowest limit ($\mu_i^{\min}$) as 0 for heterozygotes with two PCR replicates, 1/6 for those with three replicates, and the actual values for those with seven; (2) its upper limit ($\mu_i^{\max}$) as 0.25 for those with two or three repeats and actual values for those with seven repeats. The reasons for these settings for $\mu_i^{\max}$ are as below.

Firstly, Broquet et al. (2007) reviewed and calculated the error rates for the non-invasive genetic data published up to 2004 and found that the error rates (measured as allelic dropout) for all loci used but four did not exceed 50%. However, even error rate estimations of those four loci could have been biased because these loci had low amplification success rates (<50%, Smith et al. 2000; Constable et al. 2001) and some proved to be significantly affected by null alleles (Constable et al. 2001); secondly, according to the computer simulations of Taberlet et al. (1996), the maximum error rate per allele should be 0.25, so 0.25 sets a conservative standard for our study. For example, if μ_i is 0.25, the chance of obtaining the same two heterozygotes in two PCR replicates should be 0.32, but it is nearly 0.5 in our final dataset, suggesting that the ture $\mu_i^{\max}$ in our case could be less than 0.25.

Using the above method, in our study of the giant panda (Zhan et al. 2006), the probability of genotyping error for heterozygotes was 3.31E−08 to 1.32E−03 and 1.32E−04 for homozygotes. The mean genotyping error rate per locus was 1.57E−06 to 9.01E−04. Based on these estimates, we expect that there could have been no more than four erroneous genotypes in our dataset.

Our equations, for the first time, provide a new way to estimate intrinsic error rates for genetic data using the multi-tubes method, most widely used one in the noninvasive studies. In our study of the giant panda, the multi-tubes method seems to have effectively minimized genotyping errors, refuting recent criticisms of our study by Garshelis et al. (2008). Our equations also can be conveniently used for calculating error rates from pilot studies,

under which the number of repeats for heterozygotes (n) should be equal to that for homozygotes (m). Although the greater n and m are, the better the results seem to be, how many repeats needs to be carried out to guarantee noninvasive genetic research is necessarily a trade-off between cost and the scientific question being asked. For population biologists who want to use noninvasive genetics samples, we suggest in practice either to follow the multi-tubes method initially and quantify the error rates using our method, or to implement a pilot study to estimate the allelic dropout rate/false allele rate and then use our method to calculate the genotyping error rates in the final dataset and compare them with the pilot data to assess their accuracy.

Acknowledgments The study was supported by National Natural Science Foundation of China (30770399; 30620130432), the Royal Society and the Knowledge Innovation Program of Chinese Academy of Sciences.

References

Broquet T, Petit E (2004) Quantifying genotyping errors in noninvasive population genetics. Mol Ecol 13:3601–3608

Broquet T, Ménard N, Petit E (2007) Noninvasive population genetics: a review of sample source, diet, fragment length and microsatellite motif effects on amplification success and genotyping error rates. Conserv Genet 8:249–260

Constable JL, Ashley MV, Goodall J, Pusey AE (2001) Noninvasive paternity assignment in Gombe chimpanzees. Mol Ecol 10:1279–1300

Creel S, Spong G, Sands JL, Rotella J, Zeigle J, Joe L, Murphy KM, Smith D (2003) Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. Mol Ecol 12:2003–2009

Flagstad Ø, Hedmark E, Landa A, Brøseth H, Persson J, Andersen R, Segerström P, Ellegren H (2004) Colonization history and noninvasive monitoring of a reestablished wolverine population. Conserv Biol 18:676–688

Garshelis DL, Wang H, Wang DJ, Zhu XJ, Li S, McShea WJ (2008) Do revised giant panda population estimates aid in their conservation. Ursus 19:168–176

Johnson PCD, Haydon DT (2006) Maximum-likelihood estimation of allelic dropout and false allele error rates from microsatellite genotypes in the absence of reference data. Genetics 175:827–842

Mckelvey KS, Schwartz MK (2004) Genetic errors associated with population estimation using non-invasive molecular tagging: problems and new solutions. J Wildl Manage 68:439–448

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. Nat Rev Genet 6:847–859

Prugh LR, Ritland CE, Arthur SM, Krebs CJ (2005) Monitoring coyote population dynamics by genotyping faeces. Mol Ecol 14:1585–1596

Smith KL, Alberts SC, Bayes MK, Bruford MW, Altmann J, Ober C (2000) Cross-species amplification, non-invasive genotyping, and non-mendelian inheritance of human STRPs in savannah baboons. Am J Primatol 51:219–227

Solberg KH, Bellemain E, Drageset OM, Taberlet P, Swenson JE (2006) An evaluation of field and non-invasive genetic methods to estimate brown bear (*Ursus arctos*) population size. Biol Conserv 128:158–168

Taberlet P, Griffin S, Goossens B, Questiau S, Manceau V, Escaravage N, Waits LP, Bouvet J (1996) Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Res 24:3189–3194

Taberlet P, Camarra JJ, Griffin S, Uhrès E, Hanotte O, Waits LP, Dubois-Paganon C, Burke T, Bouvet J (1997) Noninvasive genetic tracking of the endangered Pyrenean brown bear population. Mol Ecol 6:869–876

Valière N (2002) GIMLET: a computer program for analyzing genetic individual identification data. Mol Ecol Notes 2:377–379

Zhan XJ, Li M, Zhang ZJ, Goossens B, Wang HJ, Chen YP, Bruford MW, Wei FW (2006) Molecular censusing doubles giant panda population estimate in a key nature reserve. Curr Biol 16:451–452

Zhan XJ, Zhang ZJ, Wu H, Goossens B, Li M, Jiang SW, Bruford MW, Wei FW (2007) Molecular analysis of dispersal in giant pandas. Mol Ecol 16:3792–3800