Individual identification from genetic marker data: developments and accuracy comparisons of methods

J. Wang

Institute of Zoology, Zoological Society of London, London NW1 4RY, United Kingdom

Left running head: J. Wang

Right running head: Marker-based individual identification

Key words: Relationship, Relatedness, Genetic Markers, Clone, Duplicates

Corresponding author

Jinliang Wang

Institute of Zoology

Regent's Park

London NW1 4RY

United Kingdom

Tel: 0044 20 74496620

Fax: 0044 20 75862870

Email: jinliang.wang@ioz.ac.uk

Abstract

Genetic marker based identification of distinct individuals and recognition of duplicated individuals has important applications in many research areas in ecology, evolutionary biology, conservation biology and forensics. The widely applied genotype mismatch (MM) method, however, is inaccurate because it relies on a fixed and suboptimal threshold number (T_M) of mismatches, and often yields self-inconsistent pairwise inferences. In this paper I improved MM method by calculating an optimal T_M to accommodate the number, mistyping rates, missing data and allele frequencies of the markers. I also developed a pairwise likelihood relationship (LR) method and a likelihood clustering (LC) method for individual identification, using poor-quality data that may have high and variable rates of allelic dropouts and false alleles at genotyped loci. The 3 methods together with the relatedness (RL) method were then compared in accuracy by analysing an empirical frog dataset and many simulated datasets generated under different parameter combinations. The analysis results showed that LC is generally one or two orders more accurate for individual identification than the other methods. Its accuracy is especially superior when the sampled multilocus genotypes have poor quality (i.e. teemed with genotyping errors and missing data) and highly replicated, a situation typical of noninvasive sampling used in estimating population size. Importantly, LC is the only method that guarantees to produce self-consistent results by partitioning the entire set of multilocus genotypes into distinct clusters, each cluster containing one or more genotypes that all represent the same individual. The LC and LR methods were implemented in a computer program COLONY for free download from the internet.

Introduction

Identification of distinct individuals and recognition of duplicated individuals from genetic marker data is important in many research areas in ecology, evolutionary biology, conservation biology and forensics. It has been used to estimate population size (or species abundance) in the traditional capture-mark-recapture (CMR) framework (Palsbøll *et al*. 1997; Schwartz *et al*. 1998; Creel *et al*. 2003; Luikart *et al*. 2010), to track individuals across different life cycle stages in studying population parameters such as survivorship (Ringler *et al.* 2015) and migration, to infer colonal reproduction rates (Escaravage *et al*. 1998; Halkett *et al*. 2005), and to trace illegally killed animals or illegal trading animal products in wildlife forensics (Alacs *et al*. 2010). It can and should also be routinely used as a data cleaning tool to remove accidentally duplicated individuals before conducting various analyses of the raw genotype data. This is because, similar to close relatives but to a greater extent, duplicated individuals inadvertently included in a genetic analysis can reduce the estimates of genetic diversity, bias the estimates of fixation indices (F_{IS}, F_{IT} and F_{ST}), induce deviations from Hardy-Weinberg and linkage equilibrium, and ruin a population structuring inference (Anderson & Dunham 2008; Rodríguez-Ramilo & Wang 2012).

When marker information is ample (i.e. many polymorphic loci) and completely reliable (i.e. no mutations and no genotyping errors), individual identification is straightforward. In this ideal situation, identical multilocus genotypes (MGs) represent duplicated individuals and non-identical MGs correspond to distinct individuals. Nowadays with the wide application of highly polymorphic markers such as microsatellites and many genomic markers of SNPs, information content is no longer considered a constraint in practice. However, data quality could be a serious problem, especially in the case of noninvasive DNA samples such as hair, feathers and scats (Taberlet *et al*. 1999; Pompanon *et al*. 2005). Due to the limited quantity and quality of DNA extracted from noninvasive samples, the presence of PCR inhibitors and DNA contaminations, noninvasive genotype data are characterized by high rates of missing data, false alleles and allelic dropouts (Bonin *et al*. 2004). Indeed, genotyping errors are a rule rather than an exception. Even genotypes obtained from DNA of high quality and quantity (e.g. extracted from fresh tissue or blood samples) are not exempt from mistypings (Pompanon *et al*. 2005). The more markers are genotyped, the higher the probability that an MG contains genotyping errors.

Unfortunately, individual identification is particularly vulnerable to genotyping errors in comparison with other genetic data analyses such as population genetic diversity or structure, because just one single error in an MG could create a false (ghost) individual. Even if genotyping errors occur at a very low rate *e* per locus, the probability that an MG contains one or more errors,

$E = 1 - (1 - e)^L$, can be high, and increases rapidly with the number of loci L. For example, a 10-, 50- and 250-locus genotype is expected to contain at least one mistyping with a probability of 1.0%, 4.9% and 22.1% respectively when e=0.001, of 9.6%, 39.5% and 91.9% respectively when e=0.01, and of 40.1%, 92.3% and 100% respectively when e=0.05. This result has prompted several researchers to suggest that individual identification should use the minimum number of loci required to attain a low probability of identity among samples from different individuals (Waits *et al*. 2001; Creel *et al*. 2003). This suggestion can reduce ghost individuals due to genotyping errors, but unfortunately it could also seriously limit the power of individual identification, especially in the difficult situation where many close relatives such as full siblings are present (Waits *et al*. 2001).

The problems of and difficulties in individual identification due to genotyping errors are made more prominent by high sample replications where many replicated samples could be collected from a single individual. Scat or hair based non-invasive samples (e.g. Creel *et al*. 2003) often exhibit massive replications with potentially tens to hundreds of replicated samples per individual. At this high level of replications, even a very small genotyping error rate could result in extreme overestimates of distinct individuals and of population size (Waits & Leberg 2000; Creel *et al*. 2003; McKelvey & Schwartz 2004). High sample replications coupled with genotyping errors and missing data can also result in numerous conflicts in pairwise inferences by any method (including the mismatch method) that compares pairs of samples (multilocus genotypes). For example, sample A may be inferred to be a duplicate of both sample B and sample C, but B and C may be inferred to come from distinct individuals.

A more robust and error-tolerant approach is to accept the presence of genotyping errors and accommodate them in recognizing individuals from MG data by the mismatch (MM) method. A common practice is that two samples having identical genotypes at all but 1 or 2 loci are accepted as being from a single individual and the mismatches are regarded as genotyping errors. This approach has been implemented in several computer programs, such as GENECAP (Wilberg & Dreher 2004). The allowance of a small threshold number, T_m, of 1 or 2 mismatches could reduce ghost individuals substantially. However, this threshold is obviously arbitrary, the optimum being dependent on factors such as the mistyping rates and the number of loci. While 1 or 2 mismatches may be sufficient to reduce ghost individuals when both mistyping rates and number of loci are low (say, e<0.05 and L<20), more mismatches should be allowed for when e or/and L are high. To overcome the problem, Galpern *et al*. (2012) proposed to determine T_m as the value where the number of individuals with more than one MG in a sample has a second minimum. Although their T_m no longer relies on a predefined value, it depends on a similarity index defined to penalize arbitrarily missing and mismatched genotypes at a locus by $1/(2L)$ and $1/L$ respectively.

Furthermore, analyses of simulated (Galpern *et al*. 2012, Table 3) and empirical data (Ringler *et al*. 2015) showed that this flexible T_m approach has a similar accuracy to the approach with a fixed T_m =2.

A more powerful approach to individual identification is via pairwise relatedness analysis. Relatedness analysis is resilient to genotyping errors (Wang 2007), and can use allele frequency as well as genotype information in identifying duplicated individuals from other competitive relationships such as full siblings (Ringler *et al*. 2015). In diploid species, two MGs are expected to have a relatedness, *r*, of 1 and 0.5 if they come from the same individual and from two first-class relatives (full sibs and parent-offspring), respectively. Therefore, MGs are inferred to represent duplicates of the same individual when their estimated relatedness is closer to 1 than to 0.5 (i.e. when their estimated relatedness is above an appropriate threshold *r* value, say T_r =0.75). Otherwise, they are inferred to represent distinct individuals.

In this study, I will improve the mismatch method by calculating and using an optimal T_m that takes into account mistyping rates, missing data, and the number and allele frequencies of markers. I also propose two new likelihood approaches to efficient individual identification from genotype data of low quality. One is based on calculating the likelihood values of two MGs for their candidate relationships of clone mates (duplicates) and close competitive relationships (full siblings and parent offspring), and the other is based on partitioning (in a likelihood framework) the entire set of MGs into clusters with each cluster containing one or more genotypes that all represent the same individual. Both approaches accommodate genotyping errors and use allele frequency information, and the likelihood clustering method abandons the pairwise approach such that the inferences are guaranteed to be consistent and are especially accurate for the difficult situation of high sample replications. The accuracy of these approaches is evaluated and compared by analysing many simulated and an empirical dataset.

Methods

Dyadic mismatch method (MM)

The threshold value of mismatches, T_m, is critical for the mismatch method. The number of distinct individuals will be overestimated and underestimated when T_m is too small and too large, respectively. The optimal T_m that minimizes falsely detected (α-error) and undetected (β-error) individuals depends on the rate of genotyping errors, the number of loci, the allele frequencies of each locus, and the actual genetic structure (i.e. the actual relationships) of the focal set of MGs.

The latter is unknown and is the target of the analysis, but the former three pieces of information are usually available and can be used to resolve an approximately optimal T_m.

Suppose locus l has K_l alleles with estimated frequencies p_{li}, where i=1, 2, …, K_l and l=1, 2, …, L. I assume that a genotype G_l at locus l may be mistyped to be a phenotype g_l due to allelic dropouts (ADO) at rate ε_{l1} and false alleles (FA) at rate ε_{l2}. ADOs and FAs are the most common genotyping errors in microsatellites (Bonin *et al*. 2004; Pompanon *et al*. 2005). For ADO, I assume each of the two gene copies in a diploid genotype has the same probability of dropping out during PCR, and double dropouts (i.e. both gene copies dropping out to produce no PCR products) are rare and negligible. Under this model, ADO affects heterozygote genotypes only, and a heterozygote G_l =$\{w,x\}$ ($w \neq x$) is observed to be a phenotype g_l =$\{w,x\}$, $\{w,w\}$ and $\{x,x\}$ with probabilities $1-2e_{l1}$, e_{l1} and e_{l1} respectively, where $e_{l1} = \varepsilon_{l1}/(1+\varepsilon_{l1})$. For FA, I assume that any allele in any genotype is independently and equally probable to be mistyped to be any one of the other alleles, at a rate e_{l2} = $\varepsilon_{l2}/(K_l-1)$.

Given allele frequencies (p_{li}) and mistyping rates (ε_{l1} and ε_{l2}), and assuming genotype frequencies at Hardy-Weinberg equilibrium (HWE), I can apply the above ADO and FA models to each of the K_l (K_l + 1)/2 genotypes twice to generate two phenotypes, and derive the probability that two phenotypes from the same genotype match, Q_l. The expression for Q_l is however a very complicated function of p_{li}, ε_{l1} and ε_{l2}, and is not enlightening. For simplicity, Q_l is determined by simulations. First, a genotype is drawn at random from a population in HWE with allele frequency p_{li}. Second, a phenotype is generated from the genotype, following the ADO model. Third, the phenotype is further modified according to the FA model. Fourth, steps 2 and 3 are repeated to generate another phenotype independently from the same genotype. Fifth, the two phenotypes are compared to determine whether they match or not. Steps 1-5 are repeated for a sufficiently large number of replicates, and the frequency of matching phenotypes gives a good estimate of Q_l.

The average number of mismatches between two phenotypes having the same underlying genotype at a set of L loci is calculated by $\sum_{l=1}^{L}(1-Q_l)$ rounded to the nearest integer. This optimal T_m value is expected to minimise both α and β errors in individual identification by the MM method. Note that T_m is calculated for each pair of MGs in a sample such that missing data can be easily accommodated. If any or both MGs have missing data at locus l, then Q_l is set to 1 for the locus in the calculation. Therefore the calculated T_m values are dyad specific and lower for dyads with more missing data. In contrast to the widely applied fixed T_m =2, this T_m value calculated from Q_l accounts for allele frequencies, mistyping rates, number of loci, and missing data. Two MGs are

inferred to be from a single and two distinct individuals when their observed number of mismatches is not and is greater than their T_m value, respectively.

Dyadic relatedness method (RL)

The genetic relatedness, *r*, between two MGs can be calculated by a marker-based moment or likelihood estimator (Wang 2007; 2014). Duplicated individuals and first-order relatives (e.g. full-sib or parent-offspring) are expected to have an *r* value of 1 or close to 1 and of 0.5 or close to 0.5, respectively, even when they have mismatches at a small fraction of loci due to genotyping errors (Wang 2007). To distinguish duplicated individuals from first-order relatives and to minimise both α- and β-error rates, I choose a threshold *r* value of T_r=0.75, which is the midpoint between the expected *r* values for duplicates and first-order relatives. Two MGs are inferred to be duplicates and distinct individuals when their *r* value is and is not greater than T_r, respectively. There are quite a few *r* estimators available (Wang 2014), among which I chose to use the one based on phenotype similarity, proposed by Lynch (1988) and improved by Li *et al*. (1993). It is chosen because it is simple to calculate and is expected to have a higher accuracy than other moment estimators when applied to close relationships such as identical twins (duplicates) and full sibs (Wang 2007).

Dyadic likelihood relationship method (LR)

It is also possible to calculate directly the likelihoods of two MGs for the candidate relationships of duplicates (clone mates or identical twins, denoted by DP), full sibs (FS), half sibs (HS), parent offspring (PO) and unrelated (UR). If DP has the highest likelihood, then the two MGs are inferred to come from the same individual. Otherwise, they are inferred to come from distinct individuals.

In contrast to pairwise relatedness estimation, relationship inference is highly vulnerable to genotyping errors. A single error could exclude truly duplicated MGs from being inferred as such. The more markers one uses, the more serious the false exclusion problem will become. The likelihood functions of FS, HS and PO are available in the literature, but they do not account for genotyping errors (e.g. Goodnight & Queller 1999) or account for ADO only (e.g. Wagner *et al*. 2006). Herein I show the general likelihood function applying to any pairwise relationship (including DP, FS, HS, PO and UR) and allowing for both ADO and FA occurring at rates variable across loci.

The genetic relationship between two non-inbred individuals is fully specified by 3 identical by descent (IBD) coefficients Δ_i, where Δ_i is the probability that the two individuals share exactly *i* (*i*=0, 1, 2) pairs of gene copies IBD at a locus. Obviously, $\Delta_0 + \Delta_1 + \Delta_2 \equiv 1$. In diploid species, Δ_0, Δ_1 and Δ_2 have values 0, 0 and 1 for DP, 0.25, 0.5 and 0.25 for FS, 0, 0.5, 0.5 for HS, 0, 1 and 0 for PO,

and 0, 0, 1 for UR. The probability of observing a phenotype $g_A=\{a,b\}$ for individual A and a phenotype $g_B=\{c,d\}$ for individual B at a locus with K codominant alleles, given their relationship defined by Δ_0, Δ_1 and Δ_2, is (Wang 2006)

$$\Pr[a,b;c,d|\Delta_0,\Delta_1,\Delta_2]=\sum_{u=1}^{K}\sum_{v=u}^{K}\sum_{w=1}^{K}\sum_{x=w}^{K}\mathrm{R}[u,v;w,x|\Delta_0,\Delta_1,\Delta_2]\Pr[a,b|u,v]\Pr[c,d|w,x], \quad (1)$$

where

$$\begin{aligned}\mathrm{R}[u,v;w,x|\Delta_0,\Delta_1,\Delta_2] &= (2-\delta_{uv})p_u p_v\Big(\Delta_0(2-\delta_{wx})p_w p_x+\tfrac{1}{4}\Delta_1(2-\delta_{wx})\big((\delta_{uw}+\delta_{vw})p_x\big)\\ &+(\delta_{ux}+\delta_{vx})p_w\Big)+\Delta_2(\delta_{uw}\delta_{vx}+\delta_{ux}\delta_{vw}-\delta_{uw}\delta_{vx}\delta_{ux}\delta_{vw}))\end{aligned} \quad (2)$$

is the probability that A and B have genotype $\{u,v\}$ and $\{w,x\}$ respectively conditional on their relationship or IBD coefficients $\Delta_0,\Delta_1,\Delta_2$, and δ_{uv} (and similarly for other δ variables) is the Kronecker delta variable with values 1 and 0 when $u=v$ and $u\neq v$, respectively. In (1), $\Pr[u,v|w,x]$ is the probability that a genotype $\{w,x\}$ shows a phenotype $\{u,v\}$ due to ADO and FA. It is derived as (Wang 2004)

$$\Pr[u,v|w,x]=\begin{cases}(1-\varepsilon_2)^2+e_2^2-2e_1e_3^2 & (u=w,v=x)\\ e_2(1-\varepsilon_2)+e_1e_3^2 & (u=v=w)\text{ or }(u=v=x)\\ (2-\delta_{u,v})e_2^2 & (u\neq w,u\neq x,v\neq w,v\neq x)\\ e_2e_3 & (\text{otherwise})\end{cases} \quad (3)$$

for a heterozygous genotype ($w\neq x$) where $e_3=1-\varepsilon_2-e_2$, and

$$\Pr[u,v|w,x]=\begin{cases}(1-\varepsilon_2)^2 & (u=v=w)\\ 2e_2(1-\varepsilon_2) & (u=w,v\neq w)\text{ or }(v=w,u\neq w)\\ (2-\delta_{u,v})e_2^2 & (u\neq w,v\neq w)\end{cases} \quad (4)$$

for a homozygous genotype ($w=x$) under the ADO and FA models described above.

Note that equations (1-4) give the likelihood of a relationship for a single locus *l*, and subscript *l* is dropped from error rates ($\varepsilon_{l1},\varepsilon_{l2},e_{l1},e_{l2}$) and allele frequencies (p_{li}) for clarity. The multilocus likelihood is simply a product of single locus likelihood values, assuming linkage equilibrium among loci.

Likelihood clustering method (LC)

The above 3 methods take a pairwise approach, which considers whether two MGs are duplicates or not in isolation of others. When an individual has more than 2 replicated MGs, pairwise approaches

may yield conflicting results. Among 3 replicated MGs A, B and C of an individual, for example, A and B as well as A and C may be inferred as DP while B and C may be inferred as distinct individuals. This happens when, for an example, A, B and C have genotypes identical at all but a single locus at which A has missing data while B and C show different alleles. The 3 pairwise inferences are obviously in conflict. The frequency of these inconsistencies increases rapidly with an increasing level of individual replications, and decreasing data information quality and quantity. Furthermore, pairwise approaches do not use marker information fully, and thus are expected to have a lower power (accuracy) than approaches that consider the relationship among all MGs simultaneously (Wang 2004).

A more desirable approach is to partition the entire set of MGs into N (unknown) individual clusters, with each cluster containing one or more MGs that all represent the same individual. To reduce both α and β errors, the clustering should be better made by considering several competitive relationships such as DP, FS and HS which could generate similar patterns of MGs. The algorithm used for sibship inference (Wang 2004) can be modified to identify individuals, as shown below.

First, assuming each MG corresponds to a distinct individual, a sibship analysis is conducted to partition the entire set of individuals into full-sib clusters. The analysis could adopt the simple monogamy model (i.e. no inference of half sibs), or the sophisticated polygamy model (i.e. inference of half sibs). The monogamy model is preferred because it runs much faster than, but has the same or very similar accuracy to, the polygamy model for individual identification. This is because DP is much closer to FS in relatedness than to HS and is thus much less likely to confuse with HS than FS. Second, each inferred FS cluster is further partitioned by a likelihood approach into a number of individual clusters, with each cluster containing one or more MGs that all represent the same individual. The first step has been described before (Wang 2004), and the second step is detailed below.

Suppose an inferred FS cluster contains M ($\geq$1) MGs. If M=1, then no further analysis is needed. Otherwise, the MGs can be divided into one of a number of B_M possible partitions (or configurations), where B_M is the Bell number. A partition contains a number of m (where $m\geq1$ and $m\leq M$) individual clusters, with each cluster containing one or more MGs that all represent the same individual. Three MGs (M=3) of A, B and C, for example, have B_3=5 different partitions, which are {(A), (B), (C)}, {(A, B), (C)}, {(A, C), (B)}, {(B, C), (A)}, {(A, B, C)} where all MGs in a pair of parentheses come from the same individual and constitute an individual cluster. Partition {(A, B), (C)}, for example, has two individual clusters which are (A, B) and (C), meaning that A and B come from one individual and C comes from another individual. Each partition is evaluated for its

likelihood which is equal to the probability of the genotype data given the partition, and the one with the maximum likelihood is returned as the best estimate. The challenge is to construct, and calculate the likelihood values of, the B_M partitions, where B_M increases explosively with M. Even for small M value of 5, 10 and 15, for example, the corresponding B_M values are 52, 115975 and 1382958545, respectively.

Instead of using the simulated annealing approach in sibship analysis (Wang 2004), I take a systematic approach to individual identification. The approach is deterministic and fast, because a FS cluster is usually small. For a FS cluster with M MGs, the algorithm starts with an initial configuration, C_0, of M individual clusters, each containing one MG. Round 1 searching works on C_0. Each of the $M(M-1)/2$ possible configurations is constructed by merging two of the M clusters, and is evaluated for likelihood. The best of these configurations, C_1, with the maximum likelihood value is then compared with C_0. If the former has a smaller likelihood, then C_0 is returned as the best estimate and the searching process terminates. Otherwise, C_0 is abandoned and C_1 is accepted, and round 2 searching is initiated to improve on C_1. Following exactly the same procedure in constructing new configurations as in round 1, round 2 returns the best configuration with $M-2$ clusters, C_2. If C_2 has a lower likelihood than C_1, then the latter is reported as the best estimate and the searching process terminates. Otherwise, C_1 is replaced by C_2, and round 3 searching is initiated to work on C_2, following the same process as in previous rounds. The whole searching process stops when, at round m, the best of the $(M-m+1)(M-m)/2$ reconfigurations, C_m, has a lower likelihood than that of the previous round, C_{m-1}, which is returned as the best estimate.

Now consider the likelihood of a configuration with m (=1~M) individual clusters, with cluster i (=1, 2, …, m) containing n_i genotypes g_{ij} (j=1, 2, …, n_i) at a locus with K alleles. All genotypes within a cluster are duplicates of the same individual, and genotypes from different clusters represent different individuals. Obviously, we have $\sum_{i=1}^{m} n_i \equiv M$. The likelihood function is

$$\sum_{u=1}^{K} p_u \sum_{v=1}^{K} p_v \sum_{w=1}^{K} p_w \sum_{x=1}^{K} p_x \prod_{i=1}^{m} \frac{1}{4} \left(\sum_{a=u,v} \sum_{b=w,x} \prod_{j=1}^{n_i} \Pr[g_{ij} | a, b] \right), \quad (5)$$

where the probability of observing a phenotype g_{ij} given its underlying genotype $G_{ij}=\{a,b\}$, $\Pr[g_{ij}|a,b]$, is calculated by (3-4). The computational cost of (5) can be much reduced by pooling all unobserved alleles in the FS cluster into a single "allele" and by pooling identical parental genotypes (e.g. $\{u,v\}$ and $\{v,u\}$) and parental genotype combinations (e.g. $\{\{u,v\}, \{w,x\}\}$ and $\{\{w,x\}, \{u,v\}\}$), as in sibship likelihood calculations (Wang 2004). For multiple loci in linkage

equilibrium, the likelihood is simply the product of single locus values calculated by (5). When one or both parents of the FS family are assigned to candidate adults with genotype data, the likelihood function is slightly more complicated and is not shown herein.

Simulations

Simulated data were generated and analysed comparatively by the above described 4 methods to evaluate their accuracies. A number of factors are expected to affect individual identifications, and are thus considered in the simulations.

First, the simulations considered the actual relatedness structures of the sampled individuals. Presence of close relatives, such as full sibs, makes individual identification more difficult by increasing β errors. I considered 3 sibship structures to reflect low, medium and high relatedness. These are denoted by 40(1, 1), 16(1, 1, 3) and 4(1, 2, 3, 4, 10), where the value before the brackets gives the number of replicate half-sib families and the values within the brackets are the sizes of full-sib families that are nested within a half-sib family. For example, 16(1, 1, 3) means there are 16 half-sib families, and each family has a single father mated with 3 mothers who give 1, 1, and 3 full siblings. Each of the 3 sibship structures yields 80 distinct individuals (genotypes) in a sample. Other close relatives such as parent-offspring may also be present in a practical sample. However, these relationships have much smaller effect on individual identification than full sibs, because the latter are more likely to generate identical or nearly identical MGs. Therefore, relatives other than full sibs are not considered in the simulations.

Second, the simulations allowed for different extents of individual replications. The number of individual genotype replications is assumed to follow a Poisson distribution with parameter λ, taking values between 0 and 5. For each of the 80 distinct individuals in a sample, a random number R~Poisson[λ] is generated and the MG of the individual is replicated by R times.

Third, the simulations considered different numbers and polymorphisms of markers. For given numbers of loci (L) and alleles (K_l) per locus, allele frequencies were drawn from a uniform distribution at each locus, and the 80 MGs in a given sibship structure were generated by assuming Hardy-Weinberg and linkage equilibrium. These MGs were faithfully replicated according to Poisson[λ] as described above. When considering the impact of K_l, I vary K_l and L simultaneously such that the total number of independent alleles across loci, $\sum_{l=1}^{L}(K_l - 1)$, is fixed at 160.

Fourth, the simulations allowed for different rates of ADO, FA and missing data at each locus. After replications, each MG is modified independently at each locus for ADOs, FAs, and missing data to generate the corresponding multilocus phenotype. Suppose ADO, FA and missing

data occur at rates ε_{l1}, ε_{l2} and ε_{l3} at locus *l*, respectively. A maximum of 3 steps are required to generate the phenotype at this locus from its genotype. In step 1, a random number *R* uniformly distributed in the range [0,1] is drawn. If $R \leq \varepsilon_{l3}$, then the phenotype becomes {0,0} to indicate missing data. Otherwise, the genotype is subject to ADO in step 2. Another random number *R* is drawn. If the genotype is a heterozygote and $R \leq \varepsilon_{l1}/(1+\varepsilon_{l1})$, then the phenotype is returned as a homozygote for an allele drawn at random from the genotype. Otherwise, the genotype has no changes in step 2. In both cases, the genotype is subject to FA in step 3. For each allele in the genotype, a random number *R* is drawn. If $R \leq \varepsilon_{l2}$, then the allele is changed to another allele randomly drawn from the *K*-1 alleles. Otherwise, no change is made to the allele.

Fifth, all methods except for RL use ADO and FA rates at each locus. In practice, these mistyping rates are usually unknown, but are estimated from duplicated genotyping or pedigree based analysis (Creel *et al*. 2003; Pompanon *et al*. 2005). It is important to know how robust these methods are to mis-specified mistyping rates. For this purpose, I simulated data with a true mistyping rate of $\varepsilon_{l1}=\varepsilon_{l2}=0.1$ for each locus *l*, but analysed the data assuming values of $\varepsilon_{l1}=\varepsilon_{l2}$ in the range of 0 to 0.2.

Accuracy assessment

Accuracy is assessed by the proportion of MG dyads in a dataset that are from a single individual but are incorrectly identified as from distinct individuals (α errors, falsely identified individuals), and that are from distinct individuals but are incorrectly identified as from a single individual (β errors, unidentified individuals). The overall accuracy including both types of errors is measured by the proportion of MG dyads in a dataset that are incorrectly inferred to be non-duplicates or duplicates, γ. These α-, β- and total-error rates were calculated for each dataset and averaged across 100 replicate datasets for a given parameter combination. Because most applications are affected by both α- and β-errors, I report the total error rate, γ, to indicate overall accuracy in this paper to save space.

Empirical data

Ringler *et al*. (2015) showed that microsatellites can be used to reliably mark amphibian larvae and to re-identify them after metamorphosis. They genotyped 1800 tadpoles of the dendrobatid frog (*Allobates femoralis*) at 14 highly polymorphic microsatellite loci before releasing them on a 5-ha river island which was previously uninhabited by this species. They surveyed the island and sampled 42 juvenile individuals six months after the release, and sampled 36 males and 31 females one year after the release. The sampled juveniles and adults were released to their capture sites after

taking DNA samples, which were genotyped at the same set of 14 loci as the tadpoles. Based on their unique ventral patterns, 20 of the 67 adults were identified to correspond to one of the 42 juveniles. These 20 individuals sampled as both juveniles and adults were mostly confirmed by relatedness analysis of marker data. Individual identification between tadpoles and juveniles or between tadpoles and adults was based on the mismatch and relatedness methods. In the present study, the genotype data are comparatively analysed by the 4 individual identification methods.

Results

Simulations under the three sibship structures yield qualitatively similar results, and thus only the analysis results for sibship structure 4(1, 2, 3, 4, 10) are reported below.

Effect of the number of markers

The optimal T_m determined by the simulation procedure gives an unbiased estimate of the average number of mismatches between duplicated MGs for different numbers of loci L (Fig. 1) and for different mistyping and data missing rates (not shown). For a given L, calculated T_m values vary because different MG dyads may have different numbers of loci at which genotype data are missing, and because different loci may have different Q_l values. However, the variation of T_m values is much smaller than the variation of the observed numbers of mismatches, and the difference increases with L. Part of the reason that the mismatch method is less accurate than other methods (see below) is the high variation of the observed number of mismatches around T_m, which results in high rates of both α- and β-errors.

With an increasing number of markers, the accuracy of mismatch (MM) method is almost constant, while that of relatedness (RL), likelihood relationship (LR), and likelihood clustering (LC) methods increases rapidly (Fig. 1). This means RL, LR and LC are statistically consistent, but MM is not, even when an optimal T_m value was calculated and used in the analysis. MM makes decreasing β-errors (undetected individuals) but increasing α-errors (falsely detected individuals) with an increasing L, as expected. As a result, the overall error rate γ is almost constant with an increasing L (Fig. 1). If a fixed value of T_m =2 were used, MM method would perform much worse with a much higher γ due to excessive β-errors when $L < 10$ or excessive α-errors when $L > 10$.

LC is the most accurate method for different numbers of markers, followed by LR. These two methods become more and more accurate than RL method with an increasing number of loci. When L=80, perfect inference ($\alpha = \beta = 0$) is obtained by both LC and LR methods.

Effect of the number of alleles

For different numbers of alleles per locus and thus different numbers of loci when the total number of independent alleles is fixed at 160, LC method always has the lowest α error rate and the second lowest β error rate (Fig. 2). LR has an α error rate only slightly larger than LC, but has the highest β error rate. MM has an α error rate much larger and a β error rate much smaller than the other three methods. Overall, LC is the most accurate, making much fewer α and β errors than the other methods.

At a fixed total number of 160 independent alleles, the overall accuracy of the 4 methods first increases and then decreases with an increasing number of alleles per locus, *K* (Fig. 2). The maximal accuracy is achieved when *K*=5 for all methods except for the mismatch (MM) method. The RL and LR methods have an indistinguishable overall accuracy, which is higher than that of MM but much lower than that of LC for different numbers of alleles per locus. The accuracy differences among methods increases with a decreasing number of alleles per locus and correspondingly an increasing number of loci.

Effect of the extent of individual replication

Contrasting behaviours of different methods are observed for different levels of individual replications, λ (Fig. 3). With an increasing λ, the accuracy of LR is almost constant, that of MM and RL decreases, while that of LC increases. When a sample contains no replicated individuals (i.e. λ=0), MM has the lowest overall error rate γ because it has no chance to falsely identify individuals (α errors) to which the method is particularly vulnerable. However, MM quickly becomes the least accurate method at a low value of λ=0.3, when roughly each of 30% individuals is replicated only once. The clustering method LC always outperforms the 3 pairwise approaches when there exist replicated individuals in a sample, and this advantage increases steadily with the replication level λ.

Effect of mistyping and missing data rates

Genotyping errors and missing data decrease marker information and increase noises. As a result, all 4 methods show a decreasing accuracy with an increasing mistyping and missing data rate (Fig. 4). The mismatch method is especially susceptible to mistyping and missing data. Its accuracy quickly reduces to the lowest when $\varepsilon_{l1}=\varepsilon_{l2}=\varepsilon_{l3}$ raises to a low value of 0.01 for each of 20 loci. For the entire range of mistyping and missing data rates from 0 to 0.16, LC has the highest accuracy, followed by LR.

Robustness to mis-specified mistyping rates

The relatedness method does not use (account for) mistyping rates and thus its accuracy is unaffected by the assumed mistyping rate $\hat{\varepsilon}$ (Fig. 5). The behaviour of MM is perplexing, as its accuracy increases slowly with an increasing $\hat{\varepsilon}$ when it is actually larger than the true simulated mistyping rate ε. This is because the dominating errors made by MM when marker information is not small are falsely identified individuals (α errors), which can be reduced by the use of an overestimated mistyping rate. The two likelihood methods, LR and LC, have the highest accuracy when $\hat{\varepsilon}$ is roughly equal to ε. Their accuracy decreases as $\hat{\varepsilon}$ deviates from ε. Relatively, LR is much more vulnerable than LC to mis-specified mistyping rates, and becomes the least accurate method when roughly $\hat{\varepsilon} > 1.25\varepsilon$. Although LC is also affected by mis-specified $\hat{\varepsilon}$, it is always the most accurate method in the range between $\hat{\varepsilon}$=0 and $\hat{\varepsilon} = 2\varepsilon$.

Results of empirical data analysis

The 1909 MGs (1800 tadpoles, 42 juveniles, 67 adults) were partitioned by LC into 1766 individual clusters, each corresponding to an inferred distinct individual. Among these clusters, 1651, 92 and 23 are singletons, dyads, and trios, each containing 1, 2 and 3 MGs, respectively. Among the 23 trios, each of 20 contains a morphologically identified juvenile-adult dyad and a tadpole, one contains 2 tadpoles and a juvenile, one contains 2 tadpoles and an adult, and one contains 3 tadpoles. The first 20 trios confirm morphological observations and are highly likely to be correct, while the last 3 trios are probably incorrect if no tadpoles are actually replicated in the sample. The last 3 trios have similar numbers of missing and mismatched genotypes to the first 20 trios.

Because juveniles and adults are subsamples of tadpoles, we expect each juvenile or adult should have a corresponding tadpole. Indeed, each of all 67 adults and each of 38 juveniles was inferred to match a tadpole, and each of the 4 remaining juveniles was inferred to match no tadpoles. This means the α error (falsely identified individuals) rate of LC for this dataset is low, only about 3.6% (4 out of 109). It is also possible to calculate β error (unidentified individuals) rate of LC for this dataset, if no individuals within a life stage (tadpoles, juveniles, adults) are actually replicated. Among the 1821186 possible dyads, only 41 dyads within a life stage were identified by LC as single individuals, giving a β error rate of 0.0000225. It turns out that all of the 41 dyads are tadpoles, and no adults and no juveniles were found duplicated. This is not surprising because tadpoles are much more numerous than juveniles and adults, and many tadpoles were inferred to come from large full sib families (data not shown).

The distributions of the numbers of loci with missing data and mismatches between a pair of MGs for various classes of dyads are shown in Fig. 6, and explain the low power and accuracy of the mismatch method. As expected, there is essentially no difference in missing data for dyads of

various relationship classes. The average number of loci with missing data for a dyad is 2, no matter the dyadic MGs come from a single individual, two full siblings, or two non-full siblings. However, the distributions of mismatches differ among dyads of different classes. A dyad coming from a single individual most often has 0, 1 or 2 mismatches, but can occasionally have a maximal number of 7 mismatches. A full sib dyad on average has 8 mismatches, but can have a minimal number of only 2 mismatches. A non-full sib dyad on average has 11 mismatches, with the minimal number of mismatches being 6. Using a threshold value of mismatches T_m =6 or 7, the mismatch method can confidently identify duplicated MGs (Fig. 6, E and F) and unrelated individuals (Fig. 6H) with a small α and β error rates. However, it has tremendous difficulties to distinguish duplicated MGs from full siblings (Fig 6G). Using the optimum T_m value of 4 or 5, it still could result in substantial α and β error rates. The analysis shown in Fig. 6 also demonstrates that the optimal T_m value is not only marker property (e.g. number, polymorphisms, genotyping error rates, data missing rates) but also sample genetic structure (i.e. distributions of relatedness among MGs) dependent. The optimal T_m value would be 6~7 and 4~5 if full siblings occur at a very low rate and at a substantial rate, respectively. It should decrease with an increasing rate of full siblings and also a decreasing rate of duplicates to minimize both α and β errors. Unfortunately, however, sample genetic structure is usually unknown, and is the focus of an individual identification study.

Results from pairwise approaches are much less accurate, as expected from the simulation results and from the fact that this dataset has a large number of individuals and contains very large full sib families. Take the LR method as an example. Among the 1821186 possible dyads, 153 dyads within a life stage were identified as single individuals, yielding a β error rate 3.73 times larger than that of LC. A serious problem with the pairwise approach is self-conflicted inferences. Fig. 7 shows the pairwise relationships among 5 MGs inferred by LR. Obviously, these pairwise inferences are incompatible. The higher the level of individual replications, the more severe will be the problem of pairwise approaches.

Discussion

Although the mismatch method is the simplest and the most widely applied method for marker-based individual identification in molecular ecology, it has unfortunately several weaknesses and as a result is the least accurate method. First, the fixed threshold, typically T_m = 1 or 2, is arbitrary. It is too small when the number of loci or/and the mistyping rate is high, resulting in too many ghost individuals. It is too large when the number of loci and mistyping rate are very low, or/and close relatives are frequent. It is also too rigid and inappropriate for pairs of MGs having missing data at

different numbers of loci. These properties of MM have been well recognized, and have led to the suggestion that the fewest possible number of markers that have sufficient power for individual identification should be used to avoid excessive mismatches and exclusions (Waits *et al*. 2001; Creel *et al*. 2003). In reality, the markers used in individual identification can be highly variable in polymorphisms and mistyping rates, and the background relationship (e.g. sibship and parentage) structure of a sample can also be highly variable. It is difficult for any fixed value of T_m to cater for all scenarios. Second, the mismatch method fails to use the mismatch information efficiently. Two single locus genotypes are regarded matched when they are identical, and mismatched when they have either one or both alleles different. Obviously, mismatched genotypes give more evidence of distinct individuals when they have both alleles rather than a single allele different. This kind of information is however unused by the mismatch method. Third, the mismatch method treats all loci equally, while they can be highly heterogeneous in information (polymorphism) and noise (mistyping) contents. The method simply counts the number of mismatches, regardless of the loci at which the mismatches occur. Obviously, mismatched MGs give more support for distinct individuals when the mismatches occur at loci with lower mistyping rates or/and higher polymorphisms.

I showed in this study that an optimal T_m value can be calculated by simulations, accommodating the number of loci, the mistyping and missing data rates and the allele frequencies at each locus. The optimal T_m gives an unbiased estimate of the average number of mismatches between truly duplicated MGs (Fig. 1). Applying the optimal T_m value determined by simulations, the mismatch method has almost a constant accuracy independent of the number of loci (L, Fig. 1). If the fixed T_m=2 were applied, the accuracy would have decreased rapidly with L when it is larger than 20 because of the excessive α errors. Compared with other methods, however, the mismatch method using the optimal T_m value is still the least accurate for various parameter combinations considered in the simulations (Figures 1-5). It is impossible for the mismatch method to use as much marker information (e.g. mistyping rates, allele frequencies) and thus to have a comparable accuracy as the other methods.

Relatedness method has rarely been used in individual identifications. However, recently Ringler *et al.* (2015) showed that it is much more accurate than mismatch method for analysing their frog data. Relatedness method has several advantages over mismatch method. First, it uses allele frequency information. For example, two matched genotypes lend more support for a single individual if they are rare (i.e. containing rare alleles) than if they are common. Second, relatedness calculation is robust to the presence of mistypings. The relatedness estimates between close relatives (such as duplicates and full sibs) are reduced only slightly by assuming perfect data when

517 they are actually not (Wang 2007). My simulations conducted for different parameter combinations
518 confirm Ringler *et al*.'s conclusion that relatedness method is more accurate than mismatch method.
519 Importantly, relatedness method is statistically consistent. With an increasing number of markers,
520 even though they suffer from genotyping errors, the method always becomes increasingly more
521 accurate (Fig. 1).

522 Like the mismatch method, the relatedness method requires a threshold value, T_r, to
523 determine the relationship between two MGs. The dyad is concluded to be a single and two distinct
524 individuals when their relatedness is greater and not greater than T_r, respectively. Ideally, the
525 optimal T_r value that minimises both α- and β-errors should be obtained by considering the
526 frequencies of DPs and the most close relationship (e.g. FS) in the sample. These frequencies are
527 usually unknown, and the close relatives are most often full siblings and parent offspring, both
528 having an expected relatedness of 0.5. Using the average relatedness of first degree relatives (0.5)
529 and DPs (1.0) as threshold, I obtained T_r=0.75 and used it in simulated data analysis. This value is
530 slightly smaller than the value obtained by Ringler *et al*. (2015), 0.8, in their frog data analysis.
531 They derived this value from the estimated relatedness of the 20 juvenile-adult pairs identified as
532 identical from morphology. In practice, whenever a sufficient number of known duplicated
533 individuals are available, Ringler *et al*.'s approach should be followed to determine a dataset
534 specific T_r. Otherwise, a generic T_r =0.75 can be used in individual identification, bearing in mind
535 that the optimal value depends on the relative frequencies of DPs and the most close relationships as
536 well as genotyping error rates and other factors (e.g. number and polymorphisms of markers).
537 Further study (via simulation or meta-analysis) is needed to investigate the optimal T_r and the
538 factors affecting it.

539 Individual identification from a pairwise likelihood relationship (LR) analysis does not
540 require a threshold. We calculate the probability of two MGs conditional on each of a number of
541 candidate relationships, and the probability is the likelihood of the relationship. We then simply
542 select the relationship that has the maximal likelihood as the best estimate. Similar to the
543 considerations in relatedness analysis, we choose FS, HS, PO as well as DP as the candidate
544 relationships. Unlike relatedness analysis, however, relationship inference is highly susceptible to
545 mistypings, and a relationship (such as PO and DP) can be erroneously excluded because of
546 genotyping errors. For this reason, I used the error models of Wang (2004) to account for false
547 alleles (FA) and allelic dropouts (ADO) separately. Overall, LR method performs slightly better
548 than, but is more susceptible to mis-specified FA and ADO rates (Fig. 5) than relatedness (RL)
549 method. Recently, researchers have recognized the ubiquitous presence of mistypings and its large
550 impact on many downstream analyses (Bonin *et al*. 2004; Pompanon *et al*. 2005), and increasingly

quantified and reported mistyping rates. Therefore, the application of LR method should increasingly less limited by the lack of mistyping information.

A common problem of the above three methods is that they consider each pair of MGs in isolation of others. These pairwise approaches waste marker information and thus have low accuracy. For an example, let's consider n+1 MGs which are identical except for a single locus at which there are n heterozygous genotypes {A,B} and 1 single homozygous genotype {A,A}. These n+1 MGs would support the hypothesis that they come from a single individual rather than two distinct individuals when ADO or FA rate is not very small at the locus showing different genotypes and when n is large. The larger the value of n, the greater is the support. However, this support is much reduced when only 2 genotypes are considered as in the pairwise approach. Confirming the reasoning, Fig. 3 shows contrasting behaviours between LC and the 3 pairwise approaches. As the replication level increases, LC becomes more accurate, while pairwise approaches either remain the same accuracy or become less accurate. As a result, the difference in accuracy between LC and pairwise approaches increases with an increasing level of individual replication.

Another common problem of the above three methods is that they frequently yield self-incompatible inferences, as shown in a real example (Fig. 7). In practice, what one needs is usually the MG clusters, each corresponding to a single individual. This means one has to go through these pairwise inferences and assemble them into individual clusters. The process is not only tedious because of so many pairwise inferences, in the order of $N(N\text{-}1)/2$ where N is the number of MGs, but may fail to produce valid clusters.

Although the simulated data contain half sibs, they were analysed by LC by assuming monogamy for both sexes such that half sibs were not inferred. This is because half sibs are not of our interest and also have much smaller effect on individual identification than full sibs. Abandoning half sib inferences can however speed up the computation substantially and is thus especially favourable for a simulation study. In analysis of real data, it is also safe to ignore half sibs when individual identification is the purpose of analysis.

Highly polymorphic microsatellites from noninvasive samples have been used in identifying individuals and estimating population size (Waits & Leberg 2000; Creel *et al*. 2003; McKelvey & Schwartz 2004). It is anticipated that SNPs would become more and more widely used in the near future because of their low cost and high automation in genotyping. Although much less informative (usually biallelic) individually than microsatellites, SNPs can be genotyped at a much larger number of loci at ease and collectively they can be much more informative. My simulations

(Fig. 2) showed that all four methods can use markers of widely different polymorphisms in individual identification. However, the performance of the mismatch method, even when improved by using an optimal T_m, deteriorates rapidly with a decreasing marker polymorphism because of the excessive false identifications of individuals (α errors). The problem is much more severe if a fixed T_m value is used. In contrast, the LC method is especially more accurate than other methods with many markers of low polymorphisms. Using a number of 160 SNPs, each having 2 alleles and a mistyping rate of 0.05, LC has an overall accuracy several orders higher than other methods.

Except for the mismatch method that uses a fixed T_m value, allele frequencies are needed in inferring duplicates. Usually these frequencies are unavailable in practice, but can be estimated from the genotype data under the assumption that all homologous genes (within and between individuals) at a locus are non-identical by descent. The assumption is obviously violated when some sampled individuals are duplicated or otherwise related. However, violation of the assumption does not seem to cause a serious problem for all 4 methods investigated in this study, even when individual replication level is high (Fig. 3). The LC method implemented in Colony program does have the ability to account for the inferred genetic structure in refining allele frequency estimates, and has been proved to be effective in improving pedigree reconstruction when the families included in a sample are highly unbalanced in sizes (Wang 2004; Wang & Santure 2009).

My simulations assumed an outbred species without inbreeding. However, inbreeding or population structure could have some effects on the inference of duplicates. While it is not immediately apparent how to extend the MM, RL and LR methods to account for inbreeding, the LC method in Colony can actually accommodate inbreeding, including selfing, in relationship inference (Wang & Santure 2009). It can estimate inbreeding and relationship jointly. However, how much improvement in individual duplicate inference can be gained by allowing for inbreeding is yet to be investigated in a further study.

The simulation results for less related family structures, 40(1, 1) and 16(1, 1, 3), are similar to those shown in Figures 1-5. All methods become slightly more accurate, because full sib frequency is smaller and thus the chance of α errors is reduced. Overall across all simulated datasets and the empirical dataset, the LC method performs substantially better than the pairwise approaches, and is highly recommended for use in practice.

The LC and LR methods are implemented and added to the computer program COLONY version 2.0.5.3, which was used in analysing the data shown in this paper. The program is downloadable from the website http://www.zsl.org/science/software/colony.

Acknowledgements

I am grateful to Eva Ringler for her stimulating paper and for sending me her frog data analysed by this MS. I thank her and four anonymous referees for valuable comments which have helped improving the MS.

References

Alacs EA, Georges A, FitzSimmons NN, Robertson J (2010) DNA detective: a review of molecular approaches to wildlife forensics. *Forensic Science, Medicine, and Pathology*, **6**, 180-194.

Anderson EC, Dunham KK (2008) The influence of family groups on inferences made with the program Structure. *Molecular Ecology Resources*, **8**, 1219-1229.

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

Creel S, Spong G, Sands JL, Rotella J, Zeigle J, Joe L *et al*. (2003) Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. *Molecular Ecology*, **12**, 2003–2009.

Escaravage N, Questiau S, Pornon A, Doche B, Taberlet P (1998) Clonal diversity in a *Rhododendron ferrugineum* L. (Ericaceae) population inferred from AFLP markers. *Molecular Ecology*, **7**, 975–982.

Galpern P, Manseau M, Hettinga P, Smith K, Wilson P (2012) ALLELEMATCH: an R package for identifying unique multilocus genotypes where genotyping error and missing data may be present. *Molecular Ecology Resources*, **12**, 771-778.

Goodnight KF, Queller DC (1999) Computer software for performing likelihood tests of pedigree relationship using genetic markers. *Molecular Ecology*, **8**, 1231-1234.

Halkett F, Simon JC, Balloux F (2005) Tackling the population genetics of clonal and partially clonal organisms. *Trends in Ecology and Evolution*, **20**, 194-201.

Li CC, Weeks DE, Chakravarti A (1993) Similarity of DNA fingerprints due to chance and relatedness. *Human Heredity*, 43, 45–52.

Lynch M (1988) Estimation of relatedness by DNA fingerprinting. *Molecular Biology and Evolution*, **5**, 584–599.

Luikart G, Ryman N, Tallmon DA, Schwartz MK, Allendorf FW (2010) Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. *Conservation Genetics*, **11**, 355-373.

McKelvey KS, Schwartz MK (2004) Genetic errors associated with population estimation using non-invasive molecular tagging: problems and new solutions. *Journal of Wildlife Management*, **68**, 439-448.

Palsbøll P, Allen J, Bérubé M, Clapham PJ, Feddersen TP, Hammond PS, Hudson RR *et al*. (1997) Genetic tagging of humpback whales. *Nature*, **388**, 767–769.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. *Nature Review Genetics*, **6**, 847–850.

Ringler E, Mangione R, Ringler M (2015) Where have all the tadpoles gone? Individual genetic tracking of amphibian larvae until adulthood. *Molecular Ecology Resources*. (In press)

Rodríguez-Ramilo ST, Wang J (2012) The effect of close relatives on unsupervised Bayesian clustering algorithms in population genetic structure analysis. *Molecular Ecology Resources*, **12**, 873-884.

Schwartz MK, Tallmon DA, Luikart G (1998) Review of DNA-based census and effective population size estimators. *Animal Conservation*, **1**, 293–299.

Taberlet P, Waits L, Luikart G (1999) Non-invasive genetic sampling: look before you leap. *Trends in Ecology and Evolution*, **14**, 323–327.

Wagner AP, Creel S, Kalinowski ST (2006) Estimating relatedness and relationships using microsatellite loci with null alleles. *Heredity*, **97**, 336-345.

Waits L, Leberg PL (2000) Biases associated with population estimation using molecular tagging. *Animal Conservation*, **3**, 191-199.

Waits L, Luikart G, Taberlet P (2001) Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Molecular Ecology*, **10**, 249–256.

Wang J (2004) Sibship reconstruction from genetic data with typing errors. *Genetics*, **166**, 1963-1979.

Wang J (2006) Informativeness of genetic markers for pairwise relationship and relatedness inference. *Theoretical Population Biology*, **70**, 300-321.

Wang J (2007) Triadic IBD coefficients and applications to estimating pairwise relatedness. *Genetics Research*, **89**, 135–153.

Wang J (2014) Marker-based estimates of relatedness and inbreeding coefficients: an assessment of current methods. *Journal of Evolutionary Biology*, **27**, 518-530.

Wang J, Santure AW (2009) Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics*, **181**, 1579–1594.

Wilberg MJ, Dreher BP (2004) genecap: a program for analysis of multilocus genotype data for non-invasive sampling and capture-recapture population estimation. *Molecular Ecology Notes*, **4**, 783-785.

685

686

687 J. Wang is interested in developing population genetics models and methods of analysis of
688 empirical data to address issues in evolutionary and conservation biology.
689

690

691 **Data accessibility**

692 The simulated genotype datasets can be found on Dryad: Dryad doi: 10.5061/dryad.2q3qh

693 The frog dataset of Ringler et al. (2015) can be found on Dryad:
694 http://dx.doi.org/10.5061/dryad.db800

695 The computer program used in the simulated and empirical data analysis, Colony, is available
696 http://www.zsl.org/science/software/colony.

697

Figure Captions

Fig. 1 Effect of the number of markers. The upper graph plots the observed (x axis) and threshold T_m (y axis) numbers of mismatches of each simulated duplicated MG dyad for different number of markers (L). The lower graph plots the error rate (γ) of 4 individual identification methods as a function of the number of markers (L). The four methods are mismatch (MM), relatedness (RL), likelihood relationship (LR), and likelihood clustering (LC). For both graphs, the parameters used in the simulations are family structure 4(1, 2, 3, 4, 10), K_l =10 and $\varepsilon_{l1}=\varepsilon_{l2}=\varepsilon_{l3}=0.05$ for each locus l (=1, 2, …, L), λ=0.5.

Fig. 2 α-, β- and total-error rates of 4 individual identification methods as a function of the number of alleles per marker (K). The four methods are mismatch (MM), relatedness (RL), likelihood relationship (LR), and likelihood clustering (LC). The parameters used in the simulations are family structure 4(1, 2, 3, 4, 10), L=160, 80, 40, 20, 10, 5 when K =2, 3, 5, 9, 17 and 33 respectively, $\varepsilon_{l1}=\varepsilon_{l2}=\varepsilon_{l3}=0.05$ for each locus, and λ=0.5.

Fig. 3 Error rate (γ) of 4 individual identification methods as a function of the extent of individual replication (λ). The four methods are mismatch (MM), relatedness (RL), likelihood relationship (LR), and likelihood clustering (LC). The parameters used in the simulations are family structure 4(1, 2, 3, 4, 10), L=10, K =10, $\varepsilon_{l1}=\varepsilon_{l2}=\varepsilon_{l3}=0.05$ for each locus, λ (x axis) varies between 0 (no replication) to 3.2 (an individual is on average replicated by 3.2 times).

Fig. 4 Error rate (γ) of 4 individual identification methods as a function of the rate of mistyping and missing data at a locus (ε). The four methods are mismatch (MM), relatedness (RL), likelihood relationship (LR), and likelihood clustering (LC). The parameters used in the simulations are family structure 4(1, 2, 3, 4, 10), L=20, K =10, λ =0.5, $\varepsilon_{l1} \equiv \varepsilon_{l2} \equiv \varepsilon_{l3}$ (x axis) varies between 0 (perfect data with no mistyping and no missing data) to 0.16 at each locus l.

Fig. 5 Error rate (γ) of 4 individual identification methods as a function of the assumed rate of mistyping at a locus ($\hat{\varepsilon}$). The four methods are mismatch (MM), relatedness (RL), likelihood relationship (LR), and likelihood clustering (LC). The parameters used in the simulations are family structure 4(1, 2, 3, 4, 10), L=20, K =10, λ =0.5, $\varepsilon_{l1} \equiv \varepsilon_{l2} = 0.1$, ε_{l3} =0.05. The analysis was conducted assuming a mistyping rate (x axis) of $\hat{\varepsilon}_{l1} \equiv \hat{\varepsilon}_{l2}$ between 0 (perfect data with no mistyping) to 0.2 at each locus l.

Fig. 6 Distributions of the numbers of loci with missing data (A-D) and mismatches (E-H) between two MGs in the frog dataset. Row 1 (A and E) is for the 60 dyads in the 20 inferred trios that contain morphologically identified juvenile-adult pairs, row 2 (B and F) is for the 106 other dyads

730 inferred to be duplicates, row 3 (C and G) is for the inferred 16620 full sib dyads, and row 4 (D and
731 H) is for the inferred 1804400 non-full-sib dyads.

732 **Fig. 7** The relationships among 5 MGs inferred by LR for the frog dataset. In the 5 MG names, "it",
733 "m" and "ij" indicate tadpoles, male adults, and juveniles respectively. Two MGs are inferred by
734 LR to come from a single individual if they are linked by a line, and from distinct individuals if they
735 are not linked by a line.

736