Estimating genotyping errors from genotype and reconstructed pedigree data

Jinliang Wang

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

Left running head: J Wang

Right running head: Inferring genotyping errors

Key words: Pedigree, markers, sibling, parentage, mistyping rates, allelic dropouts, false alleles, null alleles

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

Summary

1. Genotyping errors are rules rather than exceptions in reality, and are found in virtually all but very small datasets. These errors, even when occurring at an extremely low rate, can derail many genetic analyses such as parentage/sibship assignments and linkage/association studies.
2. Nonetheless, few robust and accurate methods are available for estimating the rate of occurrence of genotyping errors and for identifying individual erroneous genotypes at a locus. Methods based on duplicate genotyping are expensive, and estimate genotype inconsistency rather than error rate at a locus. Methods based on Hardy-Weinberg equilibrium tests have low robustness and low power, and apply only to those particular errors that cause excessive homozygosity. Methods based on pedigrees are powerful, robust and accurate. However, they rely on known and complete pedigrees that are unfortunately rarely available from natural populations in the wild.
3. I proposed a maximum likelihood method to reconstruct pedigrees from genotype data with errors occurring at a roughly estimated (presumed) rate. In this paper, I describe how to use the method and inferred pedigree in estimating allelic dropout (or null allele) rate and false allele rate jointly at each marker locus, in identifying the erroneous genotypes, and in inferring the most likely genotypes at each locus of each individual. I examine the power, accuracy and robustness of the method by extensive simulations, and demonstrate the usefulness of the method by analysing three empirical datasets.
4. It is concluded that, both pedigrees and the rates of genotyping errors at each locus can be reliably estimated from the same genotype data by the same likelihood method, when marker information is sufficient and some sampled individuals are first-degree relatives. The erroneous genotypes are however inferred conservatively, and are reliably detected only when they occur in large families and/or at highly polymorphic loci. Estimation of genotyping error rates per locus and identification of erroneous genotypes of each individual at each locus should be routinely conducted to assess and improve data quality, to highlight markers for optimization of genotyping protocols or for replacement, and to enable the integration of genotyping errors in a robust statistical analysis.

Introduction

Genotype data are imperfect. All markers suffer from mutations. Many markers also have null alleles (Pemberton *et al*. 1995), and are prone to genotyping errors due to allelic dropouts and false alleles (Bonin *et al*. 2004). Both null alleles and genotyping errors lead to an observed genotype being different from the underlying true genotype. Together with mutations, they could cause the data to depart from Mendelian inheritance laws on which many population genetics analyses are based. These marker imperfections, including mutations, null alleles, allelic dropouts and false alleles, are broadly defined as genotyping errors, and can have a profound impact on a genetic analysis. Even occurring at a low rate of 1%, these errors can cause, for examples, false parentage exclusions (e.g. Pemberton *et al*. 1995; Wang 2010), false sibship exclusions (Wang 2004), false exclusion of duplicated individuals and thus overestimation of population size (Creel *et al*. 2003; Wang 2016), biased estimates of population differentiation (Chapuis & Estoup 2007), and much increased genetic map lengths in linkage analysis (e.g. Brzustowicz *et al*. 1993).

Genotyping errors can be due to many causes and error rates vary substantially among datasets and among loci within a dataset. However, all studies that examined mistypings reported a non-negligible error rate, from 0.2% to 15% per locus (Pompanon *et al*. 2005). Unfortunately, with an ever increasing dataset size (i.e. numbers of individuals and loci) and an increasing use of high-throughput genotyping (e.g. SNP array and NGS), error rates are likely to increase. SNPs called from low coverage NGS data can have a high mistyping rate because of NGS's random sampling nature and other multiple causes such as base-calling and alignment errors (Nielson *et al*. 2011).

Genotyping errors are easy to make but are difficult to spot, and their rate of occurrence is hard to estimate (Sobel *et al*. 2002; Douglas *et al*. 2002). In molecular ecology and evolutionary biology literature, error rates are usually defined (e.g. Broquet & Petit 2004; Bonin *et al*. 2004) and estimated (e.g. Johnson & Haydon 2007) from replicated genotyping data, such as those from the multitube approach (Bonin *et al*. 2004). However, such estimated error rates are really the frequencies of inconsistencies among replicate genotypes, not mistyping rates of the consensus genotypes in the final dataset. Some (hopefully many) errors may have been identified and eliminated in reaching the consensus genotypes if many duplicate genotypes are obtained and errors are not duplicable. In such an ideal situation, inconsistency provides an overestimate of the error rate of consensus genotypes. In contrast,

inconsistency can be an underestimate of error rate of the consensus genotypes when duplications are few or when the errors due to null alleles, mutations, and short allele dominance (Wattier *et al*. 1998) are themselves duplicable. Unfortunately, it is the error rate of the consensus genotypes that are relevant to any downstream analysis, and all types of errors, duplicable or not, count.

Mistyping rates can be estimated by quantifying the deviation of observed genotype frequencies from the expectation under Hardy-Weinberg equilibrium (HWE) (e.g. Chakraborty *et al*. 1992; Brookfield 1996). This approach is effective only in detecting null alleles and allelic dropouts, which can cause directional deviations from HWE (i.e. an excess of homozygotes). It is ineffective for mistypings such as false alleles and mutations that do not cause detectable distortions of genotype frequencies from HWE. The approach makes a critical but usually unrealistic assumption that all of the many factors (e.g. nonrandom mating), except for marker abnormity, that can potentially cause deviations from HWE are absent. Even when the assumption is met, the approach has low sensitivity and power (Cox & Kraft 2006), and cannot be used to identify erroneous genotypes individually.

A pedigree-based approach was proposed both to identify erroneous genotypes and to estimate the rate of occurrence of errors at a locus. It is based on examining genotype data against Mendelian inheritance laws in a known (e.g. Sobel *et al*. 2002) or reconstructed (e.g. Wang 2004) pedigree. It is robust to the underlying assumptions (e.g. non-random mating), applicable to the estimation of all kinds of errors, and is powerful. However, most algorithms (e.g. Sobel *et al*. 2002) rely on a known and correct pedigree that is usually unavailable in wild populations in molecular ecology, evolution, and conservation biology studies. In such situations, likelihood methods were proposed to reconstruct complete (Wang 2004; Wang & Santure 2009) or partial (e.g. Hadfield *et al*. 2006; Kalinowski *et al*. 2007) two-generation pedigrees of a sample of individuals using their marker genotype data with genotyping errors. These methods also have the potential to estimate genotype errors conditioned on the inferred pedigrees (Wang 2004).

Although the methodology was partly described in Wang (2004) for identifying erroneous genotypes and estimating error rate from inferred pedigrees, no studies have been conducted to investigate how accurate and powerful the method is, how robust the method is to assumptions such as presumed mistyping rates, and what the factors are in determining the performance of the method. Clarifying these issues is essential to allow the method to be

applied as a general tool for estimating genotype errors without known pedigrees. This study undertakes to address these issues by analysing simulated and empirical datasets. The results are discussed in the general context of genotype data quality assessment, control, estimation, and integration in a downstream genetic analysis.

Materials and Methods

I briefly describe the genotyping error models and the pedigree reconstruction models that were used (Wang 2004) in reconstructing pedigrees from error-prone genotype data. I will then detail the models for detecting erroneous genotypes and estimating error rates from the reconstructed pedigrees. Last, I describe a procedure used to simulate data, and a method used to assess the accuracy of error estimation methods applied to simulated data.

GENOTYPING ERROR MODELS

An error model defines the probability of an observed genotype, or phenotype, conditional on an underlying (unknown) genotype. Without genotyping errors, this probability is 1 when the genotype and phenotype are identical and is 0 otherwise. With genotyping errors, this probability lies between 0 and 1 for each possible genotype-phenotype combination. It is difficult to derive a simple, general and accurate error model because error patterns can be complex, marker dependent, and variable with genotyping platforms and DNA sample types (Bonin *et al*. 2004). Quite a few error models, with subtle differences, were proposed and used in the literature (e.g. Ott 1993; Sobel *et al*. 2002; Wang 2004). I describe and use the two models proposed by Wang (2004) to handle the broadly defined genotyping errors.

Model of allelic dropouts

It handles false homozygotes due to allelic dropouts of microsatellites during PCR. An allelic dropout occurs when PCR fails to amplify one of an individual's two homologous genes, leading to a false homozygote phenotype when the underlying genotype is a heterozygote. For microsatellites, allelic dropouts are believed to be the most frequent type of errors and can occur at a high rate when sample DNA quality and quantity is low (Taberlet *et al*. 1996). The error patterns of microsatellites with allelic dropouts mirror those of SNPs from low-coverage NGS. They are also similar to those produced by null alleles, which cannot be amplified by PCR and thus have no detectable phenotypes because of the presence of mutations in the primer binding sequences. A null allele homozygote has no detectable phenotype, and thus is indistinguishable from missing data. A null allele heterozygote shows

a homozygote phenotype of the detectable allele it contains. Both allelic dropouts and null alleles cause an apparent excess in homozygotes. As a result, they have similar impacts on a genetic analysis, are hardly distinguishable (see below) statistically, and can be handled by the same error model in an analysis.

Assuming both homologous genes in a diploid individual drop out during PCR at the same rate ε_1 and ignoring double dropouts (Wang 2004), I obtain $\Pr(G|g = A_1A_2) = 1 - 2e_1$, e_1, and e_1 for a heterozygote genotype g=A_1A_2 being observed as a phenotype G=A_1A_2, G=A_1A_1 and G=A_2A_2 respectively, where $e_1 = \varepsilon_1/(1 + \varepsilon_1)$. Under this dropout model, a homozygote genotype is unaffected, and shows faithfully the same phenotype at a probability of 1.

Model of false alleles

Apart from allelic dropouts and null alleles, the broadly defined erroneous phenotypes can also come from mutations, false alleles (polymerase errors rendering an allele other than the true one), allele miscalling, contaminant DNA, and data entry (Bonin *et al*. 2004). These errors usually are less frequent than dropouts, affect both heterozygote and homozygote genotypes, and do not cause an apparent excess of homozygotes. Such errors are pooled, termed "false alleles", and modelled by assuming that the two homologous genes in a diploid individual are independent to be incorrectly observed, with a rate ε_2. When incorrectly observed, an allele is observed to be any other allele at an equal probability of $1/(k-1)$, where k is the number of alleles at a locus. Therefore, an allele is correctly observed at a probability of $1 - \varepsilon_2$, and incorrectly observed to be any of the other alleles at a probability of $e_2 = \varepsilon_2/(k-1)$.

Probability of a phenotype given genotype

Combining both error models (Wang 2004) yields the probability of a phenotype, $G_{u,v}$, given its genotype, $g_{w,x}$,

$$\Pr\left[G_{u,v}\middle|g_{w,x}\right] = \begin{cases} (1-\varepsilon_2)^2 + e_2^2 - 2e_1(1-\varepsilon_2-e_2)^2 & \{(u=w, v=x)\} \\ e_2(1-\varepsilon_2) + e_1(1-\varepsilon_2-e_2)^2 & \{(u=v=w); (u=v=x)\} \\ (2-\delta_{uv})e_2^2 & \{(u\neq w, u\neq x, v\neq w, v\neq x)\} \\ e_2(1-\varepsilon_2-e_2) & \{\text{Otherwise}\} \end{cases} \quad \text{eqn 1}$$

and

$$\Pr\left[G_{u,v}\middle|g_{w,x}\right]=\begin{cases}(1-\varepsilon_2)^2 & \{(u=v=w)\}\\ 2e_2(1-\varepsilon_2) & \{(u=w,v\neq w);(v=w,u\neq w)\}\\ (2-\delta_{uv})e_2^2 & \{(u\neq w,v\neq w)\}\end{cases} \quad \text{eqn 2}$$

when $g_{w,x}$ is a heterozygote ($w \neq x$) and homozygote ($w = x$), respectively. In eqns 1-2, the Kronecker δ-variable takes values 1 and 0 when allele indexes *u*=*v* and *u*≠*v*, respectively.

PEDIGREE RECONSTRUCTION MODELS

The probability of the phenotypes of all individuals in a pedigree is the likelihood of the pedigree (Thomas & Hill 2002; Wang 2004). The likelihood function of a pedigree can be rather complicated, depending on the complexity of the pedigree (Wang & Santure 2009). For illustration, let us consider a simple pedigree of a single fullsib family containing *n* children. The likelihood of this pedigree, *R*, given phenotype data *D* is

$$L(R|D)=\Pr[R]\sum_f \Pr[f]\Pr[F|f]\sum_m \Pr[m]\Pr[M|m]\prod_{i=1}^{n}\Pr[C_i|f,m], \quad \text{eqn 3}$$

where *f* and *F* are the father's genotype and phenotype, respectively, and C_i is the phenotype of child *i* (=1~*n*). $\Pr[R]$ is the prior of *R*. $\Pr[f]$ is the probability of a father genotype. Under HWE, $\Pr[f]=p_w^2$ for a homozygote $f=g_{w,w}$ and $\Pr[f]=2p_wp_x$ for a heterozygote $f=g_{w,x}$, where p_a is the frequency of allele *a* (=*w*, *x*). $\Pr[F|f]$ is the probability of phenotype *F* given its genotype *f*. When *F* is unavailable (i.e. no candidate male is assigned to the father, or the assigned candidate has a missing phenotype), $\Pr[F|f]\equiv 1$. Otherwise, it is calculated by eqns 1 and 2. To be more exact, therefore, the probability of *F* is conditional not only on *f*, but also on the error models. Corresponding terms for a mother, *m*, *M*, $\Pr[m]$ and $\Pr[M|m]$, are defined and calculated similarly to *f*, *F*, $\Pr[f]$ and $\Pr[F|f]$ for a father, respectively.

The probability of C_i given the parental genotypes, $\Pr[C_i|f,m]$, can be obtained by deriving the underlying genotypes of the child from parental genotypes under Mendelian law, and by accounting for mistyping. Suppose $f=g_{w,x}$ and $m=g_{y,z}$, then

$$\Pr\left[C_i\middle|f=g_{w,x},m=g_{y,z}\right]=\tfrac{1}{4}\left(\Pr\left[C_i\middle|g_{w,y}\right]+\Pr\left[C_i\middle|g_{w,z}\right]+\Pr\left[C_i\middle|g_{x,y}\right]+\Pr[C_i|g_{x,z}]\right), \quad \text{eqn 4}$$

where each term on the right side of eqn 4 is calculated by eqns 1 and 2.

A dataset can be explained by a combinatorically large number of possible pedigrees with varying likelihood values. A simulated annealing algorithm (Wang 2004) can be used to sift through these pedigrees to find the one with the maximum likelihood.

ERROR ESTIMATION BASED ON A RECONSTRUCTED PEDIGREE

Conditional on a reconstructed pedigree, the underlying genotype at each locus of each individual in the pedigree can be estimated probabilistically. As an example, consider child *i* in a fullsib family as considered by eqn 3. It has $k(k+1)/2$ possible genotypes at a locus with *k* codominant alleles. The likelihood of the underlying genotype being $g_{u,v}$ (where *u*, *v* = 1~*k*, and $v \geq u$) is calculated by eqn 3, with eqn 4 being replaced by

$$\Pr[C_i|f = g_{w,x}, m = g_{y,z}] =$$

$$\tfrac{1}{4}\Pr[C_i|g_{u,v}]\left(\Pr[g_{u,v}|g_{w,y}] + \Pr[g_{u,v}|g_{w,z}] + \Pr[g_{u,v}|g_{x,y}] + \Pr[g_{u,v}|g_{x,z}]\right). \qquad \text{eqn 5}$$

In eqn 5, $\Pr[g_{u,v}|g_{s,t}] = (\delta_{us}\delta_{vt} + \delta_{ut}\delta_{vs})/(1+\delta_{st})$ is the probability that child *i* has genotype $g_{u,v}$ when it inherits parental alleles *s* and *t*, where $s,t = w,x,y,z$ and the Kronecker delta δ_{st}, δ_{us}, δ_{vt}, δ_{ut} and δ_{vs} are defined as above.

Using Bayes' rule, these likelihood values can be transformed to posterior probabilities of the inferred $k(k+1)/2$ genotypes. The genotype with the maximum posterior probability is the best point estimate. If the posterior probability of the genotype identical to the observed phenotype C_i (i.e. no genotype errors) is T, then a genotype error (or an erroneous phenotype C_i) is detected at the significance level T. Erroneous phenotypes and the most likely genotypes of assigned parents are inferred similarly.

Error rates ε_j (where *j*=1,2 for allelic dropouts and false alleles) at each locus can also be estimated, conditional on a reconstructed pedigree. For each locus, the likelihood function such as eqn 3 is maximised with respect to ε_j varying in the range [0, 1] to obtain the maximum likelihood estimate of ε_j, $\hat{\varepsilon}_j$. I use Powell's (1964) conjugate direction method to find $\hat{\varepsilon}_j$, and the profile likelihood method to find the 95% confidence intervals of $\hat{\varepsilon}_j$. To avoid the algorithm converging to a local rather than the global maximum, multiple runs with different random starting values of ε_1 and ε_2 were used in maximizing the likelihood function. When the same maximum likelihood value and the same error rate estimates were obtained

repeatedly from these replicate runs, the estimates were reported and the algorithm was terminated.

The algorithm for identifying erroneous phenotypes and inferring the most probable underlying genotype for each individual at each marker locus, and the algorithm for estimating allelic dropout rates and false allele rates at each locus are implemented in the current version of software Colony, available at https://www.zsl.org/science/software/colony.

SIMULATIONS

Numerous factors influence the quality of reconstructed pedigrees (Wang 2004) and thus genotype error estimates. Herein I focus on just a few of them.

Actual relatedness

The performance of the method depends critically on the relatedness structure in a pedigree. To show the effects of pedigree relatedness, I conducted three simulations. Simulation 1 considered a sample of 160 individuals in $160/2^M$ fullsib families, each having 2^M individuals, where M=0, 1, 2, 3, 4. No candidate males and females as potential parents are available. Simulation 2 is the same as simulation 1, except that a fullsib family is replaced by a halfsib family in which all individuals share the same father but distinctive mothers. Simulation 3 considered a sample of 160 unrelated offspring, and a sample of 160 unrelated candidate males. The numbers of parent-offspring pairs between the two samples are $10\times2^{M-1}$ =5, 10, 20, 40, and 80 for M=0, 1, 2, 3, 4, respectively. Simulations 1-3 used 20 markers, each having 10 alleles in a uniform frequency distribution and each having an error rate of $\varepsilon_1 = \varepsilon_2 = 0.05$. The multilocus genotype of each individual was simulated given the pedigree and simulated allele frequencies, assuming HWE and linkage equilibrium. It was then modified according to the error models and rates to generate the multilocus phenotypes, which were analysed for pedigree and error estimation.

Marker information

It is determined by the number of loci, the number and frequency distribution of alleles per locus, as well as the mistyping and missing data rates. Simulation 4 considered the effect of the number of loci, L (=2, 4, 8, 16, 32), when each has 10 alleles, and the number of diallelic loci, L (=50, 100, 200, 400, 800). Simulation 5 considered the effect of the number of alleles per locus, k (=2, 4, 8, 16, 32), assuming a fixed total number of $Lk = 320$ alleles across loci. For both simulations, allele frequencies were drawn from a uniform distribution, and the

pedigree structure of fullsib families with each having 4 siblings, the sample size, and error rates ($\varepsilon_1 = \varepsilon_2 = 0.05$) of simulation 1 were adopted.

Prior error rates

The assumed error rates used in pedigree reconstruction may deviate from the true values, and may affect the quality of pedigree and genotyping error inferences. Simulation 6 generated genotype data at 200 SNPs at a true error rate of $\varepsilon_1 = \varepsilon_2 = 0.025$, using the other parameters as those in simulation 4. The data were analysed by assuming widely different values of prior error rate.

Actual error rates

They affect marker information quality and quantity, and thus the quality of reconstructed pedigrees and error estimates. To show the pedigree-based method can be applied to data of varying error rates, simulation 7 generated data with actual error rates varying in the wide range $\varepsilon_1 = \varepsilon_2 = [0, 0.32]$. The values of other parameters were the same as those in simulation 4, except that the number of loci was fixed at L =16.

Null allele frequency

Null alleles are not modelled but their frequency (r) can be estimated as ε_1 (see more below). Simulation 8 was conducted to check the quality of r estimates when null alleles were estimated as allelic dropouts. Twenty loci, each having 10 observable alleles and one null allele, were simulated for a sample of 160 offspring coming equally from 40 full-sib families. The null allele frequency at each locus was r, varying in the range [0, 0.32], while the other allele frequencies follow a uniform distribution. In transforming genotypes to phenotypes, a null allele homozygote was taken as missing phenotype, while a null allele heterozygote was taken as the homozygote of the observed allele. Allelic dropouts and false alleles were assumed absent ($\varepsilon_1 = \varepsilon_2 = 0$) in simulating data.

ANALYSES OF SIMULATED DATA

For each simulation described above, 100 replicate datasets were generated and analysed by the Colony program (Jones & Wang 2010) for estimating ε_1 (or r) and ε_2, and identifying erroneous genotypes. For all but simulation 6, a locus specific value, $\check{\varepsilon}_i$, was drawn at random from a uniform distribution in the range [$0.1\varepsilon_i$, $2\varepsilon_i$], where ε_i (for i=1, 2) was the actual

(simulated) error rate. $\check{\varepsilon}_i$ was taken as the roughly estimated or presumed error rate, and was used in Colony analyses. For simulation 6, either $\check{\varepsilon}_i = 0.1\varepsilon_i$ or $\check{\varepsilon}_i = 4\varepsilon_i$ was adopted for each locus in analysing the data. The simulated type of sibship (i.e. full- or half-sib families) and the default values of other parameters in Colony were used in analysing the data.

For a large random mating population without mutation, selection and migration, a marker without genotyping errors should be in HWE. Violations of these assumptions may lead to a deviation from HWE. If allelic dropouts are the sole cause of the deviation, then the observed deviation can be used to estimate ε_1. For a locus with *k* codominant alleles of population frequencies p_i (*i*=1~*k*) and with dropout rate ε_1, the heterozygosity is expected to be

$$H_E = H_e(1 - \frac{2\varepsilon_1}{1+\varepsilon_1}), \qquad \text{eqn 6}$$

where $H_e = 1 - \sum_{i=1}^{k} p_i^2$ is the expected heterozygosity under HWE without dropouts ($\varepsilon_1 = 0$). Eqn 6 shows that H_E increases monotonically with a decreasing value of ε_1 to attain its maximum $H_E = H_e$ when $\varepsilon_1 = 0$. Equating H_E to the observed heterozygosity $H_O = n_1/(n_1 + n_2)$ where n_1 and n_2 are the observed numbers of heterozygotes and homozygotes respectively, I obtain a moment estimator of ε_1,

$$\hat{\varepsilon}_1 = \frac{H_e - H_O}{H_e + H_O}. \qquad \text{eqn 7}$$

It turns out that eqn 7 is exactly the same estimator as that derived by Chakraborty *et al*. (1992) for null allele frequency *r* from a sample of individuals without (or ignoring) missing phenotypes. This shows that null alleles and allelic dropouts have the same effect on homozygosity, and therefore $\hat{\varepsilon}_1$ and $\hat{r}$ can be estimated by the same equation. In eqns 6 and 7, population allele frequencies p_i are usually unknown, but can be replaced by estimates from sample phenotype data.

QUALITY OF GENOTYPING ERROR ESTIMATORS

The mean of estimates was calculated and compared with the true simulated value to indicate the bias of an estimator. The accuracy of an estimator was measured by its root mean squared error, RMSE,

$$RMSE_i = \sqrt[2]{\frac{1}{N}\sum_{j=1}^{N}(\hat{\varepsilon}_{ij} - \varepsilon_i)^2}, \qquad \text{eqn 8}$$

where *i*=1,2 for allelic dropouts and false alleles, respectively, ε_i is the true value and $\hat{\varepsilon}_{ij}$ is the estimated value from the *j*th of *N* replicates, respectively. The bias and accuracy of a null allele frequency estimator were measured similarly.

For a simulated dataset with *M* individuals and *L* loci, the *ML* phenotypes can be partitioned into sets Φ_1 and Φ_0 containing n_1 erroneous and n_0 error-free phenotypes, with $n_1 + n_0 = ML$. Colony gives the posterior probability, $\Pr[g_x = G_x|D, R]$, of an inferred genotype g_x identical (i.e. no genotyping errors) to the phenotype G_x, given data *D* and pedigree *R*. The average of these probabilities for phenotype set *i* is

$$\rho_i = \frac{1}{n_i}\sum_{G_x \in \Phi_i} \Pr[g_x = G_x|D, R], \qquad \text{eqn 9}$$

for *i*=0,1. Thus, ρ_0 calculates the average posterior probability that a phenotype is inferred to be free of genotyping errors when there are no genotyping errors, and ρ_1 calculates the average posterior probability that a phenotype is inferred to be free of genotyping errors when there are in fact genotyping errors. Ideally, type I and type II errors of the method are minimized when ρ_0 approaches 1 and ρ_1 approaches 0.

Statistics ρ_0 and ρ_1 are sufficient for measuring the power and accuracy of the method in identifying erroneous genotypes in simulated data. In empirical datasets, however, whether a phenotype is erroneous or not is unknown and forms part of the inferences, and thus the statistics cannot be applied. For identifying erroneous phenotypes, I set a threshold *T* that a phenotype G_x is inferred to be correct and erroneous when the posterior probability $\Pr[g_x = G_x|D, R]$ is larger and smaller than *T*, respectively. *T* value affects rates of type I and type II errors, but its choice is somewhat arbitrary. Considering the conservative nature of identifying genotyping errors (Sobel *et al*. 2002) by pedigree-based analysis, I choose *T*=0.5 in calculating frequencies, F_0 and F_1, that single-locus and single-individual phenotypes that are simulated without and with genotyping errors, respectively, are inferred to be erroneous.

THREE EMPIRICAL DATASETS

An ant dataset (Hammond *et al*. 2001), a spectacled caiman dataset (Oliveira *et al*. 2014), and a sockeye salmon dataset (Hauser *et al*. 2011) were reanalysed for pedigree reconstruction, erroneous genotype identification, and genotyping error rate estimation by the method described above. Details of the three datasets are in Supporting Information.

Results

Actual relatedness

For fullsib families, both likelihood method (for $\hat{\varepsilon}_1$ and $\hat{\varepsilon}_2$) and moment method (for $\hat{\varepsilon}_1$) are almost unbiased (Figure 1). With an increasing fullsib family size, the likelihood method becomes more accurate for both ε_1 and ε_2, while the moment method becomes less accurate for ε_1. The likelihood method estimates ε_1 more accurately than ε_2 when family size is small, and the opposite is true otherwise. For $\hat{\varepsilon}_1$, the likelihood method is slightly less accurate than the moment method when all individuals are unrelated, but quickly becomes more accurate with an increasing full sib family size.

Similar results were obtained for halfsib pedigrees. The changes in accuracy (RMSE) as a function of family size are less dramatic than those of fullsib pedigrees for both methods and for both error rates. For pedigrees involving parent-offspring relationships only, the likelihood method underestimates ε_2 consistently. As a result, it gives less accurate estimates of ε_2 than those of ε_1. For different numbers of parent-offspring dyads in a dataset, the likelihood method is always less biased and more accurate than the moment method (Figure 1) in estimating ε_1.

Figure 2 shows the power and accuracy of the likelihood method for identifying erroneous genotypes in reconstructed fullsib families. As expected, the power is low in a small family with less than 3 siblings, because any sibling phenotypes are compatible with Mendelian inheritance, and have similar likelihood values. With an increase in family size, the power increases rapidly. At 16 full siblings per family, 41% of the erroneous phenotypes were detected as such while only 0.01% of the error-free phenotypes were falsely identified as erroneous. Similar results were obtained for halfsib pedigrees. Overall, it is much more difficult to identify erroneous genotypes than to estimate error rates. Only those genotyping errors that occur in parent-offspring or large sib families are detectable.

Marker information

Eqn 7 is a single-locus moment estimator of allelic dropout rate, and its performance is therefore unaffected by the number of loci (L) (Figure 3). In contrast, the likelihood estimator uses multilocus information to reconstruct pedigrees and to infer genotyping errors. Its

performance is therefore sensitive to marker information, increasing rapidly with an increase in both *L* and *k* (Figure 3). Except when marker information is rather scarce (i.e. small *Lk*), the likelihood method is more accurate and less biased than the moment method for estimating ε_1.

Prior error rates

When prior error rates are smaller than the actual values, the likelihood method yields little biased and accurate $\hat{\varepsilon}_1$ and $\hat{\varepsilon}_2$ values consistently across pedigrees of different family sizes (Figure 4), including a pedigree in which all individuals are completely unrelated. The likelihood estimates are little improved by an increasing full-sib family size. In contrast, a prior much larger than (i.e. 4 times of) the actual error rate leads to overestimated and thus inaccurate $\hat{\varepsilon}_1$ and $\hat{\varepsilon}_2$ values when few sampled individuals are siblings (Figure 4). However, with an increasing full-sib family size, the likelihood estimates improve rapidly and converge to those obtained with a prior much smaller than the actual error rate. The likelihood estimator performs well and is almost independent of the prior error rates for a sample containing some full siblings, and for a sample containing few full siblings when conservative prior error rates are adopted.

Actual error rates

For the wide range of true error rates $\varepsilon_1 = \varepsilon_2 = [0, 0.32]$, the likelihood estimates of both ε_1 and ε_2 are little biased (Figure 5). The moment method, however, is unbiased only when ε_1 is small, and underestimates ε_1 substantially when it is large. The underestimation occurs because, at high values of $\varepsilon_1 = \varepsilon_2$, homozygosity excess produced by allelic dropouts is partly destroyed by false alleles which are assumed absent by the moment method. With an increasing true error rate of $\varepsilon_1 = \varepsilon_2$, both estimators become less accurate.

Null allele frequency

Both likelihood and moment estimators give unbiased estimates of *r* in the entire range of *r*=[0, 0.32] (Figure 6). The likelihood method also yields unbiased estimates of ε_2, whose true value is 0. With a decreasing true null allele frequency, the likelihood estimator becomes increasingly more accurate than the moment estimator.

Analysis results of three empirical datasets

Colony completely recovered the actual sibship structure of the 377 ant workers, and yielded highly consistent error rate estimates at each locus (Supporting Information), regardless of the presumed values of $\varepsilon_1 = \varepsilon_2$ in the wide range of [0.001, 0.256]. Additionally, it identified eight erroneous phenotypes across the 377 individuals at six loci. These errors are all due to false alleles, and are highly reliable thanks to the strong family structure and the haplodiploid inheritance of the dataset.

For the spectacled caiman dataset, the estimated mistyping rates are generally low and false allele rates are uniformly higher than allelic dropout or null allele rates across loci (Supporting Information). At the low presumed error rate of 0.01, 6 and 2 errors were identified across the 174×6=1044 hatchling phenotypes and the 13×6=78 mother phenotypes, respectively. Overall, the power and accuracy of the likelihood method in estimating mistyping rates and in identifying erroneous phenotypes are lower for this dataset than for the ant dataset, because the family sizes are smaller and the species is diploid.

For the sockeye salmon dataset, it emerges that microsatellites have higher mistyping rates than SNPs (Supporting Information). On average across loci, the estimated null allele (or dropout) rate and false allele rate are 3.1% and 1.2% for microsatellites, 1.1% and 0.1% for SNPs. For both types of markers, false alleles are much less frequent than null alleles or allelic dropouts. This error pattern is in contrast with that in the ant dataset and the spectacled caiman dataset, where false alleles are the predominant type of genotyping errors. The likelihood method identified 39 erroneous phenotypes across loci among the 211 sampled offspring.

Discussion

This study showed that, given a sufficient, and nowadays realistic, amount of marker data, two-generation pedigrees can be reconstructed reliably. Conditional on the inferred pedigree, the underlying genotypes can be inferred for the phenotype of each individual at each locus to detect genotyping errors and to infer the most likely genotype. Similarly, maximizing the probability of the phenotype data given the inferred pedigree yields unbiased and accurate estimates of the rate of errors of each type occurred at each locus. The power and accuracy of these error estimation methods were checked by analysing simulated data, and demonstrated by analysing empirical data. The quality of error estimates based on reconstructed pedigrees depends, as partially shown in Figures 1-6, on many factors such as the true pedigree, the

quality of an inferred pedigree, the marker informativeness, and the actual error rate. Herein I briefly discuss each factor.

The types and frequencies of first-degree relatives in a pedigree affect the quality of error estimates. Loosely speaking, pedigree-based methods examine the inheritance of marker phenotypes against Mendelian laws implied by the pedigree, and a lack of conformity to the laws is interpreted as genotyping errors. A fullsib family having more than four distinct alleles at a locus, for example, is inconsistent with Mendelian law (Douglas *et al*. 2002) in diploid species, and signifies erroneous phenotypes. The greater the number of closely related individuals is in a pedigree, the more power and the higher accuracy the pedigree allows for error detection and estimation. This is verified in Figure 1, especially for the cases of fullsib and halfsib families. The good news is that, even in the extremely unfavourable case of a sample of completely unrelated individuals, the pedigree-based method still provides reasonably good estimates of both ε_1 and ε_2 (Figures 1 and 4), especially when conservative prior values of error rates are adopted.

My simulations considered three simple structures of pedigrees (full-sib, half-sib, parent-offspring) with equal family sizes. However, the method and the general conclusions apply to any two-generation pedigrees of any complexity. A genotype error in a large family has the largest detrimental effects on analyses such as pedigree reconstruction and linkage mapping, but is also the easiest to detect by pedigree-based method. For error detection and error rate estimation, therefore, the best and the worst scenarios are a pedigree with uniformly large and uniformly small family sizes, respectively, as shown in Figure 1. The intermediate scenario is a pedigree with mixed small and large families, where most genotype errors are detectable from large families but undetectable from small families. Not only the structure and size (i.e. number of individuals), but also the depth (i.e. number of generations) of a pedigree affects pedigree-based error inferences. I considered one- (full- and half-sib) and two-generation (parent-offspring) pedigrees, but pedigrees with more than two generations could allow for more accurate error inferences. Second-degree or more remote relatives (e.g. cousins) are also informative about genotype errors. How much more power and accuracy can be gained from a pedigree with more than two generations requires further investigation.

If a pedigree itself is grossly misconstrued, then its value for marker error estimation would be greatly compromised (Figure 3). There are several causes for pedigree mis-inference, the common ones being a lack of marker information (e.g. few markers with low

polymorphisms), low marker quality (e.g. many erroneous or missing genotypes), and a difficult true pedigree (e.g. scant close relationships) to reconstruct. The challenge to inferring pedigrees and genotyping errors jointly from the same data, especially when teemed with mistypings that are difficult to detect (such as those in SNPs, see below and Douglas *et al*. 2002), is enormous and can only be tackled effectively when there is sufficient marker information (Figure 3). Previous work avoided the challenge by using a known pedigree in estimating genotyping errors (e.g. Sobel *et al*. 2002). This simplified the inference greatly, and enabled the handling of pedigrees of any size and complexity and the handling of linked markers (Sobel *et al*. 2002). In evolutionary biology and molecular ecology studies of wild populations, however, pedigrees are rarely available or complete.

Marker information and quality determines to what extent a pedigree can be recovered, and whether genotyping errors are detectable or not (see below). A set of markers with few loci, low polymorphisms (Figure 3) or high typing errors (Figures 5-6) provides insufficient information to reveal the underlying pedigree of sampled individuals, resulting in poorly reconstructed pedigrees and poorly inferred genotyping errors. If marker information is deemed scarce, pedigree reconstruction and error estimation should be conducted with great caution.

The number and frequency distribution of alleles at a marker locus determines the difficulty and thus accuracy and power of error identifications. Errors in diallelic markers, such as SNPs, are much more difficult to detect than those in multiallelic markers such as microsatellites (Douglas *et al*. 2002), as shown in Figure 3. This is evident by considering the simple case of a family with three or more full siblings without parental genotypes. Any genotype combinations are consistent with Mendelian inheritance for a diallelic marker, but an increasing proportion of genotype combinations is inconsistent with Mendelian inheritance for a marker with an increasing number of alleles (Douglas *et al*. 2002).

It should be emphasized that error detection and error rate estimation are two different statistical inferences and have different applications (below). Error detection aims to identify erroneous genotypes of each individual at each locus. By nature, error detection is conservative because not all errors are detectable (Douglas *et al*. 2002; Sobel *et al*. 2002). While Mendelian inconsistent errors can be detected reliably, Mendelian consistent errors can only be detected with sufficient confidence when they occur in large families and/or at a multiallelic locus. Unfortunately, Mendelian consistent errors are usually more abundant than

Mendelian inconsistent errors, especially for markers of low polymorphisms such as SNPs (Douglas *et al*. 2002). Nonetheless, both Mendelian consistent and inconsistent errors can cause sporadic results, such as those in linkage or disease association studies. The pedigree-based method can conservatively detect errors in both categories. The high power of error detection is exemplified in the ant dataset in which sibship size is uniformly large and the species is haplodiploid (Table A1, Supporting Information). Considering the conservative nature of error detection, errors detected are probably true, but phenotypes not flagged as erroneous are not necessarily error-free.

In contrast, error rate estimation aims to give an unbiased inference of the frequency of genotyping errors at a locus, despite some genotype errors are undetectable individually. Indeed, simulations showed that, whenever the reconstructed pedigree does not deviate much from the truth, both allelic dropout (or null allele) rate and false allele rate can be accurately estimated (Figure 1, 3-6).

Although I focused on the sibship and parentage reconstruction method implemented in Colony, other methods that account for genotyping errors also have the potential to infer data errors. For example, the program MasterBayes (Hadfield *et al*. 2006) uses both genotype and behaviour data for inferring parentage in a Bayesian framework. It employs the error model of Wang (2004), but infers a partial two-generation pedigree by ignoring sibship. As a result, it can run much faster than Colony, but may yield less accurate inferences of both parentage and genotyping errors due to the ignorance of sibship. As shown in Figures 1 and 4, the power and accuracy of pedigree-based estimators depend critically on family size, and are particularly poor for false allele rate when it is inferred from individuals who are related as parent-offspring only (i.e. no siblings). A formal comparison of Colony, MasterBayes and other methods for inferring genotyping errors warrants further studies.

What can we do with the identified errors and the estimated error rates? First, they can be used to assess the quality of genotype data. Analysis of genotyping errors should be routinely conducted, and error rate for each marker should be routinely reported just like the number of alleles and the expected heterozygosity.

Second, the detected errors are useful in improving data quality. If a genotype is flagged as erroneous, one should re-examine the original image or genotyping score, and if the genotype is still unresolved and resources permit, re-genotype the particular individual at the particular locus. When it is impossible to re-genotype because the resources are

unavailable or too limited, I suggest, following Sobel *et al*. (2002), dropping the flagged genotype in performing a statistical analysis that is sensitive to data errors. Although the most likely genotype of an error-flagged phenotype is provided by the pedigree-based likelihood method, it should be treated with caution, even when its posterior probability is high. In principle, a flagged erroneous genotype should not be replaced by the inferred most likely genotype. Instead, experimental evidence (e.g. images or genotyping scores) should be re-examined or collected anew to resolve the problem. After all, missing data (by dropping flagged genotypes) are better than erroneous data (by accepting incorrectly inferred genotypes), as the former leads to a loss of precision and power while the latter to false conclusions.

Third, the estimated error rates are valuable in prioritising and highlighting markers for optimization of genotyping protocols or for replacement. A researcher usually chooses microsatellites based on their polymorphisms and ease of scoring, but seldom on their error rates. However, it could be better to choose markers of lower error rates rather than higher polymorphisms in an error-sensitive statistical analysis. More importantly, the estimated error rates can be integrated in a statistical framework to use the marker information with discretion (Wang 2004; Hadfield *et al*. 2006). It is probably easy to reduce genotyping error rate experimentally, with or without the assistance of a statistical approach shown in this study, from a high value (say, 10%) to a low value (say, 1%) by various measures (Bonin *et al*. 2004). However, it is virtually impossible or too expensive to completely eliminate all errors. Therefore, the importance of estimating data error rates and integrating them into error-tolerant statistical methods cannot be overstated. It is not a good strategy to simply drop a marker out of an analysis if it has a high error rate. Such a cautious strategy could reduce bias, but can cause a loss of precision due to the waste of information. This is especially true when only a few (say 10~20) markers are available to an analysis, and abandoning one marker means a substantial loss of information and waste of resources. A better strategy is to obtain a good estimate of the error rate of a problematic marker, and to integrate the estimate in a downstream analysis.

Acknowledgements

The author is grateful to the editor, Michael Morrissey, and two anonymous referees for insightful and constructive comments and suggestions on earlier versions of the MS.

Data accessibility

The three empirical datasets in Colony input file format are deposited in the Dryad repository: http://datadryad.org/resource/doi:10.5061/dryad.vv0gg

References

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.

Brookfield, J. (1996) A simple new method for estimating null allele frequency from heterozygote deficiency. *Molecular Ecology*, **5**, 453-455.

Broquet, T. & Petit, E. (2004) Quantifying genotyping errors in noninvasive population genetics. *Molecular Ecology*, **13**, 3601-3608.

Brzustowicz, L.M., Merette, C., Xie, X., Townsend, L., Gilliam, T.C. & Ott, J. (1993) Molecular and statistical approaches to the detection and correction of errors in genotype databases. *The American Journal of Human Genetics*, **53**, 1137–1145.

Chakraborty, R., De Andrade, M., Daiger, S.P. & Budowle, B. (1992) Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. *Annals of Human Genetics*, **56**, 45-57.

Chapuis, M.P. & Estoup, A. (2007) Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution*, **24**, 621-631.

Cox, D.G. & Kraft, P. (2006) Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. *Human Heredity*, **61**, 10-14.

Creel, S., Spong, G., Sands, J.L., Rotella, J., Zeigle, J., Joe, L., Murphy, K.M. & Smith, D. (2003) Population size estimation in Yellowstone wolves with error- prone noninvasive microsatellite genotypes. *Molecular Ecology*, **12**, 2003–2009.

Douglas, J.A., Skol, A.D. & Boehnke, M. (2002) Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *The American Journal of Human Genetics*, **70**, 487-495.

Hadfield, J.D., Richardson, D.S. & Burke, T. (2006) Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. *Molecular Ecology*, **15**, 3715-3730.

Hammond, R.L., Bourke, A.F.G. & Bruford, M.W. (2001) Mating frequency and mating system of the polygynous ant, *Leptothorax acervorum*. *Molecular Ecology*, **10**, 2719–2728.

Hauser, L., Baird, M., Hilborn, R.A.Y., Seeb, L.W. & Seeb, J.E. (2011) An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (*Oncorhynchus nerka*) population. *Molecular Ecology Resources*, **11**(s1), 150-161.

Johnson, P.C. & Haydon, D.T. (2007) Maximum-likelihood estimation of allelic dropout and false allele error rates from microsatellite genotypes in the absence of reference data. *Genetics*, **175**, 827-842.

Jones, O.R. & Wang, J. (2010) COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, **10**, 551–555.

Kalinowski, S.T., Taper, M.L. & Marshall, T.C. (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, **16**, 1099-1106.

Nielsen, R., Paul, J.S., Albrechtsen, A. & Song, Y.S. (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443-451.

Oliveira, D.P., Marioni, B., Farias, I.P. & Hrbek, T. (2014) Genetic evidence for polygamy as a mating strategy in *Caiman crocodilus*. *Journal of Heredity*, **105**, 485–492.

Ott, J. (1993) Detecting marker inconsistencies in human gene mapping. *Human Heredity*, **43**, 25–30.

Pemberton, J.M., Slate, J., Bancroft, D.R. & Barrett, J.A. (1995) Nonamplifying alleles at microsatellite loci: a caution for parentage and population studies. *Molecular Ecology*, **4**, 249-252.

Pompanon, F., Bonin, A., Bellemain, E. & Taberlet, P. (2005) Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, **6**, 847-846.

Powell, M.J.D. (1964) An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Computer Journal*, **7**, 155–162.

Sobel, E., Papp, J.C. & Lange, K. (2002) Detection and integration of genotyping errors in statistical genetics. *The American Journal of Human Genetics*, **70**, 496-508.

Taberlet, P., Friffin, S. Goossens, B. Questiau, S. Manceau, V. *et al*. (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research* **24**, 3189–3194.

Thomas, S.C. & Hill, W.G. (2002) Sibship reconstruction in hierarchical population structures using Markov chain Monte Carlo techniques. *Genetical Research*, **79**, 227-234.

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

Wang, J. (2010) Effects of genotyping errors on parentage exclusion analysis. *Molecular Ecology*, **19**, 5061-5078.

Wang, J. (2016) Individual identification from genetic marker data: developments and accuracy comparisons of methods. *Molecular Ecology Resources*, **16**, 163-175.

Wang, J. (2017) Data from: Estimating genotyping errors from genotype and reconstructed pedigree data. Methods in Ecology and Evolution doi:10.5061/dryad.vv0gg

Wang, J. & Santure, A.W. (2009) Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics*, **181**, 1579-1594.

Wattier, R., Engel, C.R., Saumitou- Laprade, P. & Valero, M. (1998) Short allele dominance as a source of heterozygote deficiency at microsatellite loci: experimental evidence at the dinucleotide locus Gv1CT in *Gracilaria gracilis* (Rhodophyta). *Molecular Ecology*, **7**, 1569-1573.