HB Harrison, P Saenz-Agudelo, S Planes, GP Jones, ML Berumen (2013). On minimizing assignment errors and the trade-off between false positives and negatives in parentage analysis. Molecular ecology, 22, 5738-5742.

On minimising assignment errors and the trade-off between false positives and negatives in parentage analysis

Hugo B. Harrison[1,2,3], Pablo Saenz-Agudelo[4], Serge Planes[3], Geoffrey P. Jones[1,2], Michael L. Berumen[4,5]

[1]School of Marine and Tropical Biology, James Cook University, Townsville, Queensland 4811, Australia. [2]Australian Research Council Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Queensland 4811, Australia. [3]Laboratoire d'Excellence "CORAIL", USR 3278 CRIOBE CNRS-EPHE, CRIOBE, BP 1013, 98729 Moorea, French Polynesia. [4]Red Sea Research Center, King Abdullah University of Science and Technology, 23955-6900 Thuwal, Kingdom of Saudi Arabia. [5]Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA.

Correspondence: hugo.harrison@my.jcu.edu.au

Running title: Minimising errors in parentage analysis

Abstract

Genetic parentage analyses provide a practical means with which to identify parent-offspring relationships in the wild. In Harrison *et al.* (2013a), we compare three methods of parentage analysis and showed that the number and diversity of microsatellite loci were the most important factors defining the accuracy of assignments. Our simulations revealed that an exclusion-Bayes theorem method was more susceptible to false positive and false negative assignments than other methods tested. Here, we analyse and discuss the trade-off between type I and type II errors in parentage analyses. We show that controlling for false positive assignments, without reporting type II errors, can be misleading. Our findings illustrate the need to estimate and report both the rate of false positive and false negative assignments in parentage analyses.

The objective of parentage analyses can vary depending on the nature of the study, though a common goal is to correctly assign each and every offspring from a population to its true mother and/or father (Jones and Arden 2003; Blouin 2003; Jones *et al.* 2010). If not all putative parents have been sampled, correct assignments and correct exclusions must be distinguished from false assignments (false positive – type I error) and false exclusions (false negative – type II error). In Harrison *et al.* (2013a), we carried out simulations to assess how the number and allelic diversity of microsatellite loci, the proportion of candidate parents sampled, and genotyping error could affect the susceptibility of different methods of parentage analysis to type I and type II errors. We showed that the number and diversity of loci were the most important factors defining the accuracy of parentage analyses. We found that full- and pairwise-likelihood methods were systematically better at minimising type I and type II errors than an exclusion-Bayes theorem approach, though all methods could accurately distinguish correct assignments and correct exclusions with 20 highly diverse loci.

In his comment, Christie (2013) cautions that an error using the exclusion-Bayes' theorem approach (Christie *et al.* 2010) led us to wrongly conclude that this method could not control the rate of false positive assignments. However, minimising *only* false positives assignments was not the objective of our study and to do so neglects other decision types of single parent assignment tests (Harrison *et al.* 2013a). We defined accuracy as the ability to distinguish correct assignments and correct exclusions from type I and type II errors; a metric that takes into account all possible decision types in parentage

analyses (Harrison *et al.* 2013a) and is the most relevant to comparative studies. We accept that applying a maximum posterior probability of assignment (*alpha*) prior to accepting putative parent-offspring pairs, as Christie (2013) has done, can control the number of false positive assignments, and that for many purposes this may be desirable. However, minimising the rate of false assignments affects the rate of false exclusions, a trade-off that is contingent on the different objectives of parentage studies. For instance**,** if the alternative goal is to maximise the number of true parent-offspring pairs that are assigned, setting *alpha* too low may inadvertently reject a large number of correct parent-offspring relationships.

To fully evaluate the effects of fixing *alpha* at different arbitrary levels, we reran all 60 simulated scenarios (Harrison *et al.* 2013b) accepting either all putative parent-offspring pairs (*alpha* = 1) or only pairs with a probability of being false below 0.01 and 0.05, and analysed the effects of such measures on the accuracy of assignments. Using the same N1000 high diversity data set with 1% genotyping error as presented in Harrison *et al.* (2013a, b), we assessed the performance of each method depending on three potential objectives of parentage analysis: **1.** Maximise the proportion of assignments that are correct. **2.** Maximise the number of true parent-offspring pairs that are identified. **3.** Obtain an accurate estimate of the proportion of true parent-offspring pairs that are present in the sample.

Fixing *alpha* at 0.05 or 0.01 did not improve the overall accuracy of the exclusion-Bayes method in our simulated scenarios unless the proportion of

candidate parents was low (Fig. 1). Across all simulated scenarios, a cut-off value of 1, as in Harrison *et al.* (2013a), resulted in an overall accuracy of 0.653 ± 0.283, whereas cut-off values of 0.05 and 0.01 resulted in an overall accuracy of 0.650 ± 0.301 and 0.599 ± 0.305, respectively. Here, reducing *alpha* results in an explicit trade-off where the decrease in type Ia and type Ib errors (falsely assigning parentage when the true parent is or isn't present in the sample of candidate parents) is outweighed by the increase in type II errors (Fig. S1-S3). Even when using this trade-off to control the rate of false positive assignments, the exclusion-Bayes method appears to be comparatively less effective at distinguishing between true and false parent-offspring pairs than either the pairwise likelihood approach implemented in FAMOZ (Gerber *et al.* 2003) or the full-likelihood approach implemented in COLONY (Wang *et al.* 2004; Jones & Wang 2010).

In some circumstances, the trade-off between type I and type II errors can be adjusted to meet specific objectives of parentage studies. For example, if the aim is to maximise the proportion of assignments that are correct (Fig. 2; **Objective 1**), using the exclusion-Bayes method with an stringent cut-off value (*alpha* = 0.01) to minimise type Ia and type Ib errors does appear to perform well compared to other methods, especially when the proportion of sampled parents and the number of loci are low. However, even in scenarios where the proportion of correct assignments equals that of FAMOZ or COLONY, it identifies comparatively fewer assignments (Fig. 3). Alternatively, if the aim is to maximise the number of true parent-offspring pairs that are identified (Fig. 2; **Objective 2**), both type Ia (falsely assigning to a parent when the true parent was in the

sample) and type II errors must be minimised. In this situation, the exclusion-Bayes method improves by allowing all putative parent-offspring pairs to be assigned (*alpha* = 1.0; Fig. 2-3). If the aim is to obtain an accurate estimate of the proportion of true parent-offspring pairs that are present in the sample (Fig. 2; **Objective 3**), the primary objective is to balance type Ib errors (falsely assigning to a parent when the true parent was not in the sample) and type II errors. The number of true parent-offspring pairs present in the sample is correctly estimated when the number of type Ib equals the number of type II error. In this case, minimising type I errors without controlling type II errors underestimates the number of true parent-offspring pairs in the sample by a factor of 2 to 4. Regardless of the objective, increasing the number or allelic diversity of loci is the most effective way to reduce both type I and type II errors (Fig 2-3, Harrison *et al.* 2013a) and increase the performance of parentage analyses. Simulations, with known parent-offspring pairs, are integral to estimating errors rates and therefore optimising the performance of parentage analyses.

The methods described by Christie *et al.* (2010) and implemented in SOLOMON (Christie *et al.* 2013) do appear well suited where marker information is scarce and where avoiding false assignments is a priority. Rejecting putative parent-offspring above a certain threshold *alpha* did not improve the overall accuracy of the exclusion-Bayes method, though it did improve its performance when the objective was to maximise the proportion assignments that were correct. This however, is not a distinct advantage over other methods such as FAMOZ or CERVUS that employ likelihood estimators (Gerber *et al.* 2003; Marshall *et al.* 1998; Kalinowski *et al.* 2007). These methods identify a threshold of

assignment based on the distributions of likelihood scores for simulated true and false parent-offspring pairs. If the distributions overlap, the threshold value is usually set at the intersection of the two distributions in order to minimise both type I and type II errors, or can be set higher (e.g. a value that is equal or higher than 95% or 99% of all simulated false pairs LOD scores) or lower in order to minimise type I or type II errors, respectively.

Clearly there can be different objectives of parentage analysis that may favour minimising false positives, false negatives or maximising overall accuracy. In some circumstances, where the cost of false positive assignments is too high, minimising type I errors to ensure that all assignments are correct may be necessary. In other cases, minimising type II to ensure that all true parent pairs are identified may be more important. In our studies, where we have used parentage analysis to examine patterns of juvenile recruitment and the reproductive success of adults in fishes (Jones *et al.* 2005; Planes *et al.* 2009; Saenz-Agudelo *et al.* 2011; Harrison et al. 2012; Berumen *et al.* 2012; Almany *et al.* 2013), we consider that minimising both type I and type II errors will provide the best estimate of these parameters. Whatever the goal or the method used, type I and type II errors should always be estimated and reported. Fixing *alpha* at the expense of type II errors, and then only reporting type I errors can be misleading and may result in false depiction of accuracy and inaccurate estimates population parameters that rely on parentage. Lastly, increasing the quantity and quality of marker information reduces both false positive and false negative assignments, which can only improve the outcome of parentage studies. We concur that in the future, with next-generation techniques for sequencing

large numbers of markers, all methods will be able to be applied with extremely high accuracy, and arguments about the relative merits of trading false positive and false negative assignments will be of marginal concern.

References

Almany, G. R., Hamilton, R. J., Bode, M., *et al.* (2013). Dispersal of grouper larvae drives local resource sharing in a coral reef fishery. *Current Biology*, **23**, 626–630.

Berumen ML, Almany GR, Planes S, Jones GP, Saenz-Agudelo P, Thorrold SR (2012) Persistence of self-recruitment and patterns of larval connectivity in a marine protected area network. *Ecology and Evolution*, **2**, 444-452.

Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology and Evolution*, **18**, 503–511.

Christie MR (2010) Parentage in natural populations: novel methods to detect parent-offspring pairs in large data sets. *Molecular Ecology Resources*, **10**, 115–128.

Christie MR (2013) Bayesian parentage analysis reliably controls the number of false assignments in natural populations. *Molecular Ecology*, in press.

Gerber S, Chabrier P, Kremer A (2003) FAMOZ: a software for parentage analysis using dominant, codominant and uniparentally inherited markers. *Molecular Ecology Notes*, **3**, 479–481.

Harrison HB, Saenz-Agudelo P, Planes S, Jones GP, Berumen ML (2013a) Relative accuracy of three common methods of parentage analysis in natural populations. *Molecular Ecology*, **22**, 1158-1170.

188 Harrison HB, Saenz-Agudelo S, Planes S, Jones GP, Berumen ML (2013b) Data
189 from: relative accuracy of three common methods of parentage analysis in
190 natural populations. Dryad Digital Repository.
191 http://dx.doi.org/10.5061/dryad.2ht96
192 Jones AG, Small CM, Paczolt KA, Ratterman NL (2010) A practical guide to
193 methods of parentage analysis. *Molecular Ecology Resources*, **10**, 6–30.
194 Jones AG, Ardren WR (2003) Methods of parentage analysis in natural
195 populations. *Molecular Ecology*, **12**, 2511–2523.
196 Jones GP, Planes S, Thorrold SR (2005) Coral reef fish larvae settle close to home.
197 *Current Biology* **15**, 1314-1318.
198 Jones OR, Wang J (2010) COLONY: a program for parentage and sibship inference
199 from multilocus genotype data. *Molecular Ecology Resources*, **10**, 551–555.
200 Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer
201 program CERVUS accommodates genotyping error increases success in
202 paternity assignment. *Molecular Ecology*, **16**, 1099–1106.
203 Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for
204 likelihood-based paternity inference in natural populations. *Molecular*
205 *Ecology*, **7**, 639–655.
206 Planes S, Jones GP, Thorrold SR (2009) Larval dispersal connects fish
207 populations in a network of marine protected areas. *Proceedings of the*
208 *National Academy of Sciences of the United States of America*, **106**, 5693-
209 5697.
210 Saenz-Agudelo P, Jones GP, Thorrold SR, Planes S (2011) Connectivity dominates
211 larval replenishment in a coastal reef fish metapopulation. *Proceedings of*
212 *the Royal Society B Biological Sciences*, **278**, 2954–2961.

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

215

216 **Supporting information**

217 Additional supporting information may be found in the online version of this
218 article.

219 Defining and measuring the performance of parentage analyses.

220 **Fig. S1** Susceptibility of three methods of parentage analysis to type Ia errors
221 under 60 independent scenarios.

222 **Fig. S2** Susceptibility of three methods of parentage analysis to type Ib errors
223 under 60 independent scenarios.

224 **Fig. S3** Susceptibility of three methods of parentage analysis to type II errors
225 under 60 independent scenarios.

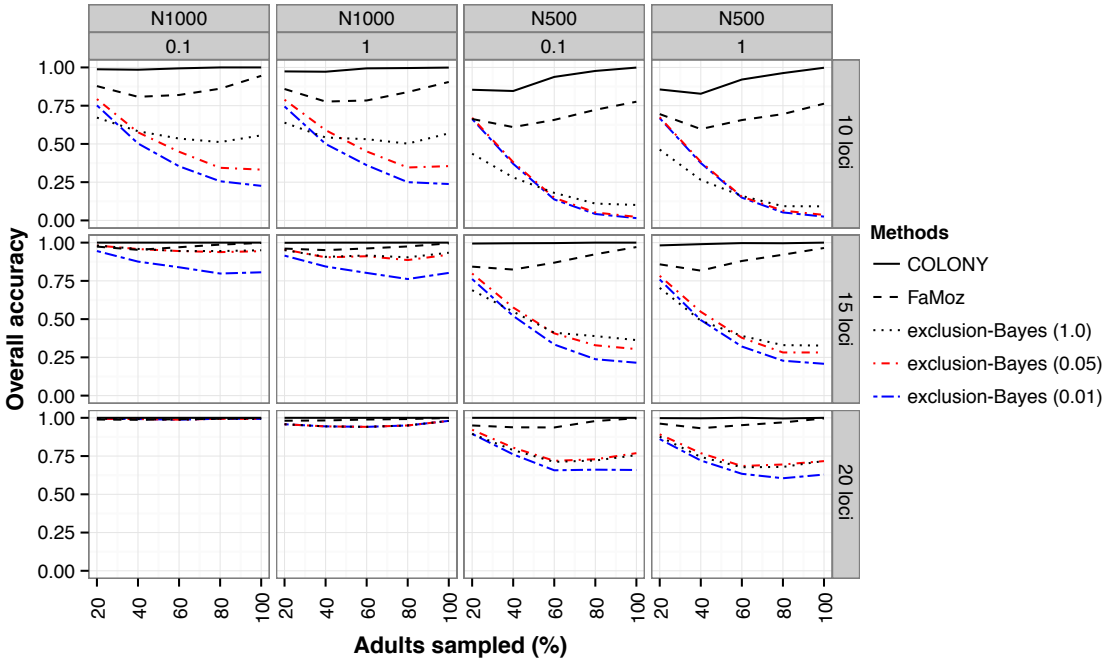
226

227 **Data accessibility**

228 Simulated data sets and R scripts deposited in the Dryad Digital Repository:
229 http://dx.doi.org/10.5061/dryad.2ht96.
230

231

232

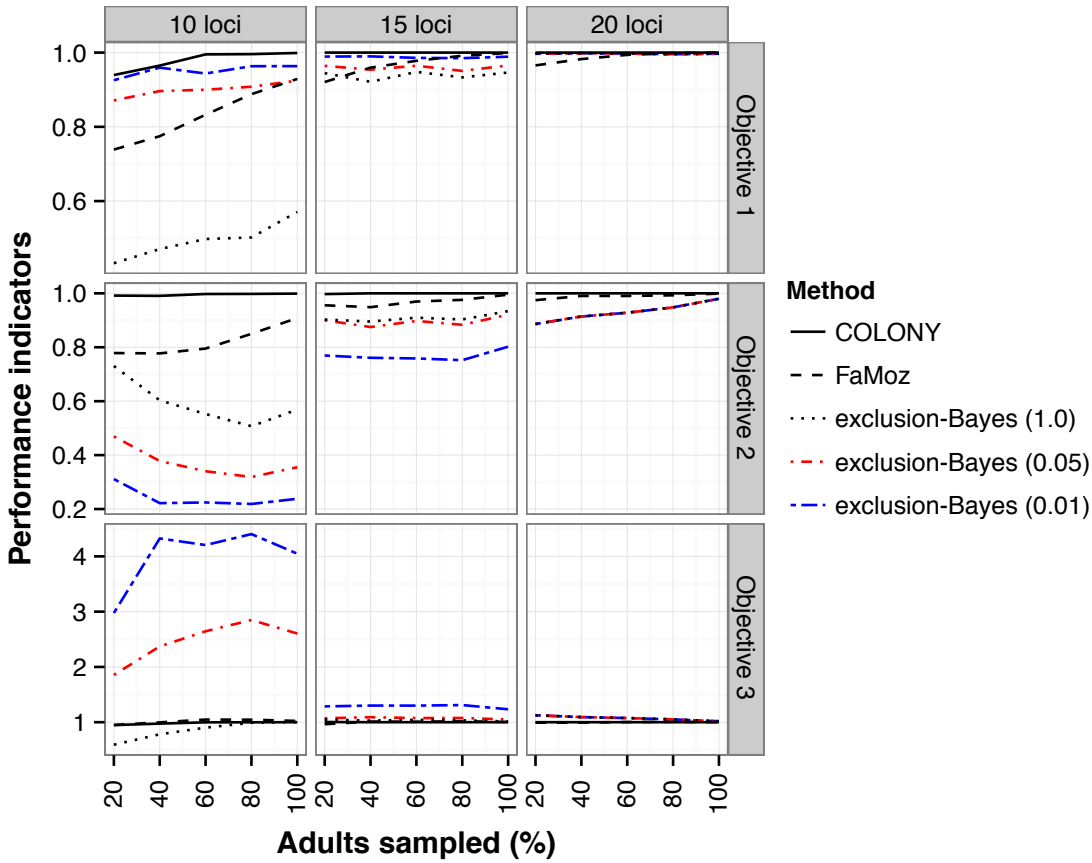


233

234 **Fig. 1** Proportion of accurate assignments of three approaches to parentage
235 analysis. Each methods was tested on high- and low-diversity simulated
236 microsatellite data sets with high (1%) and low (0.1%) levels of genotyping error
237 for varying levels of number of loci and proportion of candidate parents samples.
238 Continuous black lines correspond to results from the full-likelihood method
239 implemented in COLONY, dashed black lines are the results from the pairwise-
240 likelihood method implemented in FAMOZ and dotted black lines from the
241 exclusion-Bayes method using a cut-off value of 1.0 as presented in Harrison *et*
242 *al.* (2013). Blue and red dot-dash lines correspond to results from the exclusion-
243 Bayes method using cut-off values of 0.05 and 0.01, respectively. A value of 1.0
244 represents the optimal performance in each panel.

245

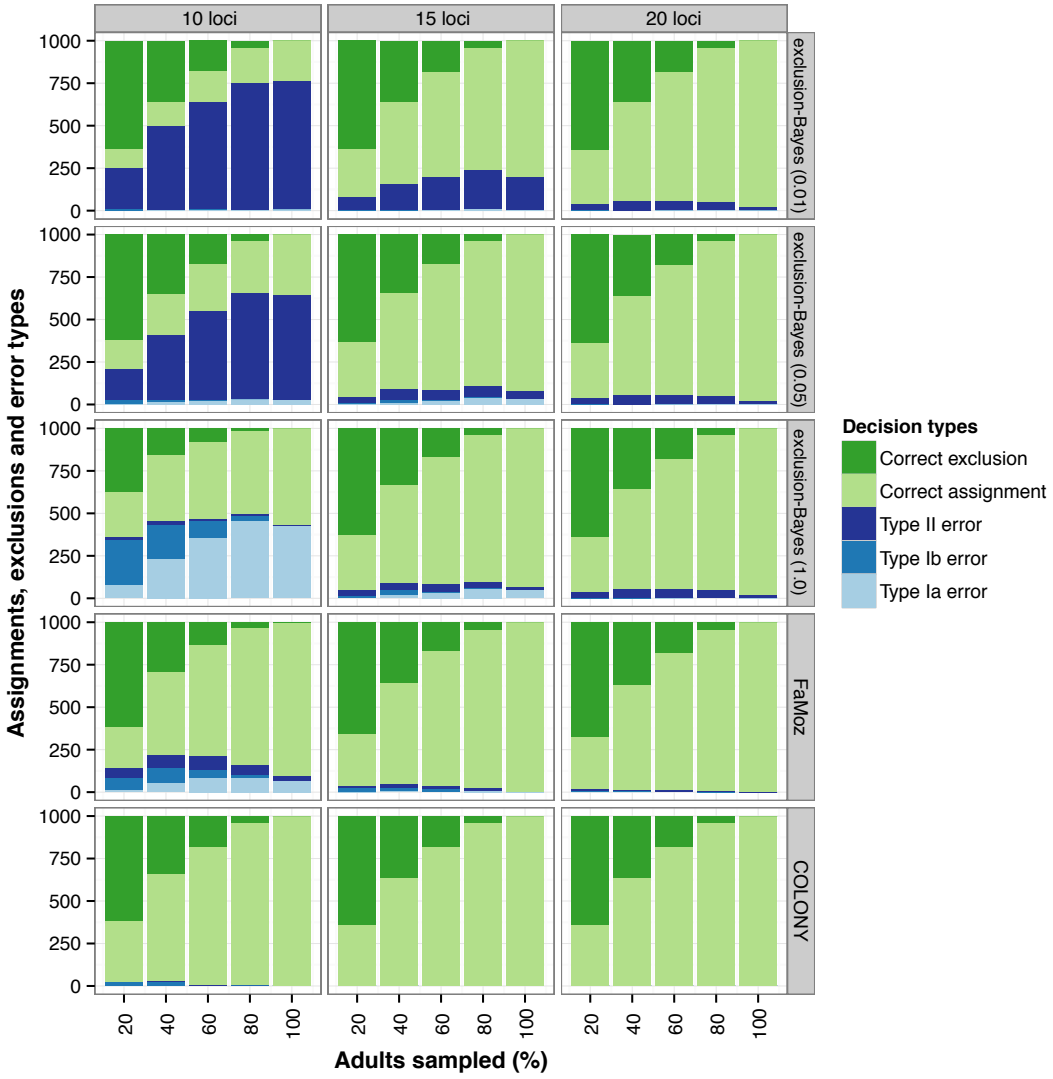
246



247

248 **Fig. 2** Performance of three methods of parentage analysis under study-specific
249 objectives. Each method was assessed using the N1000 high-diversity dataset
250 with 1% genotyping error as described in Harrison et al. (2013). The specific
251 objectives are 1) Maximising the proportion of assignments that are correct; 2)
252 Maximising the number of true parent-offspring pairs that are identified; and 3)
253 Obtaining an accurate estimate of the proportion of true parent-offspring pairs
254 that are present in the sample (see Supplementary Material for a description of
255 each performance indicator). Line representations are identical to Fig. 1. A value
256 of 1.0 represents the optimal performance in each panel.

257



258

259 **Fig. 3** Number of correct assignment, correct exclusions, false positive (Type Ia

260 and Type Ib), and false negative (Type II) assignments in the analysis of the

261 N1000, high-diversity dataset with 1% genotyping error as described in

262 Harrison *et al.* (2013). We used the exclusion-Bayes method with three different

263 cut-off values (*alpha* = 0.01, 0.05 and 1.0) and present results from FAMOZ and

264 COLONY as they were presented in Harrison *et al.* (2013).

265

Supplementary information

Defining and measuring the performance of parentage analyses

Accuracy: The accuracy of a parentage analysis is define here and in Harrison et al. (2013a) as the degree to which all relationships can be correctly resolved, whether it is assigning true parent-offspring pairs or excluding false parent-offspring pairs. Accuracy is measured as the sum of correct assignments and correct exclusion over the total number of possible assignments, which is the total number of offspring in the sample. Maximising accuracy can itself, be a potential objective of parentage analyses. Given that it takes into account of both false positive and false negative assignments it is also a valuable metric for comparisons.

$$Accuracy = \frac{No.\,of\ correct\ assignments + No.\,correct\ exclusions}{No.\,of\ offspring\ in\ the\ sample}$$

Objective 1: Maximising the proportion of assignments that are correct. Performance increases as the proportion of correct assignments approaches the number of assigned parent-offspring pairs (range 0-1).

$$P_1 = \frac{No.\,of\ correct\ assignments}{No.\,of\ assignments}$$

Objective 2: Maximising the number of true parent-offspring pairs that are assigned. Performance increases as the proportion of correct assignments approaches the number of true parent-offspring pairs present in the sample (range 0-1).

$$P_2 = \frac{No.\,of\ correct\ assignments}{No.\,of\ true\ parent-offspring\ pairs\ in\ the\ sample}$$

Objective 3: Obtaining a representative proportion of true parent-offspring pairs that are present in the sample. Performance increases as the number of assignments approaches the number of true parent-offspring pairs present in the sample (range 0-∞). A value <1, overestimates the number of true parent-offspring pairs in the sample and a value >1, underestimates the number of true parent-offspring pairs in the sample.

$$P_3 = \frac{No.\,of\ true\ parent-offspring\ pairs\ in\ the\ sample}{No.\,of\ assignments}$$

Supplementary figures

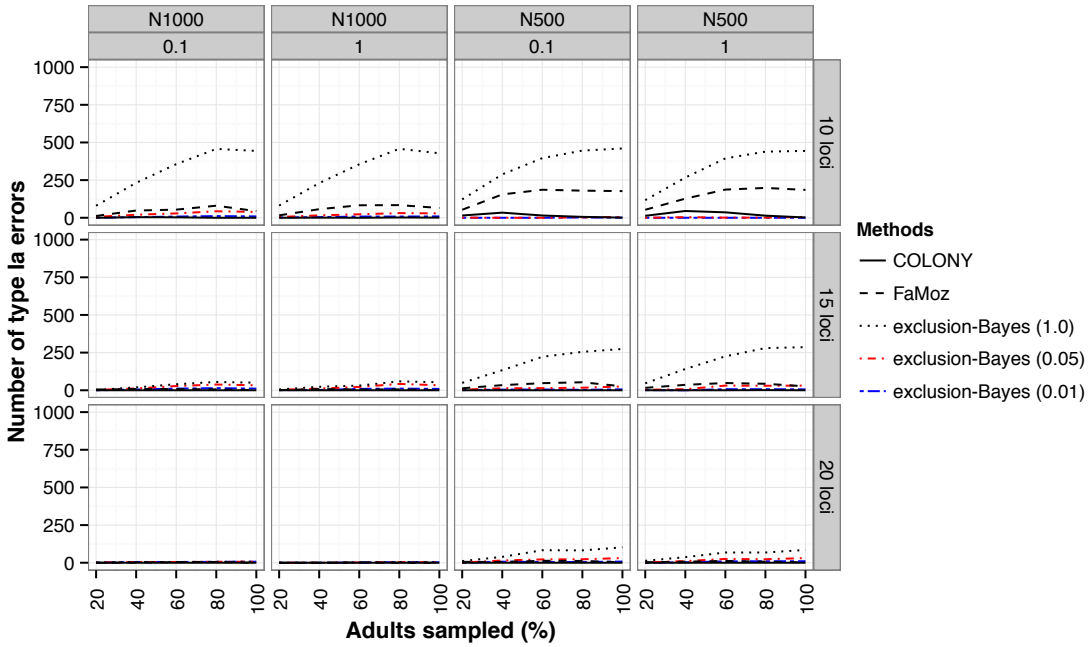
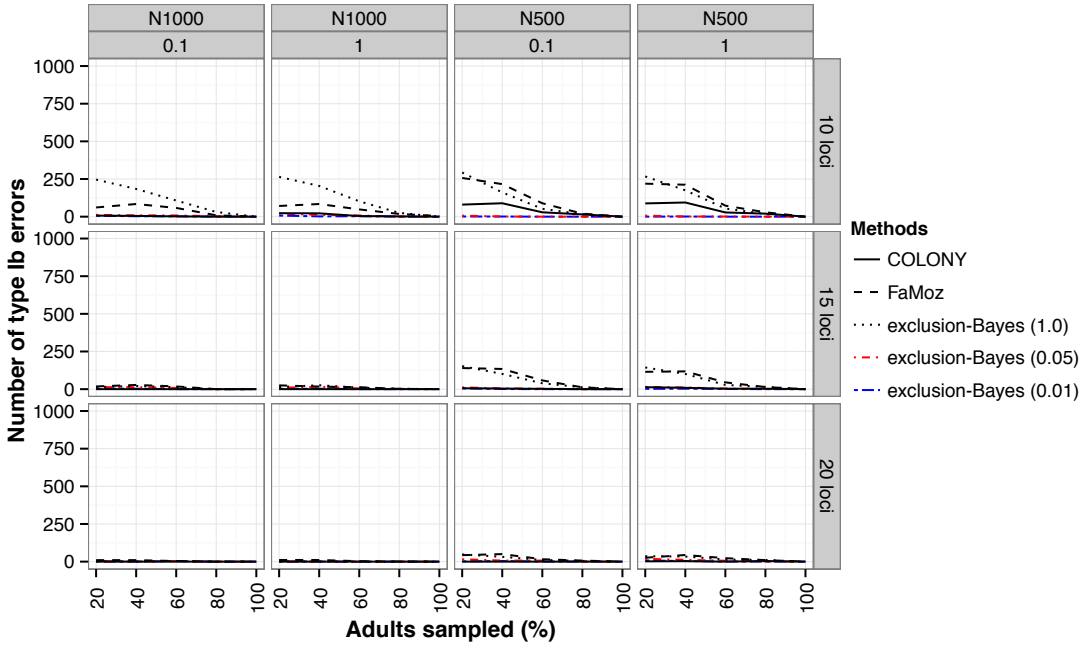


Fig. S1 Susceptibility of three methods of parentage analysis to type Ia errors under 60 independent scenarios. Each methods was tested on high- and low-diversity simulated microsatellite data sets with high (1%) and low (0.1%) levels of genotyping error for varying levels of number of loci and proportion of candidate parents samples. Continuous black lines correspond to results from the full-likelihood method implemented in COLONY, dashed black lines are the results from the pairwise-likelihood method implemented in FAMOZ and dotted black lines from the exclusion-Bayes method using a cut-off value of 1.0 as presented in Harrison *et al.* (2013). Blue and red dot-dash lines correspond to results from the exclusion-Bayes method using cut-off values of 0.05 and 0.01, respectively.

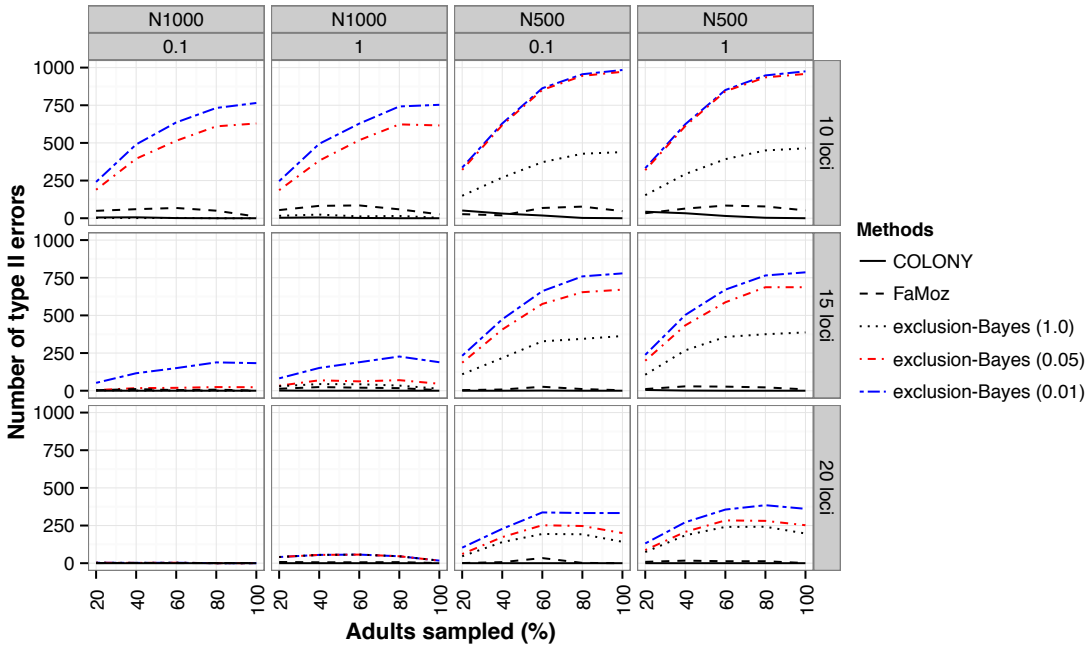


316

317 **Fig. S2** Susceptibility of three methods of parentage analysis to type Ib errors

318 under 60 independent scenarios. Data and line representations are identical to

319 Fig. S1.



320

321 **Fig. S3** Susceptibility of three methods of parentage analysis to type II errors

322 under 60 independent scenarios. Data and line representations are identical to

323 Fig. S1.