Molecular Ecology Resources
March 2020, Volume 20, Issue 2, Pages 579-590
https://doi.org/10.1111/1755-0998.13103
https://archimer.ifremer.fr/doc/00587/69866/

APIS: An Auto- Adaptive Parentage Inference Software that tolerates missing parents

Griot Ronan [1, 2], Allal Francois [3], Brard-fudulea S [1], Morvezen R [1], Haffray P [1], Phocas F [2], Vandeputte Marc [2, 3, *]

[1] SYSAAF, Station LPGP/INRA Campus de BeaulieuRennes , France
[2] GABI, INRA, AgroParisTech, Université Paris-Saclay , France
[3] MARBEC, Univ. Montpellier, Ifremer, CNRS, IRD Palavas-les-Flots , France

* Corresponding author : Marc Vandeputte, email address : marc.vandeputte@inra.fr

Abstract :

In the context of parentage assignment using genomic markers, key issues are genotyping errors and an absence of parent genotypes because of sampling, traceability or genotyping problems. Most likelihood-based parentage assignment software programs require a priori estimates of genotyping errors and the proportion of missing parents to set up meaningful assignment decision rules. We present here the R package APIS, which can assign offspring to their parents without any prior information other than the offspring and parental genotypes, and a user-defined, acceptable error rate among assigned offspring. Assignment decision rules use the distributions of average Mendelian transmission probabilities, which enable estimates of the proportion of offspring with missing parental genotypes. APIS has been compared to other software (CERVUS, VITASSIGN) on a real European seabass (Dicentrarchus labrax) SNP data set. The type I error rate (false positives) was lower with APIS than with other software, especially when parental genotypes were missing, but the true positive rate was also lower, except when the theoretical exclusion power reached 0.99999. In general, APIS provided assignments that satisfied the user-set acceptable error rate of 1% or 5%, even when tested on simulated data with high genotyping error rates (1% or 3%) and up to 50% missing sires. Because it uses the observed distribution of Mendelian transmission probabilities, APIS is best suited to assigning parentage when numerous offspring (>200) are genotyped. We have demonstrated that APIS is an easy-to-use and reliable software for parentage assignment, even when up to 50% of sires are missing.

Keywords : Parentage assignment, pedigree, SNP, microsatellites, missing parents

Introduction

Pedigree information has numerous applications, which range from selective breeding (Durel, Laurens, Fouillet, & Lespinasse, 1998; Misztal, Lawlor, Short, & VanRaden, 1992) to ecological and evolutionary studies (Foerster et al., 2007; Wilson et al., 2010). In the selective breeding of terrestrial livestock, parental information is usually recorded through individual tagging at birth. However, there are many situations where pedigree cannot be maintained with such physical tagging, either because reproduction or birth cannot be observed (as is the case in natural populations; Kruuk, 2004), or because offspring size at birth is too small for them to be tagged. This may be the case in natural populations, or in aquaculture in the case of batch spawning species or when multiple families are mixed at spawning to avoid the confusion of genetic and common environmental effects (Vandeputte & Haffray, 2014). In all these situations, an alternative method is required to identify the parents of a given offspring.

Parentage assignment using genetic markers offers an opportunity to retrieve pedigree *a posteriori* using genotypes from both the parental and offspring generations. Microsatellite markers have been widely used for this purpose (Pemberton, Slate, Bancroft, & Barrett, 1995) thanks to their high heterozygosity and number of alleles (De Woody, 2005). Depending on their variability in the population studied, highly reliable assignment can be achieved using less than 10 to in some cases 50 microsatellite markers (Glaubitz, Rhodes, & Dewoody, 2003).

More recently, the most widely used genetic markers have been single nucleotide polymorphism (SNP) markers (Morin, Luikart, Wayne, & the SNP Workshop Group, 2004; Vignal, Milan, SanCristobal, & Eggen, 2002). Thousands of SNPs can easily be generated in any species using high throughput sequencing techniques (Baird et al., 2008). Although most SNPs are only bi-allelic, and thus much less variable than microsatellites, they are widely used in parentage assignment (Anderson & Garza, 2005; Glaubitz et al., 2003). The principal reasons for this choice are their abundance throughout the genome, their lower genotyping cost and a lower genotyping error rate compared to microsatellites (Ranade et al., 2001). Because of their low variability, approximately six times more SNPs than microsatellites are required to ensure the efficiency of a particular parentage assignment (Glaubitz et al., 2003).

Several methods have been developed to assign parentage using marker genotypes from offspring and their potential parents (see the review by Jones, Small, Paczolt, & Ratterman, 2010). These methods can be divided into two main categories: exclusion and likelihood, the former being the simplest. According to Mendelian inheritance rules, the aim is to exclude all potential parent pairs of an offspring other than the true pair. In practice, all potential parent pairs are tested for Mendelian incompatibilities with the offspring for all available markers. With a sufficient number of markers, only one parent pair (the true one) should remain compatible with the tested offspring. If the set of markers is not sufficiently informative, multiple pairs will remain compatible with the offspring, leading to multiple (thus unresolved) assignments (Jones et al., 2010).

The likelihood method was developed to manage most situations where the exclusion method fails (Chakraborty, Meagher, & Smouse, 1988). With this method, the estimated likelihood of a parent pair being the true parents of a given offspring is based on the genotypes over all available markers (Marshall, Slate, Kruuk, & Pemberton, 1998). Likelihood thresholds are then estimated to infer whether the most likely parent pair is the true pair. Most software programs perform simulations in order to define these threshold values (Jones et al., 2010). A set of parents is generated from the parents present in the dataset or from allelic frequencies estimated from the parental genotypes. Offspring are then simulated based on the genotypes of the parents, according to Mendelian inheritance rules. The difference in likelihood (delta) between the most likely and second most likely parent pair is calculated, and its distribution between simulated individuals assigned to their true

parents and simulated individuals assigned to an incorrect parent pair is used to define a threshold value, usually at the 95% confidence interval, established from the delta values calculated during the simulation process. If an offspring's delta is above the threshold, the offspring is assigned to its most likely parent pair. If the offspring's delta is lower than the threshold, the offspring remains unassigned (Boichard, Barbotte, & Genestout, 2014; Kalinowski, Taper, & Marshall, 2007).

Beyond the practical considerations of cost and tissue collection, parentage assignment needs to deal with three main technical issues: the lack of power of the marker set, genotyping errors and missing parents, i.e. parents for whom no genotype is available (Jones et al., 2010). Regarding the first issue, the theoretical power of a marker set can be estimated for assignment by exclusion, as proposed by Vandeputte (2012), combining allelic frequencies in the parental population and the size of the assignment problem to be solved (number of potential male and female parents). When the power of the marker set is too low, multiple assignments will be seen under an exclusion approach, when several parent pairs remain compatible with a given offspring. In such cases, likelihood is more efficient because the most likely parent pair among those that are compatible can be identified in the majority of cases (Anderson & Garza, 2005). A likelihood threshold value must be set to minimize incorrect assignment without excessively reducing the assignment rate. Both the exclusion and likelihood methods also need to manage genotyping errors, which is achieved by allowing for some Mendelian incompatibilities. With exclusion, a certain number of mismatches is permitted empirically (Vandeputte, Mauger, & Dupont-Nivet, 2006). Under the likelihood approach, the likelihood value of a marker with Mendelian incompatibility is required as an input in the simulation process (Marshall et al., 1998) and generally needs to be small but greater than zero. A zero value will lower the total likelihood to zero, even with a single mismatch (Sancristobal & Chevalet, 1997). Offspring from "missing parents" (i.e. parents with a missing genotype) should not be assigned if the marker set is sufficiently informative, or a wrong assignment (type I error) may occur. Using likelihood approaches, an estimate of the proportion of missing parents is required for the simulation process to accurately establish the threshold values (Marshall et al., 1998). However, producing such an estimate requires good *a priori* knowledge of the population being studied. Any shift from the real values can lead to inconsistent and unreliable results (Jones et al., 2010).

We have developed a new parentage assignment method for the efficient assignment of offspring to their parents while minimizing the positive assignment error rate (the ratio of false positives to all positives) to a predefined level, set by the user. We designed this method in order to 1) be efficient even when some parents are missing, 2) be simple to use with no *a priori* information required regarding the genotyping error rate (fixed at 1%) or the number of missing parents, thus being adaptable to the characteristics of the population and the marker set. Using the observed distributions of average Mendelian transmission probability (Boichard et al., 2014) in the offspring population to be assigned, we can estimate the number of offspring for which at least one parental genotype is missing, and use this information to set robust parentage assignment rules. This method is implemented using a new software program (APIS). We compared its performance to that of existing programs: CERVUS (Kalinowski et al., 2007) as a widely used likelihood-based software and VITASSIGN (Vandeputte et al., 2006) as an exclusion-based software. We also evaluated the sensitivity of APIS to the size of the offspring population used to set the parentage assignment rules.

Materials and Methods

The APIS concept

For a given offspring and potential parent pair, APIS calculates the Mendelian transmission probability for each marker, defined as the probability of obtaining the genotype of the offspring given the genotype of the parents. These probabilities are averaged over all available markers to produce an average Mendelian transmission probability that is characteristic of the parent pair relative to a specific offspring (Boichard et al., 2014). This average Mendelian transmission probability is calculated for all potential parent pairs. The Mendelian transmission probability threshold used to declare an offspring as being assigned is determined directly from the observed distributions of Mendelian transmission probabilities in the data analyzed, which requires an offspring dataset of sufficient size to obtain meaningful distributions. The proportion of offspring with at least one missing parent is also estimated from the distributions of Mendelian transmission probabilities. The only input parameter required from the users is the positive assignment error rate (*e=false positives/all positives*) that they would accept among individuals assigned to their parents by APIS. This parameter is used to modulate the value of the Mendelian transmission probability threshold. In

the context of this study we only used SNP markers, but APIS can also handle other codominant markers, including microsatellites.

Statistical methods

For each offspring *o* and each potential parental pair (sire *s*, dam *d*), the average Mendelian probability $P_m(o,s,d)$ was calculated as the average over all genotyped loci (n) of the locus-specific (l) Mendelian transmission probability $p_m(o,s,d,l)$ for a given offspring-sire-dam trio:

$$P_m(o,s,d) = \frac{1}{n}\sum_{l=1}^{n} p_m(o,s,d,l)$$

The locus-specific Mendelian transmission probability $p_m(o,s,d,l)$ was taken from Table 1 and Table 2 as proposed by Boichard et al. (2014). Note that $P_m(o,s,d) = e^{\log(V)/n}$, where V is the log-likelihood for the same offspring-sire-dam trio.

[insert Table 1]

[insert Table 2]

We considered that the offspring population is representative of any reproductive events that have happened between the parents. Thus, when the genotype of a parent is missing at a specific locus, a reasonable estimate of the probability of allele A at this locus is the frequency of A in the offspring population, which takes account of the relative reproductive success of the potential parents. This differs from the approach adopted by Boichard et al. (2014), who used allelic frequencies in the parental population. This is reflected in the Mendelian probabilities estimated for unknown parental genotypes in Tables 1 and 2.

For a given offspring, all potential parent pairs are ordered by decreasing $P_m(o,s,d)$ and noted $PPi(o)$, where *i* is the rank of the Mendelian transmission probability of the parent pair among all possible pairs. The Mendelian transmission probability of $PP1(o)$ is noted $P1_m(o)$, that of $PP2(o)$ is noted $P2_m(o)$ and that of $PP3(o)$ is noted $P3_m(o)$. The difference between successive ordered $Pi_m(o)$ is $\Delta_i(o) = Pi_m(o) - P(i+1)_m(o)$. As APIS uses the observed distributions of $Pi_m(o)$, we defined $Pi_m(:)$ as the set of all $Pi_m(o)$ among all the offspring *o* in the dataset and $\Delta_i(:)$ as the set of all $\Delta_i(o)$. The distributions of $P1_m(:)$, $P2_m(:)$ and $P3_m(:)$, shown in Figure 1, contain information on potential missing parents (i.e. parents with missing genotypes at all loci): when the parents of all offspring are

present in the dataset, the distribution $P1_m(:)$ has a single peak (Figure 1a), but when some parents are missing, the distribution of $P1_m(:)$ has two peaks, representing two distributions (Figure 1b). The first peak (with the highest Mendelian probability value) represents the distribution of the Mendelian probabilities of the best parent pairs for offspring which have both parents genotyped in the dataset, while the second peak represents that of offspring with at least one missing parent. The distribution of $P1_m(:)$ for offspring with at least one missing parent is indeed similar to the distribution of $P2_m(:)$ for offspring which have both parents genotyped (Figure 1b).

[insert Figure 1]

The distributions of $P1_m(:)$ and $P2_m(:)$ are also highly sensitive to the power of the marker set. When the number of markers (and thus the assignment power of the marker set) is reduced, the distributions of $P1_m(:)$ and $P2_m(:)$ tend to overlap (Figure 2). Therefore, when all parents are present in the dataset, the difference between $P1_m(:)$ and $P2_m(:)$, as represented by $\Delta_1(:)$, is directly related to the power of the marker set (Figure 2). When all parents are present in the dataset, the higher becomes the power of the marker set, and the smaller the overlap of the distributions of $P1_m(:)$ and $P2_m(:)$.

By construction, $P1_m(o) \geq P2_m(o) \geq P3_m(o) \geq \ldots \geq Pn_m(o)$. Parent pair $PP1(o)$, corresponding to $P1_m(o)$, can either be the true parent pair (when both parents are genotyped and the power of the marker set is sufficiently high) or an incorrect pair (when one or both parents are missing, or the power of the marker set is low). When $PP1(o)$ is the true parent pair, $P1_m(o)$ is much greater than $P2_m(o)$. By contrast, when $PP1(o)$ is an incorrect parent pair, $P1_m(o)$ is only slightly greater than $P2_m(o)$. As a consequence, $\Delta_1(o)$ is generally high when the true parents are present and generally low when they are missing (Figure 3a and 3b, bottom panel). The values of $\Delta_i(o)$ when $i \geq 2$ are always low (Figure 3 for $i = 2$), irrespective of whether the parents are genotyped or not.

[insert Figure 2]

[insert Figure 3]

The first step to set up the Mendelian transmission probability thresholds is to estimate the number N_{miss} of offspring with at least one missing parent (i.e., a missing parent genotype at all loci). The distribution of $P1_m(:)$ for progeny with missing parents is expected to be similar to the distribution of

$P2_m(:)$ for progeny with both parents genotyped, and centered on the same median value. Thus, N_{miss} corresponds to the number of individuals within the second peak of the distribution of $P1_m(:)$.

However, as shown in Figure 2, a lack of power in the marker set also causes some overlapping of the distributions of $P1_m(:)$ and $P2_m(:)$. Thus, $P1_m(:)$ values that are higher than the median of $P2_m(:)$ may either correspond to individuals with at least one missing parent or to individuals for which both parents are genotyped. Nevertheless, it can be seen from Figure 2 that even with a low power marker set (35 markers), $P1_m(:)$ values that are lower than the median of $P2_m(:)$ are very unlikely to be values of $P1_m(:)$ from individuals with both parents genotyped, and thus come mostly from individuals with at least one missing parent. Under the assumption that the distribution of $P1_m(:)$ for progeny with at least one missing parent is symmetrical, N_{miss} can then be estimated as twice the number of offspring for which $P1_m(o)$ is lower than the median of $P2_m(:)$:

$$N_{miss} = 2 * \sum_{o=1}^{N} S_o \qquad where \begin{cases} S_o = 0 \; if \; P1_m(o) > median(P2_m(:)) \\ S_o = 1 \; if \; P1_m(o) \leq \text{median}(P2_m(:)) \end{cases}$$

where N is the total number of offspring.

If $\frac{N_{miss}}{N} \leq e$, then all the parent pairs with the highest Mendelian transmission probability are assigned as the parents of the progeny. In such a case, the estimated proportion of offspring with at least one missing parent is lower than the user accepted positive assignment error rate e, and choosing the parent pair with the highest Mendelian transmission probability offers the best solution to obtain a high assignment rate while respecting the accepted positive assignment error rate.

If $\frac{N_{miss}}{N} > e$, the assignment threshold is established as follows: $N - N_{miss}$ is the expected number of offspring with both parents genotyped. Use of the (1 – e) quantile of all the values of $\Delta_2(:)$ would control the type I error rate. However, as our aim was to control the positive assignment error rate (true positives/all positives), a more restrictive threshold was necessary, based only on offspring that could be assigned, or in other words, those with both parents genotyped. The $\Delta_2(:)$ value of such individuals was higher than those of offspring with at least one missing parent. (Supplementary Figure 1). Therefore, the highest $N - N_{miss}$ values of $\Delta_2(:)$ could be expected to be mostly those of offspring with both parents genotyped. The values of $\Delta_2(:)$ are then sorted in descending order, and the $(1 - e)$ quantile of the $N - N_{miss}$ first values is defined as the threshold value. All the offspring for which $\Delta_1(o)$ is higher than the threshold value are assigned to their $PP1_m(o)$. A consequence of using $\Delta_2(:)$ to set the assignment threshold is that the number of potential parents must be such that at least three parent pairs can be tested with each offspring.

APIS is not able to distinguish between individuals that are not assigned because of missing parents or a lack of marker power, as both missing parents and insufficient power cause an overlap between $\Delta_1(:)$ and $\Delta_2(:)$ distributions. Qualitatively, when the distribution of $\Delta_1(:)$ is composed of two separate sub-distributions, one of which is within the range of the $\Delta_2(:)$ distribution, (such as in Figure 3 with 20 missing sires and 200 markers), animals in the lower sub-distribution of $\Delta_1(:)$ are not assigned by APIS, because their parents are missing. However, when the power of the marker set is too low (Figure 3 with 35 markers), the distributions of $\Delta_1(:)$ and $\Delta_2(:)$ tend to overlap even without missing parents so that clearly distinguishing between individuals whose parents are not sampled and individuals with both parents genotyped but low $\Delta_1(:)$ is impossible.

Validation sets

A commercial seabass (*Dicentrarchus labrax*) cohort from the selected line of Ferme Marine du Douhet (FMD, la Brée les Bains, France) was genotyped on the 57K seabass SNP array DlabChip at the INRA genotyping platform GENTYANE (Clermont-Ferrand, France). The offspring cohort, comprising 1084 individuals, was obtained from 39 sires and 14 dams mated under a factorial design. After quality controls ($\mathrm{CallRate} \geq 0.9$ and $\mathrm{DQC} \geq 0.8$), 52813 SNPs were retained. From these we selected the 500 markers with the greatest minor allele frequency (MAF). This set of markers offered

an excess of information for assignment (exclusion power of 1 using the formula proposed by Vandeputte (2012)), so the pedigree obtained from it could be considered as the true pedigree. The 1084 individuals in the offspring set were assigned by exclusion using VITASSIGN (Vandeputte et al. 2006), allowing up to 25 mismatches. A total of 1068 offspring were thus assigned to a unique parent pair and constituted the reference pedigree.

Effects of the numbers of markers and missing parents

Within the genotypes of the individuals from the reference pedigree, four sets of markers (35, 42, 50 and 100 markers) were chosen at random from the 52,813 markers to reach theoretical exclusion powers (Vandeputte, 2012) of 0.90, 0.95, 0.99 and 0.99999, respectively. The corresponding datasets were then used to perform parentage assignment using APIS, CERVUS (Kalinowski et al., 2007) and VITASSIGN (Vandeputte et al., 2006). For APIS, the acceptable positive assignment error rate was set at 1% (APIS1) or 5% (APIS5). CERVUS was tested with default simulation parameters (10,000 offspring simulated with a genotyping error rate of 1% and a 95% confidence interval – CERVUS95). For VITASSIGN, we allowed for one mismatch (VITASSIGN1). The pedigree proposed by each software program was then compared with the reference pedigree.

We then tested the effects of missing parents on the results of parentage assignment. In order to test the most challenging situation, we only removed sire genotypes, as the risk of assignment error is higher when one true parental genotype is missing than when both are missing (Jamieson & C S Taylor, 1997). We created two series of datasets, one containing four missing sires (~10% missing sires) and one with 20 missing sires (~50%) chosen at random from the sires present in the reference pedigree. The sampling of missing sires was repeated ten times in both cases.

The datasets with missing sires were analyzed with the 35, 42, 50 and 100 markers chosen previously. The results obtained by each software program were then compared with the reference pedigree.

To make this comparison, we counted the number of offspring assigned to their true parents (N1), the number of offspring with both parents genotyped but assigned to an incorrect parent pair (N2), the number of offspring with both parents genotyped that were not assigned (N3), the number of offspring from missing sires assigned to an incorrect parent pair (N4) and the number of offspring from missing sires that were not assigned (N5).

The metrics used to compare the different programs were the percentages of true positives (N1/N), true negatives (N5/N), false positives or type I error ((N2 + N4)/N) and false negatives or type II error (N3/N).

Effect of the number of offspring

Because APIS uses the observed distributions of Mendelian transmission probabilities to set assignment thresholds, it could be expected that the genotypes of a minimum number of offspring are required to correctly describe these distributions. The effect of offspring number was tested by randomly choosing from 50 to 1000 offspring (by steps of 50) in twelve base datasets used for the software comparison (0, 4 and 20 missing sires with 35, 42, 50 and 100 markers). For each number of offspring tested, 100 repetitions (resampling of offspring) were performed for each base dataset, and the positive assignment error rate was calculated as the proportion of type I errors relative to positive results (assigned offspring).

$$positive\ assignment\ error\ rate = \frac{N2 + N4}{N2 + N4 + N1}$$

Effect of genotyping error rate

Using the simulated data described by Grashei, Ødegård, & Meuwissen (2018), six different datasets of 1000 salmon from 100 sires and 200 dams were chosen randomly from the 50 datasets simulated with 1% and 3% genotyping error rates. In each case, 120 markers were used to reach a theoretical assignment power of 0.99999 (Vandeputte, 2012). Each dataset was tested using APIS with an acceptable positive assignment error rate set at 1% or 5% (APIS1 and APIS5), using CERVUS with a 95% confidence interval (CERVUS95) and with VITASSIGN with 1% mismatches allowed (VITASSIGN1, two mismatches allowed as 120 markers were used). With APIS and CERVUS, when there was a 3% genotyping error in the dataset, we also tested setting the genotyping error parameter for each software program at 1% or 3%. For APIS, it was necessary to edit the code as the genotyping error is not a set user parameter and is normally fixed at 1%. Three levels of missing sires (0%, 10% and 50%) were tested, representing 0, 10 and 50 missing sires respectively.

Results

Assignment efficiency of APIS compared to other software programs

Case with no missing sires

When no parents were missing (Figure 4a), APIS and CERVUS were generally equivalent. They all produced about the same true positive and type I error rates with all marker sets. There was one exception, APIS1, which only gave 32.7% of true positives, and a 67.3% type II error when 35 markers were used. VITASSIGN1 had a high type II error rate, except when 100 markers were used (14.2%, 6.5% and 8.9% for 35, 42 and 50 markers, respectively). As the number of markers increased, the type I error rate decreased for APIS5, CERVUS95 and VITASSIGN1, from 3.5% for 35 markers to 0% for 100 markers.

Case with four missing sires

When four sires were missing (Figure 4b), APIS produced the lowest type I error rate (except when using 100 markers, where VITASSIGN1 performed equally or better), but also the lowest true positive rate. As the number of markers increased, the true positive rate of APIS increased (31.3%, 46.4%, 43.1% and 87.2% for 35, 42, 50 and 100 markers, respectively, with APIS1 and 58.8%, 73.0%, 69.5% and 89.0% for 35, 42, 50 and 100 markers, respectively, with APIS5). The type I error rate found with CERVUS95 was the highest (12.0%, 10.2%, 5.2% and 3.0% for 35, 42, 50 and 100 markers, respectively). The type II error was intermediate with VITASSIGN1 in most cases.

Case with 20 missing sires

When 20 sires were missing (Figure 4c), the behavior of all programs was comparable to the previous case with four missing sires. APIS had the lowest type I error rate except when 100 markers were used, but again generated the lowest true positive rate. CERVUS95 had the lowest true assignment rate (true positives + true negatives) as well as a very high type I error rate (35.1%, 28.4%, 20.2% and 11.4% for 35, 42, 50 and 100 markers, respectively). VITASSIGN1 had a high true positive rate but an intermediate type I error rate, except when 100 markers were used.

[insert Figure 4]

Effect of the number of offspring

In general, when the number of offspring increased, the variance of the positive assignment error rate decreased (Figure 5). When the number of markers was low and/or when some parents were missing, the positive assignment error rate displayed high variance in datasets with fewer than 500 offspring. In datasets with more than 500 offspring, the average positive assignment error rate converged to a value lower than the user-defined limits of 1% or 5%, except in three cases (Figure 5c) where it

reached 2.3% with APIS1 (20 missing sires, 42 and 50 markers) and 5.8% with APIS5 (20 missing sires, 50 markers). We also saw that the median of the positive assignment error rate converged to its asymptotic value in datasets with 200 offspring or more.

[insert Figure 5]

Effect of the genotyping error rate

The effect of genotyping error on assignments is presented in Figure 6. For a 1% genotyping error, the results were consistent with those previously obtained with the real seabass dataset, except with VITASSIGN which produced a lower true positive rate and a higher type II error rate. When the genotyping error rate increased from 1% to 3%, the type II error rate increased and true positives decreased with all programs. When the simulated genotyping error rate was 3%, APIS generated the same results whatever the genotyping error parameter of the software (1% or 3%), which was not the case for CERVUS where both the true positives and type I error rate increased (Supplementary Figure 2).

[insert Figure 6]

Discussion

This new parentage assignment method implemented under APIS and based on the observed distributions of Mendelian transmission probabilities, was designed to provide an alternative to the simulation process proposed by most likelihood-based parentage assignment software programs. When the number of markers is high and all the parents are genotyped, all programs are able to assign all offspring to their true parents without errors. However, when the situation is more complex, with fewer markers and/or missing parents in the dataset, APIS can produce more reliable results by limiting the positive assignment error rate. Such complex situations, especially involving missing parents, are quite common. Missing parents can be due to traceability issues, technical problems during genotyping (DNA degradation, extraction and genotyping issues), or an absence of identified biological samples. Such issues may or may not be known to the user. Indeed, offspring with missing parents but declared as being assigned to an incorrect parent pair will lead to pedigree errors, which in turn may cause incorrect estimations of breeding values (Banos, Wiggans, & Powell, 2001; Visscher, Woolliams, Smith, & Williams, 2002). In Visscher et al. (2002), 10% of the pedigree was incorrect, leading to a predicted loss of selection response of 2%-3%. Banos et al. (2001) showed an 11%-15%

reduction in EBV estimates when there were 11% of paternity errors in the pedigree. Pedigree errors can also have a considerable impact on the conservation of populations, leading to management errors and hence to a reduction in genetic diversity when compared to an optimum based on true pedigree information (Oliehoek & Bijma, 2009).

APIS was designed to enable users to set their specific acceptable type I error rate, although minimizing the type II error rate may also be an objective, particularly when the variable of interest is the proportion of unassigned offspring (e.g. in a stock restoration program, identifying animals arising from natural reproduction rather than restocking operations). However, a general rule is that, all things being equal, minimizing the type I error will lead to an increase in the type II error (DePoy & Gitlin, 2016). This was the case with our results (Figure 4c), so APIS is not the most appropriate software when the aim is to minimize the type II error.

Under most of the conditions tested, the positive assignment error rate observed was lower than the user-set maximal value. The one exception was scenarios where 50% of the sires were missing, but in these cases the user-set error rate was only modestly exceeded. When the proportion of offspring with at least one missing parent was high, the median $P2_m(:)$ value was shifted to lower levels. This was probably caused by the fact that in these cases, the distribution of $P2_m(:)$ is a mix of $P2_m(:)$ from individuals with both parents genotyped and from individuals with missing parents. The $P2_m(:)$ values for individuals with missing parents were indeed equivalent to the $P3_m(:)$ values for individuals with both parents genotyped, which were a little lower than the $P2_m(:)$ values for individuals with both parents genotyped. As the proportion of individuals with at least one missing parent increased, the distribution of $P2_m(:)$ shifted downwards, its median decreased and the estimate of N_{miss} tended to become lower than its real value, resulting in more relaxed thresholds and more type I errors in the results than expected. However, although the user-defined threshold of APIS was not respected in a few extreme cases, APIS performed better than CERVUS and VITASSIGN in terms of the type I error rate in those cases and overall (Figure 4c).

As APIS uses the observed distributions of Mendelian transmission probability, several factors may impact the shapes of distributions and hence assignment efficiency. We showed that the power of the marker set impacted the overlap between $P1_m(:)$ and $P2_m(:)$ and the estimate of the assignment threshold. High proportions of offspring with missing parents could also lead to misestimates of the

threshold, as discussed before. We showed that APIS required an offspring population of 500 to match the aim of controlling the positive assignment error rate (Figure 5). However, except in a few cases where the proportion of individuals with missing parents was underestimated, 200 offspring were generally sufficient for the median positive assignment error rate to reach its asymptotic value (Figure 5). Thus, APIS is not designed to handle small batches of offspring, and 200 individuals appear to be the very minimum. A fourth factor that could impact distributions and threshold estimates could be inbreeding. It has previously been shown that the type I error rate increases as the degree of relatedness between the members of a trio increases (Anderson & Garza, 2005; Marshall, Slate, Kruuk, & Pemberton, 1998). When the true parents of an offspring are full-sibs, or one of the true parents is a full-sib of the offspring, exclusion may be more efficient than a likelihood approach (Anderson & Garza, 2005). In the case of APIS, when the offspring are divided into groups with varying levels of inbreeding, their average $P1_m(:)$ will vary as a function of inbreeding. In an inbred group, AA genotypes will be more frequent in both parents and offspring, leading to higher Mendelian transmission probabilities than in an outbred group (Table 1). Depending on the population structure, different peaks may then appear in $P1_m(:)$, leading to unreliable results, mainly caused by an incorrect estimate of the number of offspring with missing parents. In the European seabass commercial cohort we tested, and even though inbreeding was not estimated, it is reasonable to consider that it was higher than in a wild population because of the history of domestication and selection. However, as the number of parents at each generation (in this case, 39 sires and 14 dams in a yearly cohort) and the number of families were quite high large, we could expect inbreeding to be quite evenly distributed in the population. And indeed, in this situation, we obtained reliable results as shown by the study.

The mating design may have an impact on the reliability of APIS in the same similar way as inbreeding. If one sex repeatedly produces very few offspring, this could lead to problems concerning distributions with multiple peaks. For this reason, the output of APIS shows the distribution of $P1_m(:)$, $P2_m(:)$, $\Delta_1(:)$ and $\Delta_2(:)$ to help the user identify these potential issues.

Another limitation to our study was our knowledge of the sex of the parents. If the sex is unknown for all potential parents and the same genotype file is used for sires and dams, the true parents will be assigned as they stand (the true sire assigned as the sire and the true dam as the dam) but also the

other way round (true sire assigned as the dam and true dam assigned as the sire). Those two symmetrical parent pairs will be the first and second best pairs, and in this situation, $\Delta_1(o)$ will be equal to zero, so APIS will not be able to determine the threshold, leading to no assignment. One solution to this is that the user should detect such situations in the APIS log file and gradually assign arbitrary sexes to the parents of offspring in which $\Delta_1(o)$ is equal to zero, until all $\Delta_1(o)$ have positive values.

Finally, APIS is not suitable for the assignment of overlapping generations. In this situation, some individuals may be in both the offspring and parental genotype datasets. If this is the case, such individuals may be assigned to themselves as a sire or dam because the Mendelian transmission probability of "parent" pairs containing the individual to be assigning is very high. Users should therefore take care not to include the same genotypes as parents and offspring.

The causes of assignment errors mainly involve genotyping errors and missing parental genotypes. In many likelihood-based assignment software programs, the genotyping error rate is required as a parameter of the model (Kalinowski et al., 2007). If this parameter is not correctly estimated, this can lead to assignment errors (Grashei et al., 2018). The genotyping error rate in APIS is fixed at 1%, so that the software is easier to use. We were able to show that using the APIS algorithm, this value was not critical for assignment efficiency as setting it at 1% or 3% produced identical results with simulated data, with a 1% or 3% genotyping error (Supplementary Figure 2), as hypothesized by Boichard et al. (2014). Nevertheless, very high error rates such as the 3% rate applied in some of the simulated datasets on salmon had a noticeable effect on assignment efficiency (Figure 6). It is interesting to test such levels of errors in order to assess the robustness of the software, but they are not expected to occur in SNP genotyping (Ranade et al., 2001).

The second important parameter is the proportion of missing parents. Under APIS, this proportion is estimated from the distributions of Mendelian transmission probabilities. Although APIS sometimes overestimates the proportion of missing parents (leading to higher type II error rates) this parameter is usually unknown to the user and an incorrect estimate (especially setting it at zero when this is not the case) can lead to incorrect assignment using other software (Oddou-Muratorio, Houot, Demesure-Musch, & Austerlitz, 2003), a situation we observed during our tests with CERVUS.

APIS was designed to limit the type I error rate. Because it does not require any input parameters other than the genotypes of the potential parents and of the offspring, as well as the maximum acceptable positive assignment error rate accepted by the user, APIS is particularly convenient and efficient when the user does not have any prior information on the population to be assigned. Although we showed that exclusion, using VITASSIGN, was more efficient in several cases (notably when numerous markers were available, where the type I error rate fell to an acceptable level), it also requires an input parameter that is the maximum allowed number of mismatches. This number is set empirically and depends on numerous factors, such as the variability of the markers used, the genotyping error and the number of markers, without existing decision rules (Vandeputte et al., 2006). In our data, we observed a significant level of type II error in the 50-marker set with VITASSIGN1 (Figure 4). This was caused by one marker, present in the 50-marker set but not in the others, that was affected by a number of genotyping errors, making the empirical limit of one mismatch ineffective in managing genotyping errors in this case, while with two mismatches these type II errors disappeared. However, when 120 markers were used and the genotyping error rate was 1% or 3%, two mismatches were not sufficient to account for genotyping errors with VITASSIGN (Figure 6). As APIS sets the thresholds itself, it is more convenient than exclusion if the user does not have any previous information on how many mismatches can be allowed.

Acknowledgments

This work received partial financial support from the GeneSea project (n° R FEA 4700 16 FA 100 0005) funded by the French Government and the European Union (EMFF, European Maritime and Fisheries Fund) under the "Appels à projets Innovants" managed by FranceAgrimer. The PhD scholarship for Ronan Griot was partially supported by the ANRT (PhD scholarship n° 2017/0731) and SYSAAF. We are grateful to the GeneSea partners Ferme Marine du Douhet and Ecloserie Marine de Gravelines-Ichtus, and particularly to Sophie Cariou for providing the European seabass cohort. We would also like to thank the INRA genotyping platform Gentyane for the production of genotype data. We thank Victoria Hawken for the English editing.

References

Anderson, E. C., & Garza, J. C. (2005). The Power of Single-Nucleotide Polymorphisms for Large-Scale Parentage Inference. *Genetics*, *172*(4), 2567–2582. doi: 10.1534/genetics.105.048074

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., … Johnson, E. A. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE*, *3*(10), e3376. doi: 10.1371/journal.pone.0003376

Banos, G., Wiggans, G. R., & Powell, R. L. (2001). Impact of Paternity Errors in Cow Identification on Genetic Evaluations and International Comparisons. *Journal of Dairy Science*, *84*(11), 2523–2529. doi: 10.3168/jds.S0022-0302(01)74703-0

Boichard, D., Barbotte, L., & Genestout, L. (2014). AccurAssign, software for accurate maximum-likelihood parentage assignment. *Proceedings, 10th World Congress of Genetics Applied to Livestock Production*.

Chakraborty, R., Meagher, T. R., & Smouse, P. E. (1988). Parentage analysis with genetic markers in natural populations. I. The expected proportion of offspring with unambiguous paternity. *Genetics*, *118*(3), 527.

De Woody, J. A. (2005). Molecular approaches to the study of parentage, relatedness and fitness: practical application for wild animals. *Journal of Wildlife Management*, *69*(4), 1400–1418. doi: 10.2193/0022-541X(2005)69[1400:MATTSO]2.0.CO;2

DePoy, E., & Gitlin, L. N. (2016). Statistical Analysis for Experimental-Type Designs. In *Introduction to Research* (pp. 282–310). doi: 10.1016/B978-0-323-26171-5.00020-3

Durel, C. E., Laurens, F., Fouillet, A., & Lespinasse, Y. (1998). Utilization of pedigree information to estimate genetic parameters from large unbalanced data sets in apple. *TAG Theoretical and Applied Genetics*, *96*(8), 1077–1085. doi: 10.1007/s001220050842

Foerster, K., Coulson, T., Sheldon, B. C., Pemberton, J. M., Clutton-Brock, T. H., & Kruuk, L. E. B. (2007). Sexually antagonistic genetic variation for fitness in red deer. *Nature*, *447*(7148), 1107–1110. doi: 10.1038/nature05912

Glaubitz, J. C., Rhodes, O. E., & Dewoody, J. A. (2003). Prospects for inferring pairwise relationships with single nucleotide polymorphisms. *Molecular Ecology*, *12*(4), 1039–1047. doi: 10.1046/j.1365-294X.2003.01790.x

Grashei, K. E., Ødegård, J., & Meuwissen, T. H. E. (2018). Using genomic relationship likelihood for parentage assignment. *Genetics Selection Evolution*, *50*(1). doi: 10.1186/s12711-018-0397-7

Jamieson, A., & C S Taylor, St. (1997). Comparisons of three probability formulae for parentage exclusion. *Animal Genetics*, *28*(6), 397–400. doi: 10.1111/j.1365-2052.1997.00186.x

Jones, A. G., Small, C. M., Paczolt, K. A., & Ratterman, N. L. (2010). A practical guide to methods of parentage analysis: TECHNICAL REVIEW. *Molecular Ecology Resources*, *10*(1), 6–30. doi: 10.1111/j.1755-0998.2009.02778.x

Kalinowski, S. T., Taper, M. L., & Marshall, T. C. (2007). Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment: CERVUS LIKELIHOOD MODEL. *Molecular Ecology*, *16*(5), 1099–1106. doi: 10.1111/j.1365-294X.2007.03089.x

Kruuk, L. E. B. (2004). Estimating genetic parameters in natural populations using the "animal model". *Philosophical Transactions of the Royal Society B: Biological Sciences*, *359*(1446), 873–890. doi: 10.1098/rstb.2003.1437

Marshall, T. C., Slate, J., Kruuk, L. E. B., & Pemberton, J. M. (1998). Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, *7*(5), 639–655. doi: 10.1046/j.1365-294x.1998.00374.x

Misztal, I., Lawlor, T. J., Short, T. H., & VanRaden, P. M. (1992). Multiple-Trait Estimation of Variance Components of Yield and Type Traits Using an Animal Model. *Journal of Dairy Science*, *75*(2), 544–551. doi: 10.3168/jds.S0022-0302(92)77791-1

Morin, P. A., Luikart, G., Wayne, R. K., & the SNP workshop group. (2004). SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, *19*(4), 208–216. doi: 10.1016/j.tree.2004.01.009

Oddou-Muratorio, S., Houot, M. L., Demesure-Musch, B., & Austerlitz, F. (2003). Pollen flow in the wildservice tree, Sorbus torminalis (L.) Crantz. I. Evaluating the paternity analysis procedure in continuous populations. *Molecular Ecology*, *12*(12), 3427–3439. doi: 10.1046/j.1365-294X.2003.01989.x

Oliehoek, P. A., & Bijma, P. (2009). Effects of pedigree errors on the efficiency of conservation decisions. *Genetics Selection Evolution*, *41*(1), 9. doi: 10.1186/1297-9686-41-9

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*(2), 249–252. doi: 10.1111/j.1365-294X.1995.tb00214.x

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Retrieved from https://www.R-project.org/

Ranade, K., Chang, M. S., Ting, C. T., Pei, D., Hsiao, C. F., Olivier, M., … Botstein, D. (2001). High-throughput genotyping with single nucleotide polymorphisms. *Genome Research*, *11*(7), 1262–1268. doi: 10.1101/gr.157801

Sancristobal, M., & Chevalet, C. (1997). Error tolerant parent identification from a finite set of individuals. *Genetical Research*, *70*(1), 53–62. doi: 10.1017/S0016672397002851

Vandeputte, M. (2012). An accurate formula to calculate exclusion power of marker sets in parentage assignment. *Genetics Selection Evolution*, *44*(1), 36. doi: 10.1186/1297-9686-44-36

Vandeputte, M., & Haffray, P. (2014). Parentage assignment with genomic markers: a major advance for understanding and exploiting genetic variation of quantitative traits in farmed aquatic animals. *Frontiers in Genetics*, *5*. doi: 10.3389/fgene.2014.00432

Vandeputte, M., Mauger, S., & Dupont-Nivet, M. (2006). An evaluation of allowing for mismatches as a way to manage genotyping errors in parentage assignment by exclusion. *Molecular Ecology Notes*, *6*(1), 265–267. doi: 10.1111/j.1471-8286.2005.01167.x

Vignal, A., Milan, D., SanCristobal, M., & Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution*, *34*(3), 275. doi: 10.1186/1297-9686-34-3-275

Visscher, P. M., Woolliams, J. A., Smith, D., & Williams, J. L. (2002). Estimation of Pedigree Errors in the UK Dairy Population using Microsatellite Markers and the Impact on Selection. *Journal of Dairy Science*, *85*(9), 2368–2375. doi: 10.3168/jds.S0022-0302(02)74317-8

Wilson, A., Réale, D., Clements, M. N., Morrissey, M., Postma, E., Walling, C. A., … Nussey, D. H. (2010). An ecologist's guide to the animal model. *Journal of Animal Ecology*, *79*(1), 13–26. doi: 10.1111/j.1365-2656.2009.01639.x

Data accessibility

APIS is coded in R (R Core Team, 2017) and Fortran. The APIS R package is available on CRAN. The simulated data are available on request from the authors of Grashei et al. (2018). The anonymized European seabass cohort genotypes are available as example data in the APIS R package.

Table 1: Mendelian transmission probabilities for one marker when the offspring is homozygous (AA), conditional on the sire and dam genotypes. Adapted from Boichard et al. (2014)

Sire\dam	AA	AC	CC	Missing
AA	1	0.5	0.01	fa
AC	0.5	0.25	0.01	0.5*fa
CC	0.01	0.01	0.01	0.01
Missing	fa	0.5*fa	0.01	fa^2

C = any allele different from A

fa = frequency of allele A in the offspring population analyzed

0.01 = arbitrary value for genotyping error

Table 2: Mendelian transmission probabilities for one marker when the offspring is heterozygous (AB), conditional on the sire and dam genotypes. Adapted from Boichard et al. (2014)

Sire\dam	**AA**	**AB**	**AC**	**BB**	**BC**	**CC**	**Missing**
AA	0.01	0.5	0.01	1	0.5	0.01	fb
AB	0.5	0.5	0.25	0.5	0.25	0.01	0.5(fa + fb)
AC	0.01	0.25	0.01	0.5	0.25	0.01	0.5*fb
BB	1	0.5	0.5	0.01	0.01	0.01	fa
BC	0.5	0.25	0.25	0.01	0.01	0.01	0.5*fa
CC	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Missing	fb	0.5(fa + fb)	0.5*fb	fa	0.5*fb	0.01	2*fa*fb

C = any allele that is not A or B

fa = frequency of allele A in the offspring population analyzed

fb = frequency of allele B in the offspring population analyzed

0.01 = arbitrary value for genotyping error

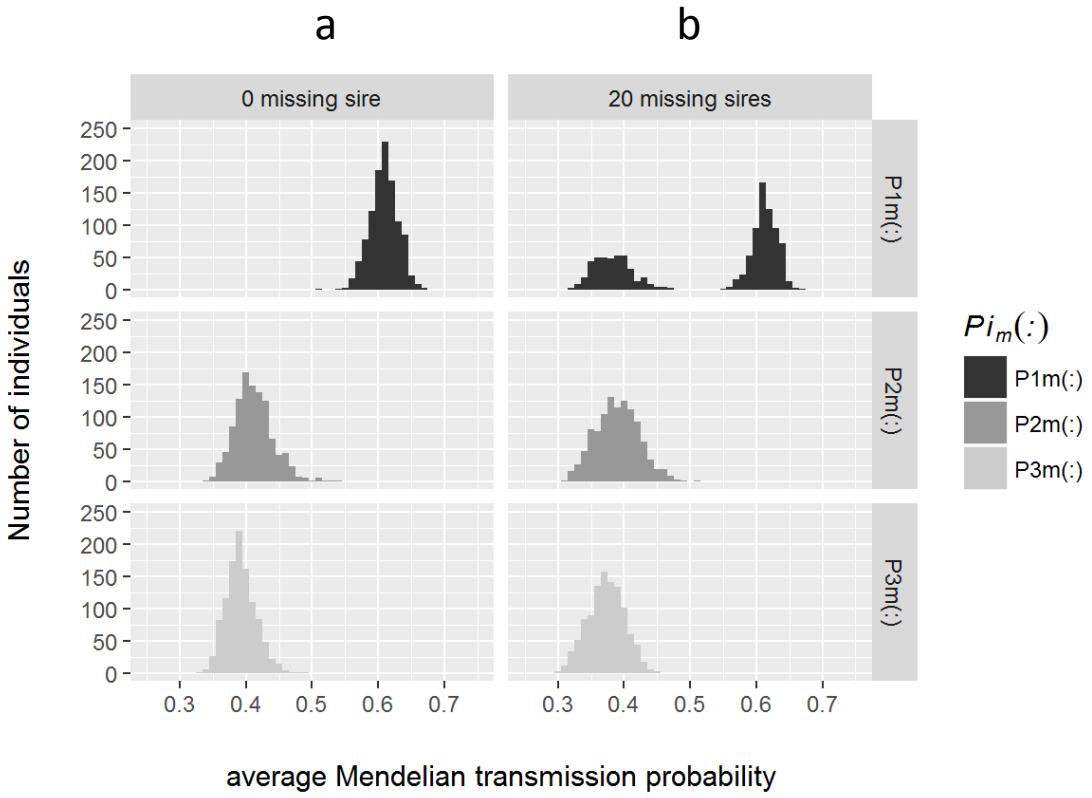

Figure 1: Effect of missing parental genotypes on Mendelian transmission probability distributions in 1068 offspring of European sea bass from 39 sires and 14 dams genotyped for 200 SNPs. Distributions of Mendelian transmission probabilities of the best parent pairs ($P1_m(:)$, in dark grey), of the second best parent pairs ($P2_m(:)$, in medium grey) and of the third best parents pairs ($P3_m(:)$, in light grey) in the case of (a) no missing parents. (b) 20 missing sires out of 39.

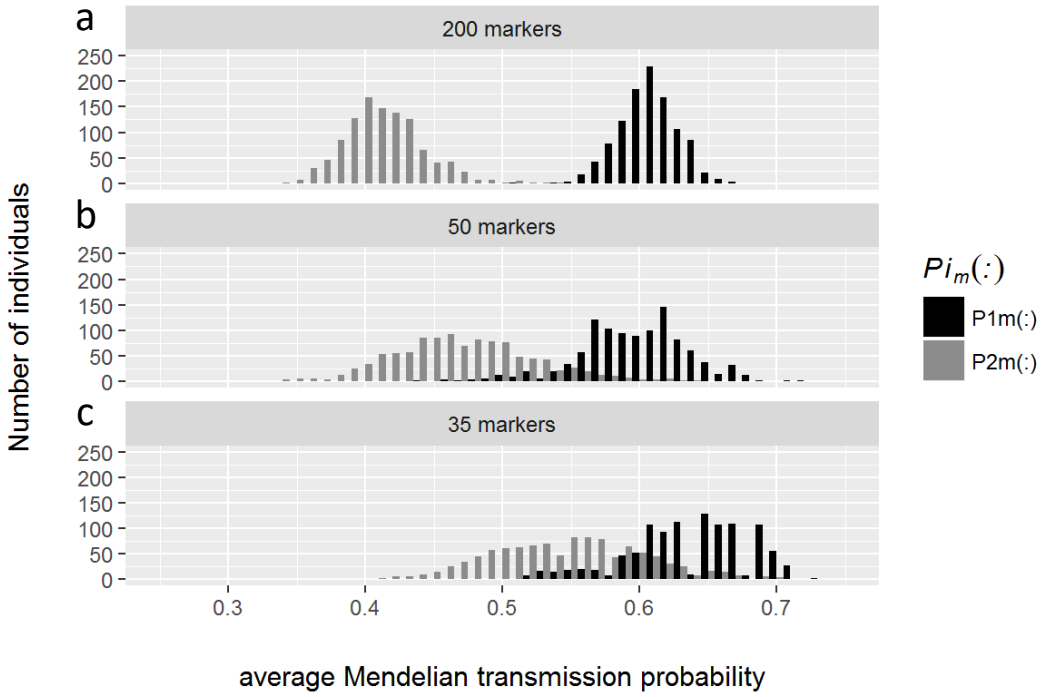

Figure 2: Effect of the number of markers on Mendelian transmission probability distributions in a cohort of 1068 offspring of European sea bass from 39 sires and 14 dams genotyped for 200 (a), 50 (b) or 35 (c) SNPs. The distributions of Mendelian transmission probabilities for the best parent pairs ($P1_m(:)$) are represented in black and those of the second best parent pairs ($P2_m(:)$) are represented in grey.

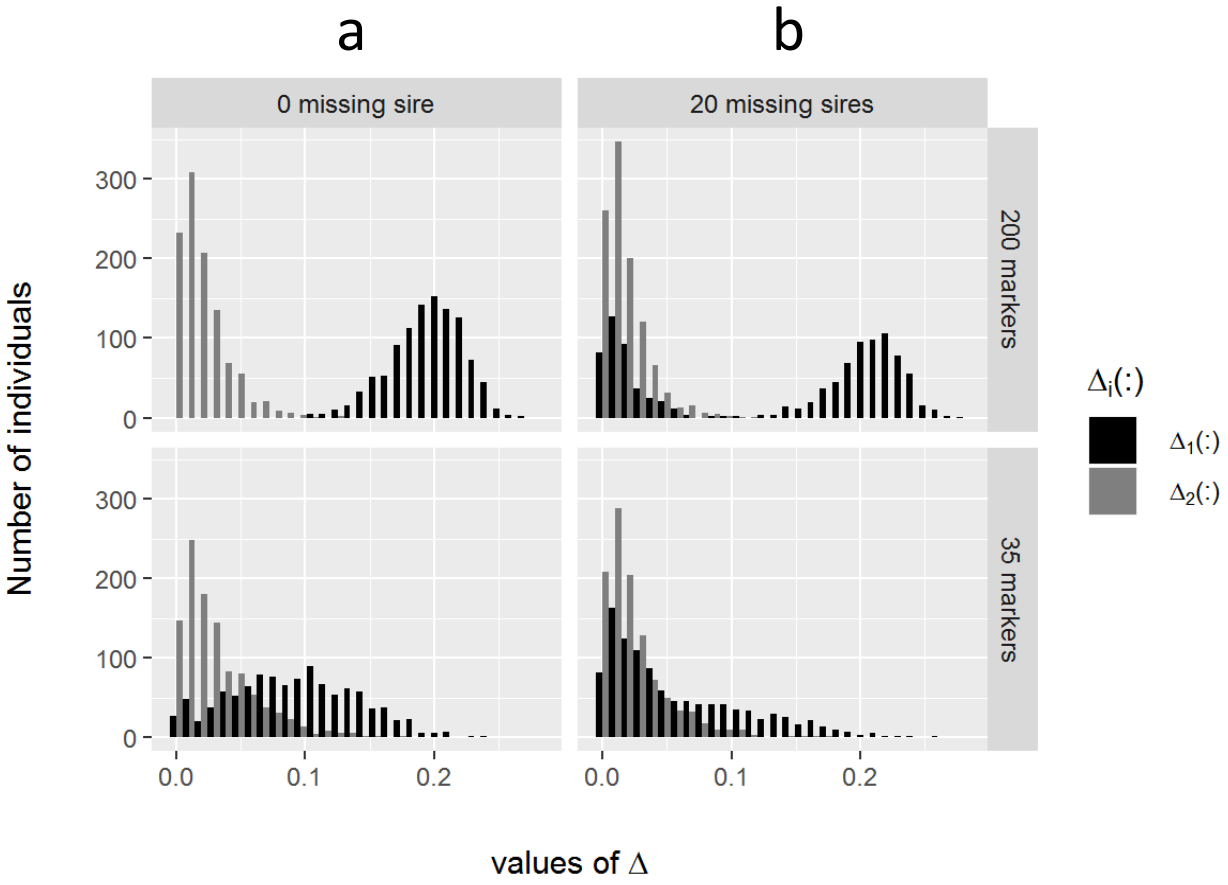

Figure 3: Effect of the number of markers and of missing parental genotypes on Δ distributions in 1068 offspring of European sea bass from 39 sires and 14 dams, with no missing sire (a) or 20 missing sires (b), and with 35 markers (top panels) or 200 markers (bottom panels). $\Delta_1(:)$, in black, is the difference in Mendelian transmission probabilities between the best and second best parent pairs of an offspring, and $\Delta_2(:)$, in grey, is the difference in Mendelian transmission probabilities between the second and third best parent pairs of an offspring.

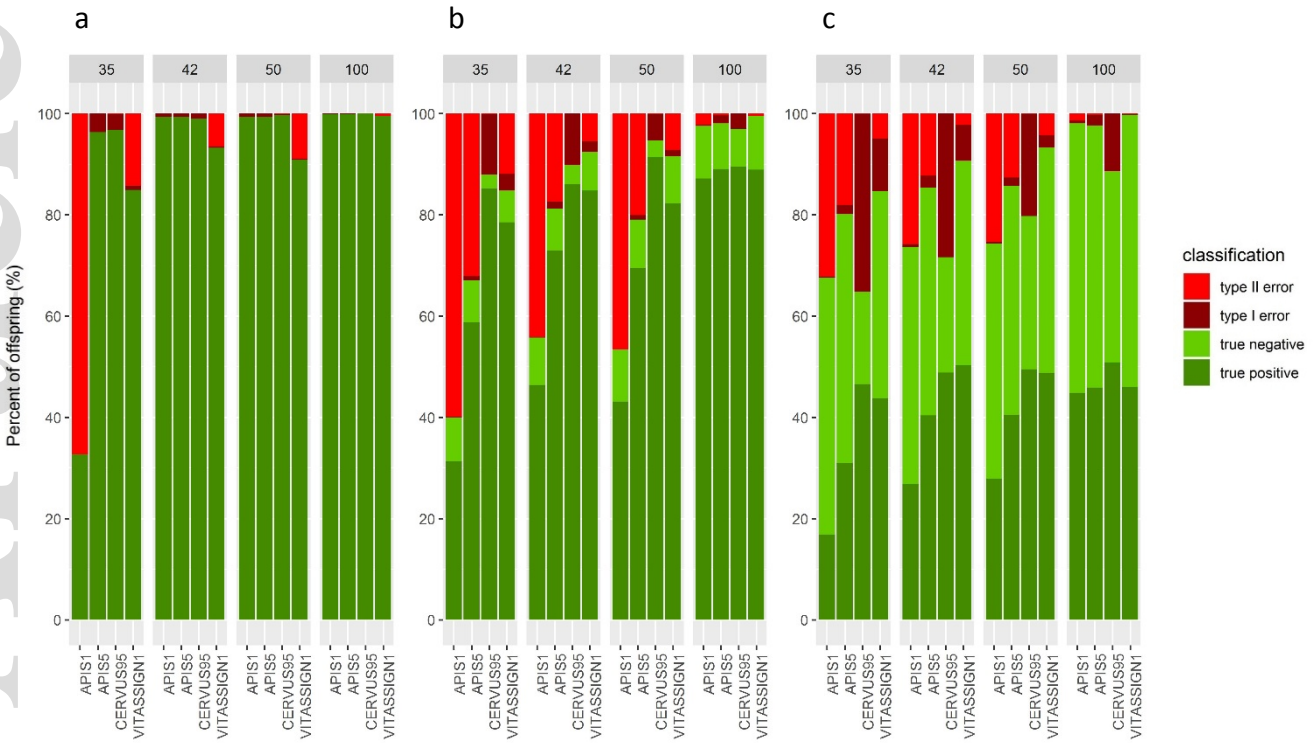

Figure 4: Parentage assignment efficiency with zero (a), four (b) or twenty (c) missing sires in a cohort of 1068 sea bass from 39 sires and 14 dams using APIS1 (1% user-set acceptable positive assignment error rate), APIS5 (5% user-set acceptable positive assignment error rate), CERVUS95 (95% confidence level) and VITASSIGN1 (one mismatch allowed), with four sets of 35, 42, 50 and 100 SNP markers. In panels (b) and (c), the proportions are the means of 10 trials with a random resampling of missing sires.

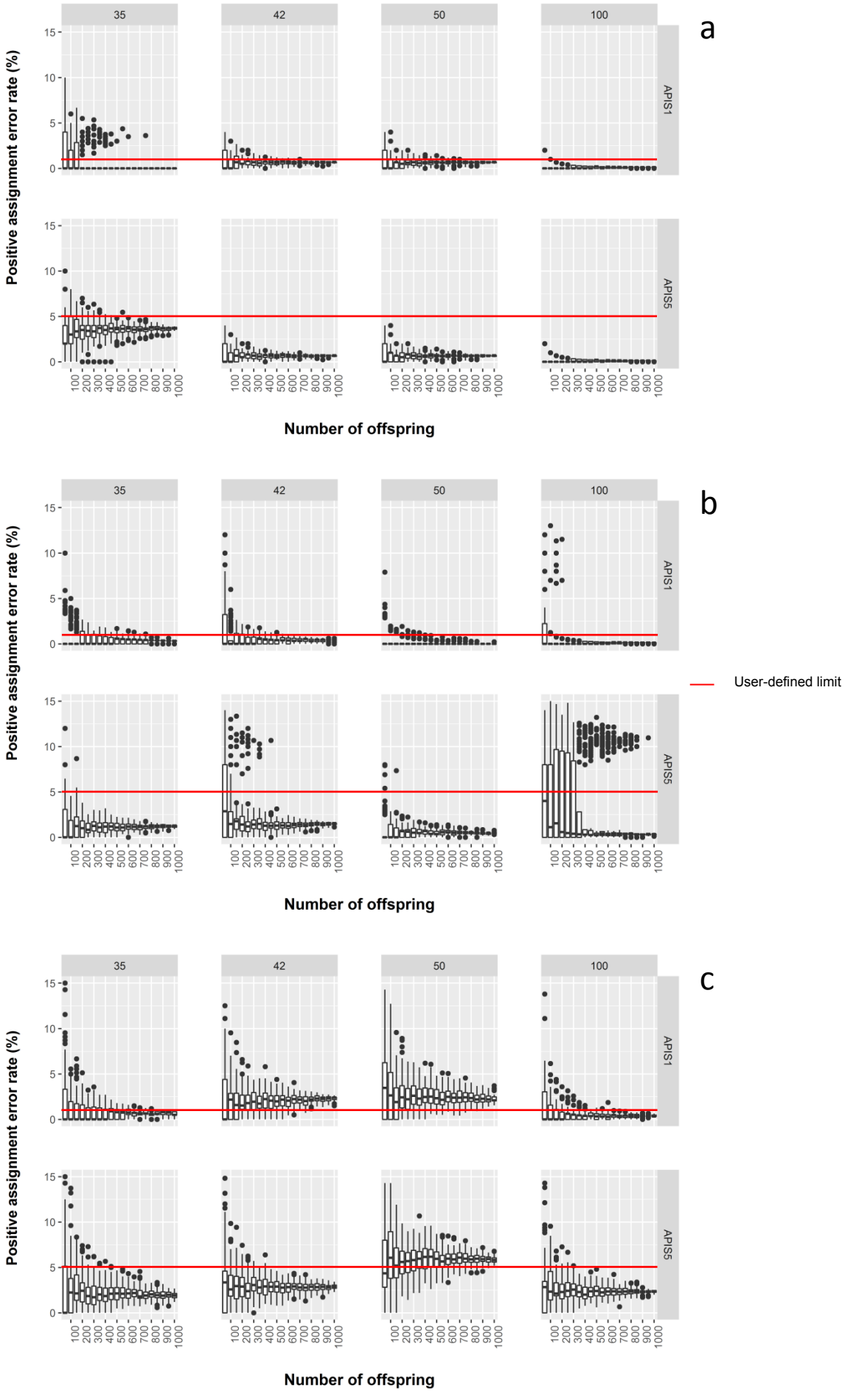

Figure 5: Effect of the number of offspring on the positive assignment error rate using APIS, with zero (a), four (b) or twenty (c) missing sires in a cohort of 1068 sea bass from 39 sires and 14 dams. Each boxplot represents the results of 100 random samples of offspring for each number of offspring tested. Each panel is divided into two rows (top: 1% user-defined maximum positive assignment error rate %, bottom: 5% user-defined maximum positive assignment error rate) and four columns (35, 42, 50 and 100 SNP markers used).The red line represents the user-defined maximum positive assignment error rate.

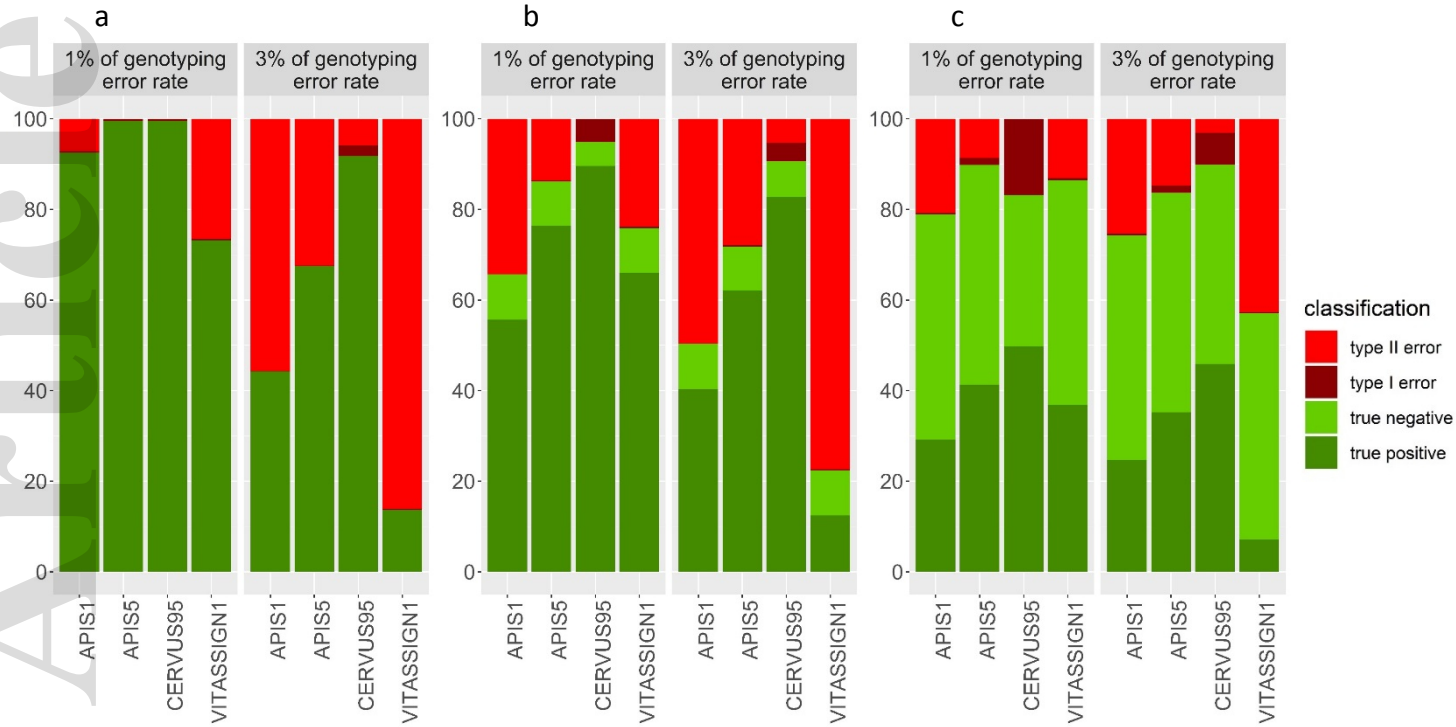

Figure 6: Parentage assignment efficiency with zero (a), ten (b) or fifty (c) missing sires using simulated data from Grashei et al. (2018), with 1000 salmon from 100 sires and 200 dams genotyped on 120 randomly chosen SNP markers, with 1% and 3% genotyping errors tested using APIS1 (1% user-set acceptable positive assignment error rate), APIS5 (5% user-set acceptable positive assignment error rate), CERVUS95 (95% confidence level, 1% genotyping error setting) and VITASSIGN1 (1%, so two mismatches allowed). Assignment efficiency is averaged over six simulated data sets.