Individual-based genotype analysis in studies of parentage and population assignment: how many loci, how many alleles?[1]

Louis Bernatchez and Pierre Duchesne

Abstract: We developed multivatiate analytical models to predict the probability of assigning offspring to parental couples as a function of population size, number of loci, and allelic diversity and the relationships between the probability of allocating individuals to their population of origin as a function of number of loci and allelic diversity. The parentage model predicts that the number of loci and number of alleles contribute interactively to increase assignment success. Given sufficient allelic diversity, a relatively low number of loci is required to achieve high allocation success even for relatively large numbers of possible parents. In contrast, the population model predicts an additive contribution of the number of loci and alleles. There appears to be no significant gain in increasing allelic diversity beyond approximately 6–10 alleles per locus in population assignment studies. Such predictive models should contribute to maximizing the returns of population and parentage assignment studies by increasing our understanding of interactions among the various variables affecting allocation success and by allowing the adjustment a priori of the required level of resolution and, consequently, optimizing the costs–benefits ratio in the use of molecular markers.

Résumé : Nous présentons ici deux modèles analytiques multivariés prédisant la probabilité d'assigner des rejetons aux couples parentaux en fonction du nombre de parents potentiels, du nombre de locus, de la diversité allélique et la relation entre la probabilité d'assigner des individus à leur population d'origine en fonction du nombre de locus et de la diversité allélique. Le modèle parental prédit une contribution interactive entre le nombre de locus et d'allèles sur l'accroissement du succès d'assignation. Étant donnée une diversité allélique suffisante, un nombre relativement faible de locus est requis pour atteindre un succès d'assignation élevé, même quand le nombre de parents possibles est grand. Inversement, le modèle populationnel prédit une contribution additive du nombre de locus et d'allèles. Peu de gain significatif est obtenu en augmentant la diversité allélique au-delà d'environ six à 10 allèles par locus dans les études d'assignation populationnelle. De tels modèles prédictifs devraient contribuer à maximiser le rendement d'études d'assignation parentale et populationnelle en améliorant notre compréhension des interactions entre les différentes variables pouvant influencer le succès d'allocation et en permettant d'ajuster a priori le niveau requis de résolution et donc, d'optimiser le rapport coût–bénéfice de l'utilisation des marqueurs moléculaires.

Introduction

The use of molecular genetics is increasingly contributing to our knowledge of fundamental issues in evolutionary biology of aquatic organisms. For instance, phylogeographic studies have been instrumental in elucidating patterns and processes of postglacial recolonization and highlighting the importance of historical events in shaping the genetic diversity of contemporary populations (Bowen and Avise 1995; Stanley et al. 1996; Colbourne et Hebert 1996; Bernatchez and Wilson 1998; Taylor and McPhail 1999). When used in combination with ecological approaches, the use of molecular genetics also contributed to our understanding of the role of evolutionary forces involved in population divergence and, ultimately, speciation events (e.g., Taylor and Bentzen 1993; Bernatchez et al. 1999; Lu and Bernatchez 1999; Turgeon et al. 1999). Such information, in turn, has been of paramount relevance for applied purposes, such as in defining evolutionary significant units for conservation (Dizon et al. 1992; Bernatchez 1995; Mayden 1995) and optimizing fisheries management (Utter and Ryman 1993; Carvalho and Hauser 1994; Ward and Grewe 1994).

Typically, most applications of molecular genetics rely on estimation of demographic parameters of diversity and differentiation that are derived from averaging the genetic composition over populations. It has been recognized for nearly 20 years, however, that further knowledge of relevance to both evolutionary biology and management may be obtained from the analysis of individual-based genotypic information (Foltz and Hoogland 1981; Hanken and Sherman 1981; Smouse et al. 1982). Yet, the potential of such applications remained relatively unexplored outside the studies of human and plant populations until recently (but see Jordan and Youngson 1992, and references therein). The most likely explanation for this is that the levels of accuracy and precision required in such studies were beyond the reach of available genetic markers (Smouse and Chevillon 1998). The blooming development of new genetic markers over the last de-

Received July 27, 1999. Accepted November 5, 1999.
J15271

L. Bernatchez[2] and P. Duchesne. GIROQ, Département de biologie, Université Laval, Sainte-Foy, QC G1K 7P4, Canada.

[1]Based upon the J.C. Stevenson Memorial Lecture presented at the Canadian Conference for Fisheries Research held in Edmonton, Alberta, in January 1999.

[2]Author to whom all correspondence should be addressed.
e-mail: Louis.Bernatchez@bio.ulaval.ca

cade, namely variable number of tandem repeat loci (VNTRs; microsatellites and minisatellites), has, however, revived a major interest in studies based on the definition of individual multilocus genotypes and opened exciting avenues of research and applications (reviewed in Estoup and Angers 1998; Davies et al. 1999).

Besides genetic mapping, studies based on the analysis of individual multilocus genotypes can be grouped into two broad categories of applications: parentage and population assignments. The former includes studies necessitating the assessment of precise parental relationships within populations, which may be achieved in various ways, including the use of exclusion probability, likelihood methods, and categorical and fractional parental assignment (reviewed in Marshall et al. 1998; see also Meagher and Thompson 1986). This may allow the defining of social structure (Amos et al. 1993), mating patterns (Clapham and Palsbøll 1997; Jones and Avise 1997), kinship (Fontaine and Dodson 1999), and quantification of reproductive success (Rico et al. 1992; Jones et al. 1998). Such analyses may also contribute improvement of the efficiency of selective breeding programs in domesticated populations (Herbinger et al. 1997; Estoup et al. 1998; Ferguson and Danzmann 1998). Studies of population assignments necessitate the determination of population membership of single individuals. This consists in assigning an individual to the population in which its multilocus genotype has the highest probability of occurring, assuming reliable allelic representation, Hardy–Weinberg equilibrium, and locus independence. Such estimation may be relevant to more precisely quantify gene flow and follow movements of individuals (Waser and Strobeck 1998; Palsbøll et al. 1997), determine the degree of differentiation among populations (Paetkau et al. 1995), and establish relationships among individuals within and among populations or higher taxonomic groupings (Nielsen et al. 1997; Roques et al. 1999). An extension of these aproaches is to detect the contribution of stocked fish in natural populations or to detect an admixture of populations in a sample of individuals of unknown origin (Banks et al. 1996; Nielsen et al. 1997; Tessier and Bernatchez 1998; Roques et al. 1999).

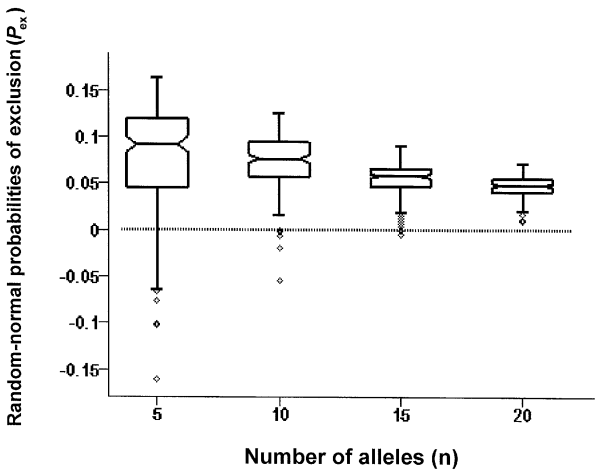
In contrast with the efforts made to find and develop markers suitable for individual-based multilocus genotype analyses, relatively little attention has been paid to determine how the characteristics of genetic markers may affect their usefulness for such purposes. Namely, it would be of particular interest to know a priori what combinations of numbers and types (especially in terms of allelic diversity) of loci should be used in order to reach a required level of resolution and optimize the use of such applications. The few studies that undertook such evaluation were based mainly on specific empirical observations or simulation studies derived from a finite set of loci with particular allelic distributions (e.g., Shriver et al. 1997; Estoup et al. 1998; Marshall et al. 1998). Such studies are certainly instructive; however, it remains unclear how much they can be generalized and used to make predictions in other contexts. This may be better achieved by a simulation exploration that allows a broader coverage of the vast spectrum of possible combinations of factors.

In this study, we used such an approach with the main objective of proposing multivariate analytical functions that could predict population and parentage assignment success as a function of number and properties of loci. For parentage analysis, we were interested in assessing the relationships between the probability of assigning offspring–parental couples in a close population as a function of population size, number of loci, and number and distribution of alleles per locus. We decided to focus on this particular issue for three reasons. First, this represents the most general situation of parentage analysis from which more specific cases can be derived. Second, exploration of parentage analyses at the population level and involving the identification of offspring–parental couples are scarce compared with the bulk of studies concerned with paternity analyses in which the mother–offspring is generally known (reviewed in Marshall et al. 1998). Such situations, however, apply only to those species providing parental care. In contrast, situations where both maternal and paternal parents, as well as their mating patterns, are unknown likely apply to a wider range of aquatic species, particularly in fishes. For population assignment, we were particularly interested in assessing the general relationships between the probability of allocating individuals to populations of origin, as well as its variance, as a function of number of locus and allelic diversity. We decided not to extend this investigation to the exploration of the effect of number of populations and their extent of divergence, which would justify a detailed analysis by itself. However, this has been partly investigated previously using a different procedure (Smouse et al. 1982). Although this paper focuses mainly on aquatic organisms, the models that we propose have general applicabilities and may be used for any sexually reproducing organisms.

Parentage assignment

We want to determine the probability Ps of successfully allocating an individual offspring to its parents based on its multilocus genotype. This is accomplished by the maximum likelihood method detailed in Sancristobal and Chevalet (1997). Briefly, this consists in computing the probability of occurrence of a given offspring genotype among the potential offsprings of each possible parental pair in a population. Once the probability of occurrence of the multilocus genotype of a given offspring is obtained, it is allocated to the parental couple showing the highest probability of producing it. We assume a Wright–Fisher type of reproduction and no prior knowledge on the sex of parents. The number of loci is N, the average number of alleles per locus is n, and the true parental pair belongs to a set of Ng potential parents. The Ps may be viewed as a random variable for a given combination of (n, N, Ng). Given this, we seek to obtain an analytical expression $P(N, n, \text{Ng})$ that will reflect the mathematical structure of relationships between Ps and these three variables. Next, we want to determine analytically, based on P, minimal conditions on the values of (N, n, Ng) so that Ps can be trusted to be at least 90% most of the time. Given any triplet (N, n, Ng), one would then be able to predict with confidence whether or not the 90% success rate threshold will be reached. The search for such minimal conditions led us to build P as a near lower bound for the distribution of Ps(n, N, Ng), thus representing a conservative modelling. We also

Fig. 1. Box plots representing empirical distributions of the differences in probability of exclusion P_{ex} between random allelic distributions and the normal allelic distribution associated with n = 5, 10, 15, and 20 alleles. For each number of alleles, 1000 random distributions were obtained by Monte-Carlo simulations using $U(0, 1)$. Almost all random distributions showed larger values of P_{ex} than the normal distribution for the same number of alleles. To obtain a normal allelic distribution for n (odd) alleles, we build the following list: $\{1/(n-1), 2/(n-2),..., n/(n-1)\}$. Each number is then evaluated by the normal density function $N(1/2, 0.15)$. Finally, each member of this new list is divided by the sum of the list to get $\Sigma = 1$.



aimed at maintaining a small discrepancy (≈ 0.05) between $P(n, N, \text{Ng})$ and the smallest Ps values.

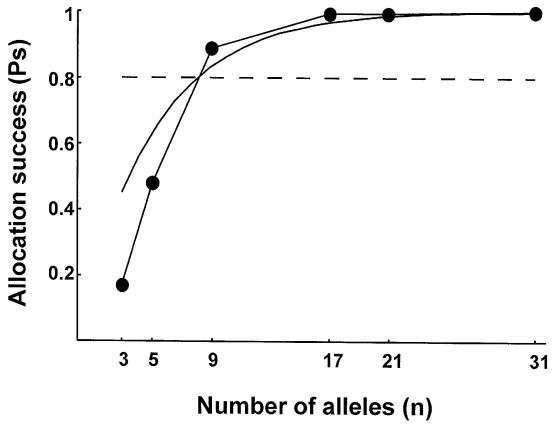
Methodological outlines

We first modellized the expectancy of Ps for allelic distributions with poor information content. To this end, a normal allelic distribution (detailed in the caption to Fig. 1) was constructed for any number of alleles per locus. The distribution of differences between the probability of exclusion P_{ex} (Smouse and Chakraborty 1986) obtained from random and normal allelic distributions of a same number of alleles illustrates the generally poor content of information of a normal allelic distribution for parentage analysis (Fig. 1). Thus, we expected that conservative modelling of the means of Ps for a normal distribution would lead to a reliable lower bound P for random allelic distributions.

To obtain an analytical expression $P(n, N, \text{Ng})$ to be used as a conservative lower bound for the distribution of $\text{Ps}(n, N, \text{Ng})$, we performed the following steps. First, simulations were done to collect estimates of the expectancy of Ps, $E(\text{Ps})$, for each of a series of specific (N, n, Ng) combinations. We then proceeded to a conservative modelling of these estimates as an analytical expression, $P(N, n, \text{Ng})$. Finally, the predictive potential of P as well as its structural similarity to Ps for random allelic distributions was tested by comparing the fit between independent data obtained by simulations and values predicted by the model.

Precisely, individual Ps values were computed with the following procedure: (1) construction of the normal allelic distribution for n alleles, (2) random generation of Ng parental genotypes for N loci with the allelic distributions obtained from step 1, (3) random generation of 100 offsprings from step 2, (4) parentage assignment of offsprings among all possible parent pairs using the maximum likelihood method, and (5) estimation of the proportion of correct allocations. The $E(\text{Ps})$ values were estimated from the average of Ps values based on 20 such realizations for all combinations of the following parameter values: Ng = 10, 20, 30, 40, 50, and 70, N = 1, 2, 3, 4, 5, 6, and 7, and n = 3, 5, 9, 17, 21, and 31. All procedures were performed using programs written with the algebric computer system Maple V, version 5. A detailed description of the procedure used for estimating parentage assignment of offsprings is provided in Appendix 2.

Fig. 2. Observed relationship between estimates of the expectancy of allocation success Ps and number of alleles n together with the corresponding analytical curve in the case of five loci and 20 parents.



Stepwise modelling of E(Ps) for normal allelic distributions

The modelling of $E(\text{Ps})$ as a function of n, N, and Ng was done in four steps. We first established the relationship between the number of alleles n and $E(\text{Ps})$. The empirical data show that $E(\text{Ps})$ is an increasing function of n that converges to unity (Fig. 2). The rate of increase of $E(\text{Ps})$ is regularly diminishing with n. A simple, straightforward way of modelling such behaviour is

$$(1) \qquad p(n) = 1 - (1/s)^n.$$

Because we were mainly interested in high values of $E(\text{Ps})$, we choose s so that $p(n_0) = 0.8 = \text{Ps}(n_0)$. In other words, the desired s makes the empirical and analytical curves cross precisely when they reach 0.8 (Fig. 2). Using this procedure, we computed s for each number of loci (N = 1–7). As exemplified in Fig. 2 for N = 5, $p(n)$ remained under the empirical data for all values over 0.8, in agreement with our search for a lower bound for Ps. The same pattern was observed for other numbers of loci.

The speed parameter s is dependent on N and consequently may be viewed as a function of the number of loci. The second step was thus to modellize the functional relationships between the number of loci and s. The empirical relationship between s and N (varying from 1 to 7) has an obvious linear component (Fig. 3). The number of parents Ng clearly influences $s(N)$ (Fig. 3). For each of Ng = 10, 20, 30, 40, 50, and 70, we fitted the parameters m and b to their

Fig. 3. Empirical relationships between the number of loci *N* and the speed parameter *s* as defined in eq. 1 for various numbers of parents (10, 20, 30, 40, 50, and 70).

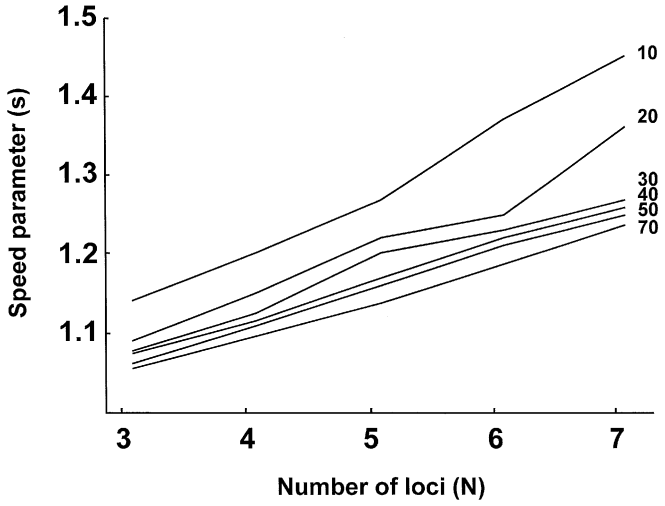
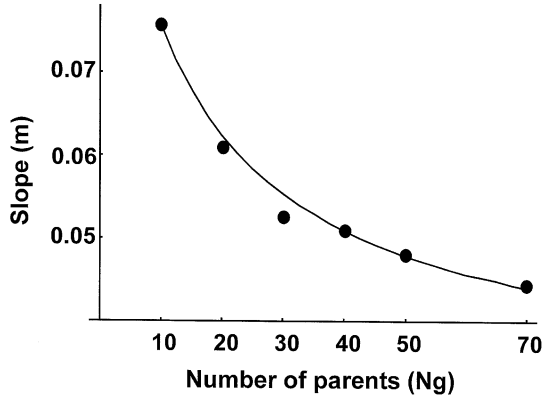


Fig. 4. Empirical relationship between the number of parents Ng and the slope parameter *m* as defined in eq. 2 along with the fitted analytical curve.



respective empirical *s*(*N*) curves. In all cases, *b* was estimated at ≈ 0.915 such that

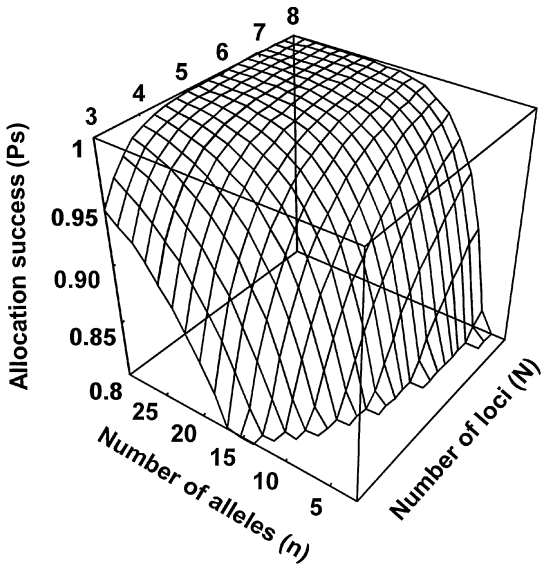
$$(2) \qquad s(N) = mN + 0.915.$$

The third step was thus to modellize the functional relationships between the number of parents Ng and *m*. The empirical relationships between *m* and Ng (varying from 10 to 70) show a decreasing trend in the form 1/*x*. A satisfactory fitting between empirical data and the model (Fig. 4) was reached with the following equation:

$$(3) \qquad m(\mathrm{Ng}) = \frac{1}{(\mathrm{Ng}+13)} + 0.032.$$

Equating eqs. 1, 2, and 3, we now define *P*(*n*, *N*, Ng), the relationships between parentage assignment success as a function of population size (Ng), number of loci (*N*), and number of alleles per locus (*n*), as follows:

Fig. 5. According to the analytical model *P*, the surface representing the functional relationship between allocation success *P*, number of loci *N*, and average number of alleles *n* per locus in the case of 20 parents.



$$(4) \qquad P(n, N, \mathrm{Ng}) = 1 - \left(\frac{1}{\left(\frac{1}{\mathrm{Ng}+13} + 0.032 \right) N + 0.915} \right)^{n}.$$

Figure 5 illustrates the relationships between the predicted parentage assignment success *P*, *n*, and *N* for Ng = 20. The graph illustrates that for any given number of loci, it is always possible to tend towards *P* = 1 by increasing the number of alleles per locus. Clearly, this indicates that increasing the number of alleles is highly advantageous in parentage assignment. Also, the rate at which *P* approaches unity depends strongly on the number of loci. Equation 4 also predicts that parentage assignment success will decrease with an increasing number of parents, but not linearly (see below).

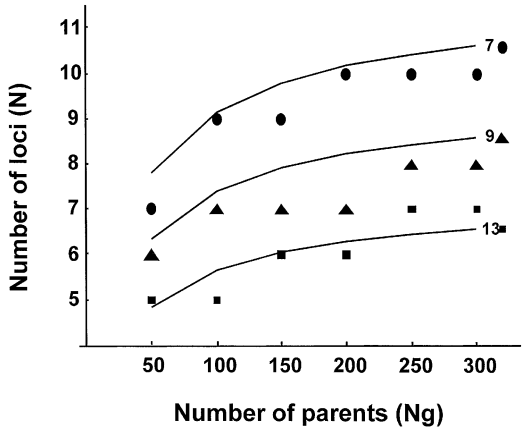
Modelling *E*(Ps) for uniform allelic distributions

In order to explore the generality of predicting parentage assignment from a simple analytical function, we use the stepwise procedures described above to model *E*(Ps) as a function of *n*, *N*, and Ng for uniform allelic distributions (identical frequencies of 1/*n* for all alleles). A modelling procedure identical to the one applied for a normal allelic distribution led to the following analogous expression:

$$P_{\mathrm{uni}}(n, N, \mathrm{Ng}) = 1 - \left(\frac{1}{\left(\frac{1}{\mathrm{Ng}+6} + 0.057 \right) N + 0.88} \right)^{n}.$$

which reflects a mathematical structure identical to that obtained for a normal allelic distribution. Given that P_{ex} values

Fig. 6. Empirical points and analytical curves representing minimal number of loci N to reach allocation success Ps ≥ 0.90 as a function of number of parents Ng for 7, 11, and 13 alleles.



obtained for most randomly generated allelic distributions stand between P_{norm} and P_{uni}, their structural similarity may indicate the generality of the relationships between parentage assignment success (Ps) as a function of (N, n, Ng).

Predictive potential

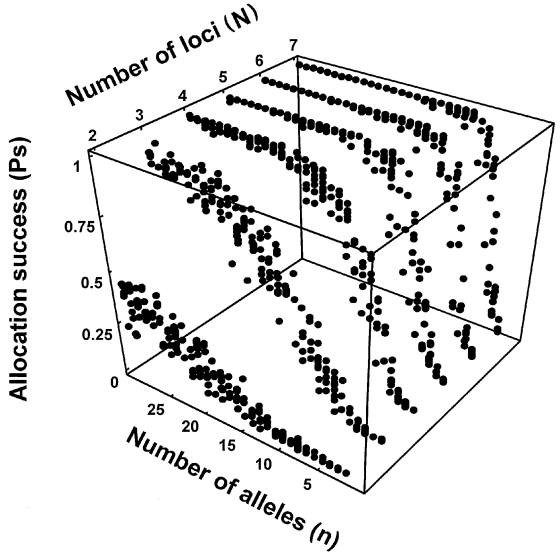
As stated above, our goal was to identify (N, n, Ng) triplets for which Ps will be at least 0.90. Thus, we empirically searched for the minimal N values that, given a normal allelic distribution and specific values of n and Ng, were large enough to ensure that Ps ≥ 0.90. This quantity was denoted $N90$. Simulations were done to estimate $N90$ in all combinations of the following number of alleles (n = 7, 9, 11, and 13) and number of parents (Ng = 50, 100, 150, 250, and 300). Each estimate of $N90$ resulted from the random generation of parents and 100 offspring as described above. Then, we compared these empirical $N90$ values with analytical N derived from $P(N, n, \text{Ng})$. Analytical N was evaluated as an explicit function of (n, Ng, P) derived from eq. 4:

$$(5)\quad N_{norm}(n, \text{Ng}, P) = \frac{0.125\left(1000\left(\frac{1}{1-P}\right)^{\frac{1}{n}}\text{Ng} - 915\text{Ng} + 13\,000\left(\frac{1}{1-P}\right)^{\frac{1}{n}} - 11895\right)}{(177 + 4\text{Ng})}.$$

We found that N evaluated at P = 0.9 generally exceeded $N90$ for any combination of n and Ng. We reasoned that if the conservative nature of eq. 4 was the main source of discrepancy, then we should be able to find a smaller value for P such that $N90$ and N would agree most of the time. In fact, we found that P = 0.83 produced a very close approximation to empirical $N90$ (Fig. 6). In other words, given specific numbers of alleles and number of parents, if one sets $N = N_{norm}(n, \text{Ng}, 0.83)$, then the triplet ($N$, n, Ng) will be minimal in the sense that an increase in either N or n or a decrease in Ng will only raise its allocation power and consequently the confidence that Ps > 90%.

In order to further investigate the generality of the above equation to assess P as a lower bound for individual realisations of Ps, we performed simulations with random allelic distributions involving random sets of 20, 70, and 100 parents. For each N loci varying from 2 to 7 (from 2 to 8 in the case of Ng = 100), 200 n values (restricted between 2 and 31) were randomly generated. One hundred offspring were also randomly generated for each combination of (N, n, Ng), from which the proportion of correct allocations (Ps) was computed. It can be seen from Fig. 7 that P and the random realizations of Ps have very similar structures. We then divided the empirical triplets (N, n, Ng) into two categories: those with equal or greater power than minimal triplets and those with lesser power than minimal triplets. For both categories and each number of loci, we computed the proportion of realizations having Ps ≥ 0.90. In the high-power category, we found very high proportions of Ps ≥ 0.9 for all numbers of loci except N = 2 and N = 3. The overall percentage of high values was 84% for Ng = 20, 98% for Ng = 70, and 100% for Ng = 100. Excluding triplets with N = 2 or 3, we obtained 95% for Ng = 20, 100% for Ng = 70, and 100% for Ng = 100. Conversely, the low-power category yielded low percentages of high values (Ps ≥ 0.90) for Ng = 20 (7%), Ng = 70 (10%), and Ng = 100 (12%). Because the procedure to determine analytical minimal triplets was based on data with Ng as high as 300, we believe that similar results would be obtained for higher numbers of parents as well.

Fig. 7. Three-dimensional representation of 1200 triplets (N, n, Ps), each composed of a number of loci N chosen at random from 2 to 7, a number of alleles n chosen at random from 2 to 31, and of the allocation success Ps resulting from a Monte-Carlo simulation in the case of 70 parents.



Practical use of minimal triplets (N, n, Ng)

Minimal triplets may be used to determine the minimal number of loci to reach a parental assignment success P ≥ 0.90. For instance, if the number of parents is 200 and the average number of alleles per locus is 7, then a minimal N would be $N_{norm}(7, 200, 0.83) \approx 10$. Conversely, one may seek a minimal number of alleles per locus, given an upper limit on the number of loci. Take the case of N = 6 and Ng = 250. Using eq. 5, one finds four minimal triplets satisfying these

initial conditions: (6, 13, 250), (6, 14, 250), (6, 15, 250), and (6, 16, 250). Similarly, it can be verified that given 9 loci and 8 alleles per locus, any triplet with a number of parents in the range of 150–300 parents will be minimal. Because such calculations may be tedious, we programmed an electronic spreadsheet to compute minimal triplets over a user-defined array of numbers of parents. Appendix 1 provides a table of selected output values. This spreadsheet is available at the following internet address: http://www.bio.ulaval.ca/LBernatchez.html. Note that P should be viewed as a preanalysis predictor. A more accurate after-analysis estimate of Ps should be sought by running simulations with the established genetic data.

Population allocation

We want to determine the probability Ps of successfully allocating an individual to its population of origin based on its multilocus genotype, and population-specific allelic distribution, which is accomplished by a maximum likelihood algorithm (Shriver et al. 1997). This consists in assigning an individual to the population in which its multilocus genotype has the highest probability of occurring, assuming reliable allelic representation, Hardy–Weinberg equilibrium, and locus independence. Once the probability of occurrence of the multilocus genotype of a given offspring is obtained, it is allocated to the population showing the highest probability of producing it. The Ps will depend on the number of loci N and the average number of alleles per locus n. Given N and n, and considering the distribution of Ps over all possible pairs of allelic distributions between two populations, Ps may be viewed as a random variable. Given this, we first want to explore the relationships of the distribution of Ps as a function of number of loci and average number of alleles per locus with a random allelic distribution. We then seek a bivariate analytical expression $P(N, n)$ that will predict the behaviour of the expectancy of Ps as a function of number of loci and alleles, given all possible allelic distributions.

Methodological outlines

We first estimate the means of Ps for a number of pairs (N, n), which serve as a basis for modelling $P(N, n)$. Then, P is validated through Monte-Carlo simulations with values of (N, n) lying outside the initial modelling domain. Finally, we investigate the precision of $P(N, n)$ as a predictor of individual realisations of Ps for given values of N and n. Precisely, individual Ps values were computed for a given combination of (N, n) using the following procedure: (1) generation of two random allelic distributions, (2) construction of all possible genotypes for each allelic distribution obtained in step 1, and (3) computation of the proportion of successful allocations from step 2. Note that given a specific pair of allelic distributions, the value of Ps obtained is exact and does not depend on a random production of genotypes (Roques et al. 1999). Thus, the probability of correct allocation for any given multilocus genotype g can be exactly computed according to the formula

$$p(g) = \frac{\mathrm{Max}\,(p_1(g), p_2(g))}{p_1(g) + p_2(g)}$$

where $p_1(g)$ is the probability of genotype g in population 1 and $p_2(g)$ is the probability of genotype g in population 2. The expected probability of allocation success over all genotypes is exactly equal to the weighted sum of probabilities:

$$P_G = \sum_{g \epsilon G} p(g)\left(\frac{1}{2}\,p_1(g) + \frac{1}{2}\,p_2(g)\right).$$

To exhaustively run through all possible multilocus genotypes, we first generated this set symbolically. The $p_1(g)$ and $p_2(g)$ values were then computed by substitution of allelic frequencies for the allele symbols. Besides providing precision, this approach proved to be fast enough to generate a large number of pairs of allelic distributions for each combination (N, n). Thus, the expectancy of Ps, E(Ps), was estimated from the average of 1500 pairs of allelic distributions for each of the following combinations: 1 locus, 2–28 alleles; 2 loci, 2–10 alleles; 3 loci, 2–6 alleles; 4 loci, 2–4 alleles; 5 loci, 2 and 3 alleles; 6 loci, 2 alleles; 7 loci, 2 alleles. Increasing the number of combinations was limited by the explosive increment in number of possible multilocus genotypes with increasing number of loci and alleles. All of these procedures were performed using programs written with the algebric computer system Maple V, version 5.

Stepwise modelling of E(Ps)

The modelling of E(Ps) as a function of n and N was done in three steps. We first established the relationships between the number of alleles n and E(Ps). For each number of loci N, E(Ps) is an increasing function $p(n)$ of the number of alleles, which is bounded by an asymptotic value α_N smaller than 1 (Fig. 8, top panel), that is dependent on the number of loci. These properties, combined with data compatibility, led to the following analytical expression:

$$(6) \qquad p_N(n) = \alpha_N - \frac{1}{2}^{n+2}.$$

Second, we modellized the relationship between the number of loci and E(Ps) for $n = 2$. The empirical data indicate that the function $\tilde{p}_2(N)$ has an asymptotic value of 1 (Fig. 8, bottom panel). Moreover, since it is increasing monotonically, a suitable analytical model is $1 - \left(\frac{1}{a}\right)^{N+b}$, $a > 1$. A satisfactory fitting for N ranging from 1 to 8 (Fig. 8, bottom panel) was reached for $a = 1.25$ and $b = 4$ such that

$$(7) \qquad \tilde{P}_2(N) = 1 - \frac{4}{5}^{N+4}.$$

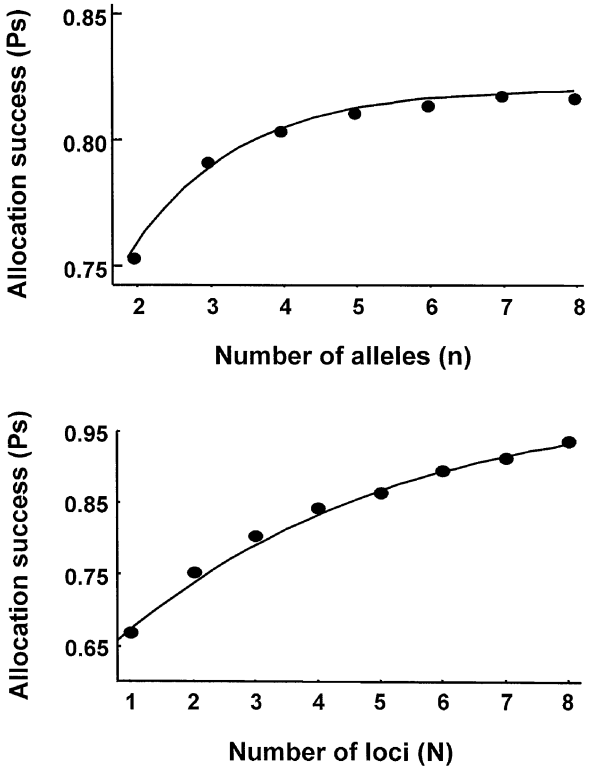
Third, we established the relationships between E(Ps) and pairs of (N, n). If α_N is quantified, then eq. 6 can provide an explicit bivariate function $P(N, n)$. This can be solved by equating eqs. 6 and 7, which have to be equal at $n = 2$:

$$\alpha_N = \frac{17}{16} - \left(\frac{4}{5}\right)^{N+4}.$$

which is then substituted into eq. 6 to define $P(N, n)$, the population assignment success as a function of number of loci and number of alleles per locus, as follows:

$$p(N, n) = \frac{17}{16} - \left(\frac{4}{5}\right)^{N+4} - \left(\frac{1}{2}\right)^{n+2}.$$

Fig. 8. Top panel: empirical data and analytical curve describing the relationship between the expectancy of allocation success Ps and the number of alleles n in the case of two loci. Bottom panel: empirical data and analytical curve describing the relationship between Ps and the number of loci N in the case of two alleles.



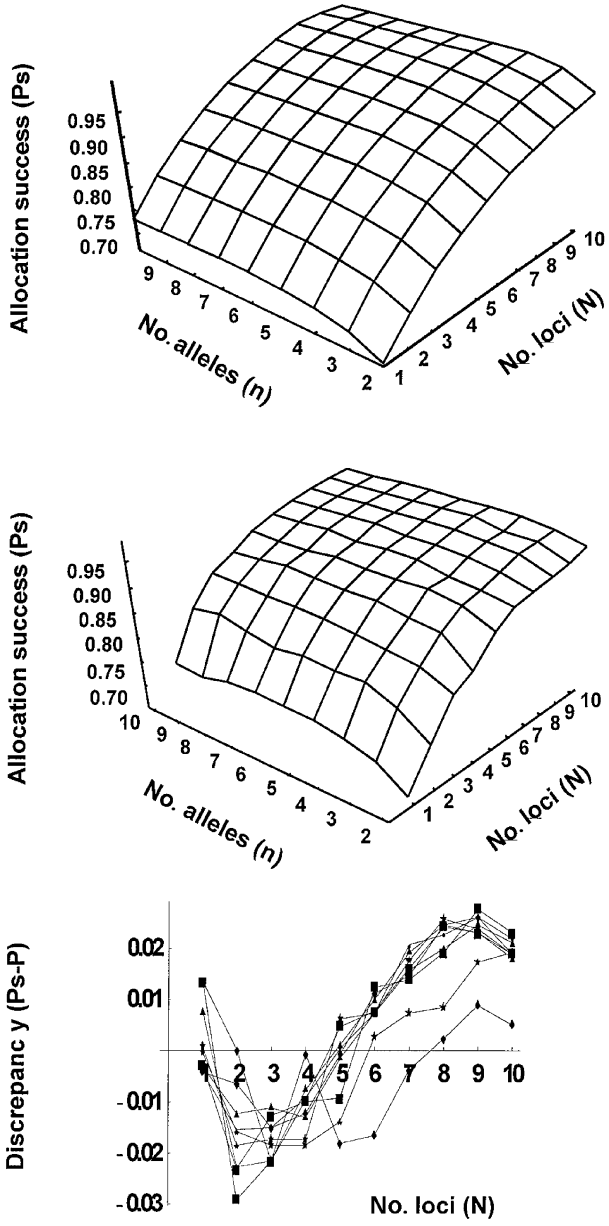
Because the least upper bound of $P(N, n)$, 17/16, is slightly larger than unity, we make a minor correction to the above expression to generate the final model:

$$(8) \qquad P(N,n) = \mathrm{Min}\left(1, \frac{17}{16} - \left(\frac{4}{5}\right)^{N+4} - \left(\frac{1}{2}\right)^{n+2}\right).$$

One important feature of $P(N, n)$ is the additive nature of the respective contributions of number of loci and number of alleles. In other words, according to the proposed model, there is no interaction between N and n. Figure 9 (top panel) illustrates the relationships between the predicted population assignment success P, n, and N. It can be seen that P increases rapidly as a function of N and that it will always be possible to nearly reach $P = 1$ by increasing N for any value of n. Given a moderate average number of alleles per locus, the expected minimal number of loci to reach high levels of allocation success ($P > 0.90$) between two populations is not overwhelming. In contrast, P is relatively independent of the average number of alleles per locus. Thus, for any number of loci, P first increases by augmenting the number of alleles per locus but rapidly reaches an asymptot at approximately $n = 6$. Recall that for any fixed number of loci N, the asymptotic value of P is

$$\alpha_N = \frac{17}{16} - \left(\frac{4}{5}\right)^{N+4}.$$

Fig. 9. Top panel: analytical model $P(N, n)$ shown as a surface over the domain (N = 1–10) × (n = 2–10) of values of number of loci N and number of alleles n. Middle panel: three-dimensional grid structure representing estimates obtained from Monte-Carlo simulations of expectancies of allocation success Ps over the domain (N = 1–10) × (n = 2–10) of values of N and n. Bottom panel: family of curves showing the discrepancies y between Ps and analytical values obtained from the model $P(N, n)$. Each curve traces E(Ps) – $P(N, n)$ as a function of N given a specific n ranging from 2 to 10.

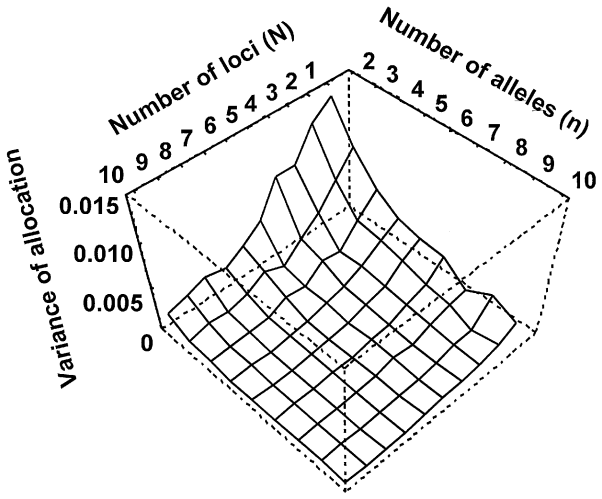


Hence, the maximum gain in P from increasing the number of alleles is $\alpha_N - P(N, 2)$, or

$$\left(\frac{17}{16} - \frac{4}{5}^{N+4}\right) - \left(\frac{17}{16} - \frac{4}{5}^{N+4} - \frac{1}{2}^{2+2}\right) = \left(\frac{1}{2}\right)^4 \approx 0.06$$

for any N value. This maximum gain is reached as soon as $n = 6$ since

Fig. 10. Three-dimensional grid structure representing estimates obtained from Monte-Carlo simulations of variances of allocation success over the domain (N = 1–10) × (n = 2–10) of values of number of loci N and number of alleles n.



$$\left(\frac{1}{2}\right)^{2+2} - \left(\frac{1}{2}\right)^{6+2} \approx 0.06.$$

Clearly, increasing the number of alleles per locus beyond this approximate number of alleles adds little to population assignment success.

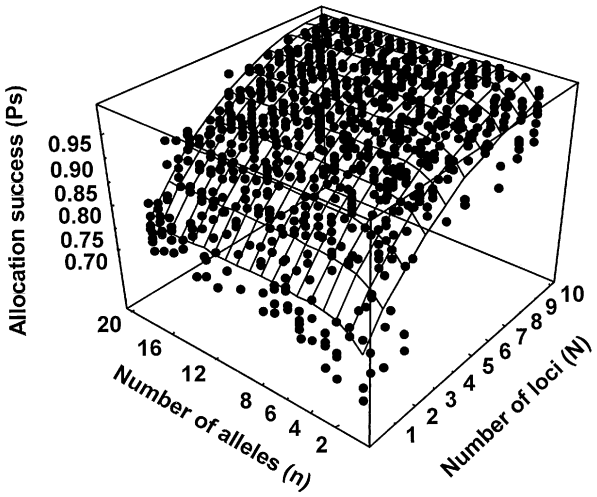
Validation of $P(N, n)$

In order to validate $P(N, n)$, estimates of E(Ps) were calculated for all possible pairs (N, n) for N varying between 1 and 10 and n varying between 2 and 10. For each pair, we performed 100 iterations of the following Monte-Carlo simulation procedure: (1) production of a pair of random allelic distributions, (2) random generation of 500 specimens for each population with allelic distribution generated in step 1, and (3) computation of the proportion of allocation success Ps. For each (N, n), E(Ps) was estimated from the average of the 100 individual Ps values. Note that the above procedure is distinct from the one previously described to generate the model. The main distinctive feature is the exactness of the former, due to the systematic generation of the whole set of genotypes rather than a partial random set. Figure 9 (middle panel) illustrates the strong structural similarity between the three-dimensional web of estimates of E(Ps) and the predicted surface $P(N, n)$. The discrepancy between both estimates is shown in Fig. 9 (bottom panel). The $P(N, n)$ underestimates E(Ps) for smaller numbers of loci ($N < 6$) whereas E(Ps) is overestimated for $N \geq 6$. The discrepancy curves for various values of n are very similar. Hence, the discrepancy appears to stem essentially from the number of loci N. Globally, however, the fit between the model and validation data was excellent, since the maximum absolute difference in allocation success between both methods was less than 0.03.

Predictive potential of $P(N, n)$

So far, we have been focusing on the mean values of Ps distributions to generate and validate the predictive model.

Fig. 11. Three-dimensional representation of 1000 triplets (N, n, Ps), each composed of a number of loci N chosen at random from 1 to 10, a number of alleles n chosen at random from 2 to 20, and of the allocation success Ps resulting from a Monte-Carlo simulation.



To assess the potential of $P(N, n)$ for predicting random values of Ps, however, we quantified the dispersion of the Ps distributions. The data that served to compute mean Ps estimates in the validation procedure also served to estimate variances of Ps (Fig. 10). The dispersion of Ps decreases rapidly as a function of increasing values for both N and n. Hence, Ps distributions with high mean values will have lower dispersions, and consequently, higher values of P will tend to be better predictors. This also indicates that while increasing the number of alleles per locus has little effect on increasing Ps, this may contribute to lowering the discrepancy between predicted and observed values of assignment success.

To evaluate specifically the predictive power of P, we ran additional simulations to produce 1000 individual values of Ps that were generated as follows: (1) random generation of N loci, varying between 1 and 10, (2) random generation of n alleles, varying between 2 and 20, (3) production of a pair of allelic distributions with N loci and n alleles per locus, (4) production of 500 random genotypes for each of both allelic distributions from step 3, and (5) computation of the proportion of allocation success Ps.

The distribution of the 1000 triplets (N, n, Ps) thus obtained is illustrated in Fig. 11. They show high similarity with the model $P(N, n)$. The predictive power of P can be quantified as the percentage of Ps values lying inside the interval ($P - \varepsilon$, $P + \varepsilon$), where ε is a discrepancy value. This showed that Ps has a probability of at least 0.80 of lying inside ($P - 0.05$, $P + 0.05$) for any value of P over 0.85. That is, P predicts Ps ± 0.05 whenever P is relatively high. Note that the 0.05 level compounds the error due to the model and the inherent variability of Ps.

Practical use of $P(N, n)$

Suppose one wants to estimate the minimum number of loci that is necessary to ensure high level of allocation success. First, we solve $P(N, n) = P$ for N from eq. 3:

$$N(n, P) = \frac{12\ln(2) - 4\ln(5) - \ln(17 - 4 \cdot 2^{(-n)} - 16P)}{\ln(5) - 2\ln(2)}.$$

As an example, take an average number of alleles $n = 6$ and a desired level of success Ps = 0.90. Taking into account the error level 0.05, a conservative strategy is to aim at Ps = 0.90 + 0.05 = 0.95. An estimated minimal number of loci is $N(6, 0.95) \approx 6$. One may also be interested in the minimal number of alleles per locus n, given some upper limit on the number of loci. For instance, given $N = 7$, using the above technique and a few trial-and-error evaluations of $N(n, P)$, one finds that $n = 3$ is sufficient to reach Ps ≥ 0.90 in the majority of cases. We programmed an electronic spreadsheet to compute minimal triplets (N, n, Ps) over user-defined arrays of values of n and desired Ps. This spreadsheet is available at the following internet address: http://www.bio.ulaval.ca/LBernatchez.html. As for parental allocation, P should be viewed as a preanalysis predictor. A more accurate after-analysis estimate of Ps should be sought by running simulations with the established genetic data.

Discussion

Our main objective was to use a simulation approach in order to develop multivatiate analytical functions that could predict population and parentage assignment success as a function of number and properties of loci. In both cases, we were able to generate a model with a relatively high predictive power over a wide range of possible situations. Previous studies have described similar trends in relationships between the various parameters that we used in this study. In parentage assignment, for instance, Estoup et al. (1998) observed an increase in allocation success as a function of number of loci and their probability of exclusion, which is highly correlated with allelic diversity. In this study, allocation success also decreased with an increase in the possible number of matings. Similar trends were also reported by Marshall et al. (1998) for paternity analysis. The conclusions of these studies, however, were restricted to observations made from simulations derived from a restricted set of loci with specific allelic distribution. For population assignment studies, Smouse and Chevillon (1998) also reported an increase in allocation success with an increasing number of loci, although they remained relatively vague on the effect of numbers of alleles per locus. Furthermore, none of these studies attempted to explore the multavariate effects of various parameters (e.g., allocation success as a function of number of loci and numbers of alleles per locus). As such, the present study represents to our knowledge a first attempt to generate more general models that can be used to explore the interactive effects of various parameters on allocation success and also to predict allocation success over a wide range of conditions, in both parentage and population assignment studies.

For parentage assignment, the analytical model was built so as to reflect the structure of relationships involving the proportion of successful parental allocation Ps and the triplet of variables (N, n, Ng). In situations where these parameters can be evaluated at least approximately before undertaking a given study, the predictions derived from the model may be used to decide on a minimal number of loci or average number of alleles per locus to be used to reach a satisfactory level of allocation success. Ideally, it would have been desirable to take into consideration finer genetic information, such as specific allelic distribution. In our view, however, this can hardly be incorporated in preanalysis decision making processes. Consequently, our strategy was to develop a model specifically from normal allelic distributions. Despite their poor information content, these showed high structural homology with the whole set of possible allelic distributions. This means that our model may serve as a general approach for predicting the functional relationships between Ps and (N, n, Ng) as well as a lower bound for Ps for all possible allelic distributions.

A first important prediction of the model was that for any number of loci, one can always increase the average number of alleles to reach Ps ≈ 1. The model also predicts that N increases the rate with which n is acting on Ps. Hence, the effects of n and N are closely linked. A second prediction is that the number of parents in the population exerts a dampening effect over the number of loci, but this effect decreases with increasing Ng. Practically, this means that the addition of an extra locus compensates for ever larger increases in the number of parents, such that a reasonable number is required to achieve high allocation success even for relatively high numbers of possible parents, given sufficient allelic diversity. For instance, seven loci with an average number of alleles of 13 would be required to reach an allocation success of 0.90 in a population of 300 possible parents (Appendix 1). There are few empirical data available at this time that could be compared with the predictions of our model. In an ongoing study of parentage assignment in Atlantic salmon (*Salmo salar*), we found that six loci were sufficient to reach an allocation success of $P = 0.90$ in a population of Ng = 75, in close accordance with predictions (D. Garant et al., unpublished data). Estoup et al. (1998) observed that approximately four and five loci were necessary to reach $P = 0.90$, in situations where the number of parents was approximately 30, and mean allelic diversity was 9 and 14, respectively, for two different species. This is also in close accordance with the predictions of our model.

Clearly, however, this model may be improved to take into account other parameters. Namely, the model is based on simple assumptions of the Wright–Fisher random mating model. Note, however, that this mating pattern represents the most stringent situation for allocation success, since the number of possible matings for a given number of parents would be reduced for any other type of mating scheme. A first and obvious extension would be to consider sexed parents, unequal sex ratio, and various reproductive patterns, such as factorial and paternity retrieval scheme (Sancristobal and Chevalet 1997; Estoup et al. 1998). The model could also be modified specifically for paternity studies in which maternal–offspring relationships are known or not (Marshall et al. 1998). In microsatellite studies, allelic scoring errors are not rare (O'Reilly et al. 1998) and may significantly decrease the proportion of allocation success (Sancristobal and Chevalet 1997; Marshall et al. 1998). At this time, we cannot predict the effect of the rate of scoring error on Ps, given specific N, n, and Ng values. Such knowledge would be of great use for predictive as well as theoretical purposes. To minimize the negative impact of scoring errors, this investi-

gation should be led within the framework of an error tolerant allocation procedure (Sancristobal and Chevalet 1997), which excludes few, if any, potential pairs, such that a partly mistaken offspring genotype may still be correctly allocated. Finally, the proposed model assumes that all possible parents are identified and genotyped. Clearly, this assumption will not always be satisfied in wild populations. It will certainly be of practical concern to be able to predict the loss of success rate based on at least approximate estimates of the proportion of missing potential parents (Marshall et al. 1998).

For population assignment, the expectancy of the proportion of population allocation success has been modelled as $P(N, n)$, a function of number of loci and average number of alleles per locus in the specific case of two populations. We observed a relatively high concordance between predicted and observed proportion of allocation success. Thus, the differences between predicted and observed absolute values did not exceed 0.03 over the range of number of loci and alleles covered in the study. One important feature of the model is that it predicts an additive contribution of the number of loci and number of alleles per locus. For a fixed number of loci, the contribution of number of alleles to the total proportion of allocation success never exceeds 6%, and this value is reached soon, approximately at $n = 6$. Clearly, the contribution to allocation success of the number of alleles is largely outweighed by that of the number of loci N. The structure of relationships between N, n, and Ps is thus very different from the one prevailing in the case of parentage assignment where we observed that the number of alleles could always be made large enough to get $P \approx 1$ and a strong interactivity between N and n. In brief, while the use of loci with high allelic diversity is highly advantageous in studies of parentage assignment, there is no apparent gain of doing so beyond a given level (approximately $n = 6$) in population assignment studies. This prediction only stands in a situation of two populations. Although not explored here, our prediction is that maximum gain will be obtained with values of $n > 6$ in situations with more than two populations. However, the asymptotic relationships between P and n would most likely remain.

While the contribution of number of loci to allocation success largely outweighs that of the number of alleles, we observed that the variance in Ps quickly decreased with an increase in both n and N, these two variables being practically interchangeable in this respect. This means that the predictive precision of the model grows with allocation success itself. Consequently, more than 80% of Ps values fell within the interval $(P - 0.05, P + 0.05)$ when $P \geq 0.85$. Thus, an increase in average number of alleles, while adding little to E(Ps), reduces the uncertainty in predicting allocation success. On the other hand, this positive asset may potentially be counterbalanced by sampling errors of low frequency alleles that could reduce allocation success when using maximum likelihood methods (Smouse and Chevillon 1998; Roques et al. 1999). Consequently, we conclude that the best strategy for optimizing allocation success in studies of population assignment is to use loci with moderate allelic diversity, with n varying between 6 and 10 as a rule of thumb.

As for parentage assignment, the model that we developed to predict allocation success in studies of population assignment should be improved to take into account other parameters. In practice, such studies will often involve more than two populations, and consequently, our model could be extended to any reasonable number of populations. A preliminary exploration by simulation approach indicated that the general structure of relationships between Ps, N, and n is not altered with a varying number k of populations. This means that the generalization of $P(N, n)$ to a suitable $P(N, n, k)$ should be possible. Also, we did not attempt to predict allocation success as a function of various levels of population divergence. Clearly, allocation success is expected to increase as a function of population divergence (Paetkau et al. 1995; Smouse and Chevillon 1998; Roques et al. 1999).

To conclude, there is no doubt that in the years to come, the use of individual multilocus information will increasingly contribute importantly to our knowledge of fundamental issues of the biology of aquatic organisms. Such information will in turn contribute to improving management, conservation, and production practices. The use of predictive tools, such as the first generation of models developed here, should contribute to maximizing the returns of such applications by increasing our understanding of interactions among the various variables affecting allocation success and by allowing the adjustment a priori of the required levels of resolution and, consequently, optimizing the costs–benefits ratio in the use of molecular markers.

Acknowledgments

We thank two anonymous referees for their constructive comments on an earlier version of the paper. Research in L.B.'s laboratory has been financially supported over the years by the Natural Sciences and Engineering Research Council of Canada and Fonds FCAR (Quebec).

References

Amos, B., Schlötterer, C., and Tautz, D. 1993. Social structure of pilot whales revealed by analytical DNA profiling. Science (Washington, D.C.), **260**: 670–672.

Banks, M.A., Baldwin, B.A., and Hedgecock, D. 1996. Research on chinook salmon (*Oncorhynchus tshawytscha*) stock structure using microsatellite data. Bull. Natl. Res. Inst. Aquacult. **2**(Suppl.): 5–9.

Bernatchez, L. 1995. A role for molecular systematics in defining evolutionarily significant units in fishes. *In* Evolution and aquatic ecosystem: defining units in population conservation. *Edited by* J.L. Nielsen. Am. Fish. Soc. Symp. **17**: 114–132.

Bernatchez, L., and Wilson, C.C. 1998. Comparative phylogeography of nearctic and palearctic fishes. Mol. Ecol. **7**: 465–452.

Bernatchez, L., Chouinard, A., and Lu, L. 1999. Integrating molecular genetics and ecology in studies of adaptive radiation: whitefish, *Coregonus* sp., as a case study. Biol. J. Linn. Soc. **68**: 173–197.

Bowen, B.W., and Avise, J.C. 1995. Conservation genetics of marine turtles. *In* Conservation genetics: case histories from nature. *Edited by* J.C. Avise and J.L. Hamrick. Chapman and Hall, New York. pp. 190–237.

Carvalho, G.R., and Hauser, L. 1994. Molecular genetics and the stock concept in fisheries. Rev. Fish Biol. Fish. **4**: 326–350.

Clapham, P.J., and Palsbøll, P.J. 1997. Molecular analysis of paternity shows promiscuous mating in female humpback whales (*Megaptera novaeangliae*, Borowski). Proc. R. Soc. Lond. Ser. B, Biol. Sci. **264**: 95–98.

Colbourne, J.K., and Hebert, P.D.N. 1996. The systematics of North American *Daphnia* (Crustacea: Anomopoda): a molecular phylogenetic approach. Philos. Trans. R. Soc. Lond. B, Biol. Sci. **351**: 349–360.

Davies, N., Villablanca, F.X., and Roderick, G.K. 1999. Determining the source of individuals: multilocus genotyping in nonequilibrium population genetics. Trends Ecol. Evol. **14**: 17–21.

Dizon, A.E., Lockyer, C., Perrin, W.F., Demaster, D.P., and Sisson, J. 1992. Rethinking the stock concept: a phylogeographic approach. Conserv. Biol. **6**: 24–36.

Estoup, A., and Angers, B. 1998. Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations. *In* Advances in molecular ecology. *Edited by* G.R. Carvalho. NATO Sciences Series, IOS Press, Amsterdam, The Netherlands. pp. 55–79.

Estoup, A., Gharbi, K., SanCristobal, M., Chevalet, C., Haffrey, P., and Guyomard, R. 1998. Parentage assignment using microsatellites in turbot (*Scophtalmus maximus*) and rainbow trout (*Oncorhynchus mykiss*) hatchery populations. Can. J. Fish. Aquat. Sci. **55**: 715–725.

Ferguson, M.M., and Danzmann, R.G. 1998. Role of genetic markers in fisheries and aquaculture: useful tools or stamp collecting? Can. J. Fish. Aquat. Sci. **55**: 1553–1563.

Foltz, D.W., and Hoogland, D.W. 1981. Analysis of the mating system in the black-tailed prairie dog (*Cynomis ludovicianus*) by likelihood of paternity. J. Mammal. **62**: 706–712.

Fontaine, P.M., and Dodson, J.J. 1999. An analysis of the distribution of juvenile Atlantic salmon (*Salmo salar*) in nature as a function of relatedness using microsatellites. Mol. Ecol. **8**: 189–198.

Hanken, J., and Sherman, P.W. 1981. Multiple paternity in Belding's ground squirrel litters. Science (Washington, D.C.), **212**: 351–353.

Herbinger, C.M., Doyle, R.W., Taggart, C.T., Lochmann, S.E., Wright, J.M., Brooker, A.L., and Cook, D. 1997. Family relationships and effective population size in a natural cohort of cod larvae. Can. J. Fish. Aquat. Sci. **54**: 11–18.

Jones, A.G., and Avise, J.C. 1997. Microsatellite analysis of maternity and the mating system in the Gulf pipefish *Syngnathus scovelli*, a species with male pregnancy and sex-role inversal. Mol. Ecol. **6**: 203–213.

Jones, A.G., Kvarnemo, C., Moore, G.I., Simmons, L.W., and Avise, J.C. 1998. Microsatellite evidence for monogamy and sex-biased recombinaition in the Western Australian seahorse *Hippocampus angustus*. Mol. Ecol. **7**: 1497–1505.

Jordan, W.C., and Youngson, A.F. 1992. The use of genetic marking to assess the reproductive success of mature male Atlantic salmon parr (*Salmo salar* L.) under natural spawning conditions. J. Fish Biol. **41**: 613–618.

Lu, G., and Bernatchez, L. 1999. Correlated trophic specialization and genetic divergence in sympatric lake whitefish ecotypes (*Coregonus clupeaformis*): support for the ecological speciation hypothesis. Evolution, **53**: 1491–1505.

Marshall, T.C., Slate, J., Kruuk, L.E.B., and Pemberton, J.M. 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. **7**: 639–655.

Mayden, R.L. 1995. *In* Evolution and aquatic ecosystem: defining units in population conservation. *Edited by* J.L. Nielsen. Am. Fish. Soc. Symp. **17**: 114–132.

Meagher, T.R., and Thompson, E. 1986. The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. Theor. Popul. Biol. **29**: 87–106.

Nielsen, E.E., Hansen, M.M., and Loeschcke, V. 1997. Analysis of microsatellite DNA from old scale samples of Atlantic salmon *Salmo salar*: a comparison of genetic composition over 60 years. Mol. Ecol. **6**: 487–492.

O'Reilly, P.T., Herbinger, C., and Wright, J.M. 1998. Analysis of parentage determination in Atlantic salmon (*Salmo salar*) using microsatellites. Anim. Genet. **29**: 363- 370.

Paetkau, D., Calvert, W., Stirling, I., and Strobeck, C. 1995. Microsatellite analysis of population structure in Canadian polar bears. Mol. Ecol. **4**: 347–354.

Palsbøll, P.J, Allen, J., Bérubé, M., Clapham, P.J., Feddersen, T.P., Hammond, P.S., Hudson, R.R., Jørgensen, H., Katona, S., Larsen, A.H., Larsen, F., Lien, J., Mattila, D.K., Sigurjònsson, J., Sears, R., Smith, T., Sponer, R., Stevick, P., and Øien, N. 1997. Genetic tagging of humpback whales. Nature (Lond.), **388**: 767–769.

Rico, C., Kuhnlein, U., and Fitzgerald, G.J. 1992. Male reproductive tactics in the threespine stickleback — an evaluation by DNA fingerprinting. Mol. Ecol. **1**: 79–87.

Roques, S., Duchesne, P., and Bernatchez, L. 1999. Potential of microsatellites for individual assignment: the North Atlantic redfish (genus *Sebastes*) species complex as a case study. Mol. Ecol. **8**: 1703–1718.

Sancristobal, M., and Chevalet, C. 1997. Error tolerant parent identification from a finite set of individuals. Genet. Res. **70**: 53–62.

Shriver, M.D., Smith, M.W., Jin, L., Marcini, A., Akey, J.M., Deka, R., and Ferrell, R.E. 1997. Ethnic-affiliation estimation by use of population-specific DNA markers. Am. J. Hum. Genet. **60**: 957–964.

Smouse, P.E., and Chakraborty, R. 1986. The use of restriction fragmment length polymorphisms in paternity analysis. Am. J. Hum. Genet. **38**: 918–939.

Smouse, P.E., and Chevillon, C. 1998. Analytical aspects of population-specific DNA fingerprinting for individuals. J. Hered. **89**: 143–150.

Smouse, P.E., Spielman, R.S., and Park, M.-H. 1982. Multiple-locus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. Am. Nat. **119**: 445–463.

Stanley, H.F., Casey, S., Carnahan, J.M., Goodman, S., Harwood, J., and Wayne, R.K. 1996. Worldwide patterns of mitochondrial DNA differentiation in the harbor seal (*Phoca vitulina*). Mol. Biol. Evol. **13**: 368–382.

Taylor, E.B., and Bentzen, P. 1993. Evidence for multiple origins and sympatric divergence of trophic ecotypes of smelt (*Osmerus*) in northeastern North America. Evolution, **47**: 813–832.

Taylor, E.B., and McPhail, J.D. 1999. Evolutionary history of an adaptive radiation in sticklebacks (*Gasterosteus*) inferred from mitochondrial DNA variation. Biol. J. Linn. Soc. **66**: 271–291.

Tessier, N., and Bernatchez, L. 1998. Contribution des différentes populations de ouananiches à la pêche sportive en lac pour les années 1994 à 1996. Rapport présenté par l'Université Laval à la Corporation de l'Activité Pêche Lac Saint-Jean (CLAP). Université Laval, Sainte-Foy, Qué.

Turgeon, J., Estoup, A., and Bernatchez, L. 1999. Species flock in the North American Great Lakes: molecular ecology of lake Nipigon ciscoes (Teleotei: Coregonidae: *Coregonus*). Evolution, **53**: 1857–1871.

Utter, F., and Ryman, N. 1993. Genetic markers and mixed fisheries. Fisheries (Bethesda), **18**: 11–21.

Ward, R.D., and Grewe, P.M. 1994. Appraisal of molecular genetic techniques in fisheries. Rev. Fish Biol. Fish. **4**: 300–325.

Waser, P.M., and Strobeck, C. 1998. Genetic signatures of interpopulation dispersal. Trends Ecol. Evol. **13**: 43–44.

Appendix 1. Minimum number of loci to reach at least 90% of successful allocations as $N_{norm}(n, Ng, 0.83)$ rounded to the nearest integer

Number of alleles	Number of parents					
	50	100	150	200	250	300
2	32	37	40	41	42	43
3	19	22	23	24	25	25
4	13	16	17	18	18	18
5	11	12	13	14	14	14
6	9	10	11	12	12	12
7	8	9	10	10	10	11
8	7	8	9	9	9	9
9	6	7	8	8	8	9
10	6	7	7	8	8	8
11	5	6	7	7	7	7
12	5	6	6	7	7	7
13	5	6	6	6	6	7
14	5	5	6	6	6	6
15	4	5	6	6	6	6
16	4	5	5	6	6	6
17	4	5	5	5	5	6
18	4	5	5	5	5	5
19	4	4	5	5	5	5
20	4	4	5	5	5	5

Appendix 2. Maximum likelihood parental allocation procedure

Compute the likelihood of an offspring genotype for each pair of parental genotypes
Given the offspring genotype

$$\text{Go} = ([R1, R2], [R3, R4], [R5, R6],...)$$

and the two parental genotypes

$$\text{Par1} = ([M1, M2], [M3, M4], [M5, M6],...)$$

$$\text{Par2} = ([F1, F2], [F3, F4], [F5, F6],...)$$

we first compute the following for locus 1:

$$\text{Pr} = (1/2\text{pr}(M1 \rightarrow R1) + 1/2\text{pr}(M2 \rightarrow R1)) \times (1/2\text{pr}(F1 \rightarrow R2) + 1/2\text{pr}(F2 \rightarrow R2)) + (1/2\text{pr}(F1 \rightarrow R1) + 1/2\text{pr}(F2 \rightarrow R1)) \times (1/2\text{pr}(M1 \rightarrow R2) + 1/2\text{pr}(M2 \rightarrow R2))$$

where

$\text{pr}(A \rightarrow B) = 1$ if $A = B$, $= 0$ if $A \neq B$, assuming mendelian transmission.

The likelihood $L1$ of Go genotype at the first locus among the possible progeny of the parental pair (Par1, Par2) is

$L1 = \text{Pr}$ if locus 1 is a heterozygote for Go

$L1 = \text{Pr}/2$ if locus 1 is a homozygote for Go.

The global likelihood of Go, given the parental pair (Par1, Par2), is the product of all single locus likelihoods:

$$L(\text{Go}) = L1 \times L2 \times L3 \times ... \times Ln.$$

The above computation has to be done for each potential parental pair.

Allocation of offspring Go

Once the set of all likelihoods (one to each parental pair) has been calculated, the highest one is retained. If this maximum belongs to a single parental pair, the offspring is allocated to the latter; otherwise, it is not allocated. Lack of allocation is equivalent to an incorrect allocation when estimating Ps by running Monte-Carlo simulations.