Computer Note

Parentage Analysis with Few Contributing Breeders: Validation and Improvement
PIERRE DUCHESNE, TORBEN MELDGAARD, AND PATRICK BERREBI

From the Département de biologie, Université Laval, Ste-Foy, Québec, Canada G1K 7P4 (Duchesne); the Laboratoire Ecosystèmes Lagunaires, UMR 5119, cc093, Université Montpellier 2, place Bataillon, 34095 Montpellier, Cedex 05, France (Meldgaard and Berrebi); and the Institut des Sciences de l’Evolution, UMR 5554, cc065, Université Montpellier 2, place Bataillon, 34095 Montpellier, Cedex 05, France (Berrebi).

Address correspondence to P. Duchesne at the address above, or e-mail: pierre.duchesne@bio.ulaval.ca.

Validation of parental allocation using PAPA software (Duchesne P, Godbout MH, Bernatchez L. 2002. PAPA (package for the analysis of parental allocation): a computer program for simulated and real parental allocation. Mol Ecol Notes. 2:191–193.) was investigated under the assumption that only a small proportion of potential breeders contributed to the offspring sample. Inbreeding levels proved to have a large impact on allocation error rate. Consequently, simulations from artificial, unrelated parents may strongly underestimate allocation error, and so, whenever possible, simulations based on the actual parental genotypes should be run. An unexpected and interesting finding was that ambiguity (the highest likelihood is shared by several parental pairs) rates below 10% stood very close to exact allocation error rates (true proportions of wrong allocations). Hence, the ambiguity rate statistic may be viewed as a ready-made indicator of the resolution power of a specific parental allocation run and, if not exceeding 10%, used as an estimate of allocation error rate. It was found that the PAPA simulator, even with few contributing breeders, can be trusted to output reasonably accurate estimates of allocation error as long as those estimates do not exceed 15%. Indeed, most discrepancies between exact and estimated error then stood below 3%. Reproductive success variance had little impact on error estimate discrepancies within the same range. Finally, a (focal set) method was described to correct the estimated family sizes computed directly from parental allocations. Essentially, this method makes use of the detailed structure of the allocation probabilities associated with each parental pair with at least 1 allocated offspring. The allocation probabilities are expressed in matrix form, and the subsequent calculations are run based on standard matrix algebra. On average, this method provided better estimates of family sizes for each investigated combination of parameter values. As the size of offspring samples increased, the corrections improved until a plateau was finally reached. Typically, samples comprising 250, 500, and 1000 offspring would bring corrections in the order of 10–20%, 20–30%, and 30–40%, respectively.

Parentage analyses have become increasingly necessary in studies of evolutionary ecology and conservation genetics (Avise 1994; Hughes 1998; Jones and Ardren 2003), and the development of highly polymorphic microsatellite markers has improved the statistical power of such analyses (Luikart and England 1999). The objective of a parental allocation process based on genetic information is to find parental genotypes corresponding to the true parents of each of a set of offspring genotypes (Duchesne and Bernatchez 2007). This, in turn, makes it possible to investigate the “genetic mating system” of the species, to evaluate the actual reproductive success of breeders (Pemberton et al. 1992; Coltman et al. 1999; Bekkevold et al. 2002), and to construct pedigrees. Parentage analysis allows one to confirm monogamy in some species (Ribble 1991; Brotherton et al. 1997) and to demonstrate extrapair copulation in others (Girman et al. 1997; Goossens et al. 1998). They are especially useful when direct mating observations are difficult (Clapham and Palsboll 1997) such as in studies of different spawning strategies in fish (Garcia-Vazquez et al. 2001).

PAPA is a parental pair allocation and simulator program specifically designed to perform parental analysis when all contributing parents belong to the set of scored potential parents (closed systems). Its allocation method is based on the likelihood of each candidate parental pair producing the multilocus genotype found in the offspring being tested. The choice of the PAPA parental analysis software in this study was motivated by a combination of several distinctive statistical and structural features. First, PAPA outputs ambiguity and “correctness rate” statistics (see definitions in the Glossary). Second, the PAPA simulator mimics the entire allocation procedure although the allocation procedure itself does not involve the simulator. This uncoupling of the simulator and allocation procedures favors reliable performance assessments of parental analysis. Also significantly enhancing the reliability of the PAPA simulator is the possible use of the real parental genotypes (the ones collected by the user) as opposed to program-generated, therefore unrelated, parental genotypes. Due to the very circumstances under which genetic parental allocations are run, the proportion of correct allocations over all allocations (the correctness rate)
cannot generally be assessed out of direct, empirical observations. Therefore, one has to resort to simulations in order to estimate correctness rates within specific parental analyses. Besides generating “Poisson” distributions of reproductive success, most current simulators generate offspring from artificial, totally noninbred, parental sets.

Goals of This Study

Within several reproduction systems, both in natural and aquaculture settings, it seems that only a small proportion of potential breeders have any offspring at all. Studies involving a wide range of species such as Pacific oyster (Li and Hedgecock 1998), salmon (Garant et al. 2001), cheetah (Kelly 2001), rainbow trout (Palti et al. 2006), abalone (Hara and Sekino 2007), Bullock’s oriole (Richardson and Burke 2001), giant Galapagos tortoise (Milinkovitch et al. 2004), black rat snake (Blouin-Demers et al. 2005), brown rat (Heiberg et al. 2006), Nile tilapia (Fessehaye et al. 2006), red flour beetle (Bernasconi et al. 2006), and cod (Wesmajervi et al. 2006) have found high variance levels of reproductive success that are totally incompatible with the Poisson model. Studies involving a wide range of species such as Pacific oyster (Li and Hedgecock 1998), salmon (Garant et al. 2001), cheetah (Kelly 2001), rainbow trout (Palti et al. 2006), abalone (Hara and Sekino 2007), Bullock’s oriole (Richardson and Burke 2001), giant Galapagos tortoise (Milinkovitch et al. 2004), black rat snake (Blouin-Demers et al. 2005), brown rat (Heiberg et al. 2006), Nile tilapia (Fessehaye et al. 2006), red flour beetle (Bernasconi et al. 2006), and cod (Wesmajervi et al. 2006) have found high variance levels of reproductive success that are totally incompatible with the Poisson model of the distribution of reproductive success among potential breeders. The Poisson model is based on the assumptions that breeders mate at random, have the same reproductive capability, and produce equally fit offspring.

The problem of potential effects of a strong departure from the Poisson model over both the reliability and the validation of parental analysis came up in the context of an experimental study bearing on the hybridization of marble trout (MT) with brown trout (BT) (see details in Meldgaard et al. 2007). This experimentation from which the genotype data were obtained is based on 35 potential parents, 126 offspring, and 9 microsatellite loci. It was carried out in order to improve knowledge on MT and stocked BT hybridization in the Soca River (Slovenia), which is threatening the native local form (Berrebi et al. 2000).

Parental analysis can and often does provide the cornerstone for research bearing on the ecology of small and sometimes endangered populations. Enhanced knowledge of the specifics of endangered populations is certainly an asset from a conservation perspective. However, breeders within small natural populations are often inbred, sometimes to a very high degree. This is one reason to try to better understand the effects of high inbreeding levels over parental allocation performance. Moreover, populations that are kept in captivity for various purposes such as food production, strain selection, and supplementation programs are sometimes monitored for trait inheritance and fecundity through parental analysis. Those populations also tend to be highly inbred with possible consequences for the quality of both the allocations and the estimation of how good those allocations are. As of now, those consequences still remain to be investigated.

The consequences of the discrepancy between inbred parental structure and assumed parental unrelatedness are evaluated here when only a small proportion of potential breeders contribute offspring (thereafter referred to as the SPC condition). Another aspect that is investigated under SPC is the impact over parental analysis of the variance of reproductive success among the few contributing breeders.

More generally, this paper investigates various statistical properties of the PAPA parental allocation system under the SPC condition. A second goal is to describe and assess a method (the “focal set method”) to improve estimates of family sizes over the ones obtained directly from allocation runs.

This study has been conducted under the assumptions that scoring is error free and that all contributing parents belong to the set of scored potential parents (closed parental system).

**Methods**

All artificial parental as well as offspring genotypes were generated based on real BT and MT specimens sampled in the Driselpoh stream (Slovenia). Nine loci have been scored, and a summary of these microsatellite data is provided in Table 1. Note that the relationships that are investigated in this paper can be generalized to any genetic data set as long as they are interpreted in a qualitative, structural way. For

**Table 1.** Summary of microsatellite data within the BT and MT sampled in September 2001

<table>
<thead>
<tr>
<th>Species marker</th>
<th>Ssa197</th>
<th>Mst543</th>
<th>Mst85</th>
<th>One9</th>
<th>Omm1105ARS</th>
<th>SsaT47</th>
<th>LEE</th>
<th>Omy21</th>
<th>DIAS</th>
<th>Sfo1</th>
<th>SsoSL 311</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Ho</td>
<td>0.719</td>
<td>0.464</td>
<td>0.647</td>
<td>0.504</td>
<td>0.746</td>
<td>0.757</td>
<td>0.624</td>
<td>0.311</td>
<td>0.767</td>
<td></td>
<td></td>
</tr>
<tr>
<td>He</td>
<td>0.962</td>
<td>0.577</td>
<td>0.846</td>
<td>0.846</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.385</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P (HW)</td>
<td>0.0017*</td>
<td>0.3980</td>
<td>0.0988</td>
<td>0.0003*</td>
<td>0.0043</td>
<td>0.0001**</td>
<td>0.0001**</td>
<td>0.5449</td>
<td>0.0001**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>11</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ho</td>
<td>0.735</td>
<td>0.651</td>
<td>0.412</td>
<td>0.534</td>
<td>0.835</td>
<td>0.340</td>
<td>0.804</td>
<td>0.568</td>
<td>0.821</td>
<td></td>
<td></td>
</tr>
<tr>
<td>He</td>
<td>0.944</td>
<td>0.722</td>
<td>0.500</td>
<td>0.444</td>
<td>0.833</td>
<td>0.333</td>
<td>0.667</td>
<td>0.667</td>
<td>0.883</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P (HW)</td>
<td>0.1025</td>
<td>0.5397</td>
<td>1.000</td>
<td>0.2576</td>
<td>0.3895</td>
<td>0.3961</td>
<td>0.0367</td>
<td>0.5293</td>
<td>0.8325</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Com</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* P < 0.05.
** P < 0.01.

\[\text{HEDD}, \text{expected heterozygosity}; \text{HO}, \text{observed heterozygosity}; \text{P (HW)}, \text{value of HW tests within species and locus}; \text{Com}, \text{number of alleles in common between the 2 species}. \]
the sake of clarity, the methods pertaining to the 2 distinct goals will be described in 2 sections.

**Allocations: Error Rates and Ambiguity**

*Simulation Method*

First, several sets of parental genotypes were constructed. Offspring were generated from each set and allocated according to the PAPA algorithm. The exact correctness rates were computed. Estimated correctness rates were obtained from the PAPA parental simulator that uses the parental genotypes. Thus, exact and estimated correctness rates could be compared under various parameter conditions. We now proceed to a detailed description of genotype generation (parents and offspring) and of the allocation procedure.

*Genotype Generation*

Four sets of 35 parents (19 males and 16 females) were generated. Three of these sets were obtained as offspring from grandparental sets of various sizes $S = 4, 8$, and 16, each comprising equal numbers of males and females generated based on the allelic frequencies found in the trout population (see Case Study below). These parental sets will be denoted as grp22, grp44, and grp88, respectively. A fourth parental set, the *unrel* set, was generated directly from these same allelic distributions. In addition, the grandparental sets were given a Russian doll structure so that, for instance, the largest grandparental set (16) included the second largest (8) and the latter the next largest (4). This construction ensured that the larger the grandparental set, the less inbred their offspring, that is, the corresponding parental set. The $\text{M}_x\text{y}$ values, the number of shared alleles per locus averaged over all loci, and all parental pairs (Blouin et al. 1996), for parental sets grp22, grp44, and grp88, were 0.465, 0.380, and 0.353, respectively. The $\text{M}_x\text{y}$ statistic is especially pertinent in the context of parental analysis because allele sharing is directly linked to higher levels of ambiguity and allocation error. The size of the grandparental set will thereafter be considered as the “inbreeding parameter.”

*Allocation Procedure*

The allocation procedure involved 3 components: an allocation matrix, a family size array, and the allocation of random offspring.

*Allocation matrix.* For each of the 4 parental sets, 10 000 offspring were generated using the PAPA parental simulator based on the first 5, 7, and 9 available loci. This resulted in $4 \times 3 = 12$ sets of offspring that were then analyzed in terms of the 19 x 16 = 304 possible parental pairs. Given a set of offspring originating from a parental pair $P_\alpha$, the number of offspring not allocated due to ambiguity and the number of offspring (family size) allocated to each parental pair, including $P_\alpha$, was then recorded. Subsequently, those numbers were divided by their sum to become allocation proportions associated with ambiguity or with $P_\alpha$. Each of those allocation proportions was interpreted as an estimate of the probability that the $P_\alpha$ offspring are nonallocated (ambiguous) or allocated to $P_1$, $P_2$, $P_\eta$, $P_\iota$, respectively. The estimates were denoted as $\text{Pr}_{P_\alpha \rightarrow P_1}$ and put together to form the allocation matrix $<\text{M}_\lambda>$:

$$
\begin{bmatrix}
\text{Pr}_{P_1 \rightarrow P_1} & \text{Pr}_{P_2 \rightarrow P_1} & \cdots & \text{Pr}_{P_\eta \rightarrow P_1} \\
\text{Pr}_{P_1 \rightarrow P_2} & \text{Pr}_{P_2 \rightarrow P_2} & \cdots & \text{Pr}_{P_\eta \rightarrow P_2} \\
\vdots & \vdots & \ddots & \vdots \\
\text{Pr}_{P_1 \rightarrow \text{amb}} & \text{Pr}_{P_2 \rightarrow \text{amb}} & \cdots & \text{Pr}_{P_\eta \rightarrow \text{amb}}
\end{bmatrix}
$$

Thus, 12 such allocation matrices were built.

*Family size arrays* $<N> = <N_1, N_2, \ldots, N_\iota>$. The artificial offspring sample sizes within each allocation run involved 125, 250, 500, 1000, 2500, or 5000 offspring. These offspring were distributed among 18 (6% of all 304 possible pairs) randomly chosen parental pairs according to one of 3 types of family size arrays: Poisson, “aqua,” and “trout,” in ascending order of reproductive success variance. The Poisson 500 offspring sample was obtained from a Poisson random number generator. The aqua 500 offspring sample was distributed following the family size distribution found in a recent aquaculture investigation that involved allocating 573 rainbow trout offspring to 20 potential parental pairs (Palti et al. 2006). The trout 500 offspring samples were distributed very much as in the Case Study presented below. Offspring samples of other sizes were distributed in the same fashion but in proportion to the 500 samples of the same type. Overall, 6 (sample sizes) x 3 (types: Poisson, aqua, and trout) = 18 family size arrays $<N>$ were used in association with each allocation matrix. Note that each artificial sample was collected based on a new, distinct set of 18 randomly chosen parental pairs. Family size arrays are partially represented in Figure 1.

*Allocation of random offspring based on $<\text{M}_\lambda>$. Given a specific allocation matrix $<\text{M}_\lambda>$, associated with*
a parental set and a number of loci, each run of allocations was performed according to the following procedure:

- Eighteen parental pairs (6%) were chosen at random from among the 304 possible pairs.
- A family size was ascribed at random to each of the 18 parental pairs according to some family size array \(<N>\).
- Each offspring “bred” by 1 of the 18 chosen pairs, say \(P_n\), was randomly allocated to 1 of the 304 possible pairs with probability \(Pr_{Pi \rightarrow Pi}\) or \(Pr_{Pi \rightarrow amb}\) as found in the current allocation probability matrix \(<MA>\).

All calculations involved in this section were performed by running a specifically designed Maple© v9.5 code (MapleSoft 1981–2004). The sampling and allocation of a single offspring within this procedure are schematically depicted in Figure 2.

### PAPA Estimates

The PAPA estimates of ambiguity and allocation error associated with each of the 4 parental sets \((grp22, grp44, grp88,\text{ and } unrel)\) and each number of loci (5, 7, and 9) were obtained directly from the corresponding allocation matrices. Note that these estimates involved all possible 304 pairs as would be the case if one ran the PAPA simulator.

### Output Variables

Data were collected for exact allocation error rate, ambiguity rate, and discrepancy between exact allocation error rate and estimated error rate from the PAPA simulator. These variables were investigated in relation to genetic contents, as measured by number of loci (5, 7, and 9), inbreeding level \((grp22, grp44, grp88,\text{ and } unrel)\), and reproductive success (Poisson, aqua, and trout).

### Focal Set Method: Family Size Estimates

**Description**

Small proportions of contributors will generally translate into allocations being ascribed to a small proportion of possible parental pairs. This may be turned into an asset by investigating the detailed allocation structure of the set of parental pairs that obtained at least 1 allocation, hereafter referred to as the “focal set.” Practically, this means estimating, through simulations, the proportion of \(P_i\) offspring that are allocated (correctly) to \(P_i\) or (erroneously) to any of the other pairs belonging to the focal set of parents. In other words, the first step in applying the focal set method is not to correct the allocations themselves. A complete description of the focal set method is provided in the Appendix.

**Example**

The following example serves to illustrate the main idea behind the focal set method. Suppose one knew that a bunch of 100 specimens belong to either population X or Y. After some classification procedure, 70 specimens are allocated to X and 30 to Y. Now the problem is that the procedure is not flawless. In fact, a simulation study has shown that 80% of (true) X specimens are allocated to X and 20% are (mis)allocated to Y. The allocation probabilities for true Y are 70% and 30%, respectively. Based on the above data, one can therefore write the following linear equation system:

\[
NX \times 0.8 + NY \times 0.3 = 70 \quad \text{and} \quad NY \times 0.7 + NX \times 0.2 = 30,
\]

where \(NX\) and \(NY\) refer to the true numbers of specimens belonging to X and Y.

The above system can be solved easily by running some standard computer solver as found in Maple© (MapleSoft 1981–2004) or Excel© packages. In the above case, the solution is \(NX = 80\) and \(NY = 20\).

Under most circumstances, the corrected numbers will be closer to the true numbers than the original allocation numbers. However, the method provides no indication as to which allocations are correct or not, and so the allocations per se are not corrected. It is worth noting that the improvement in allocation number estimates are obtained without any additional information bearing on the specimens such as improved criteria or enhanced measurement.
Table 2. Allocated \((n_a\text{ before slash})\) and estimated \((N_a\text{ after slash})\) family size to each pair of marble (M) trout and brown (B) trout

<table>
<thead>
<tr>
<th>Female/ Male</th>
<th>Bm1</th>
<th>Bm3</th>
<th>Bm4</th>
<th>Bm5</th>
<th>Bm6</th>
<th>Bm7</th>
<th>Bm8</th>
<th>Bm10</th>
<th>Mm5</th>
<th>Mm7</th>
<th>Mm8</th>
<th>Mm9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bf1</td>
<td>38/38*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8/6</td>
<td>(9847)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bf2</td>
<td></td>
<td>32/32*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bf3</td>
<td></td>
<td></td>
<td>16/0</td>
<td>79/127</td>
<td>6/3</td>
<td>(9709)</td>
<td>25/−3</td>
<td>(7170)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mf1</td>
<td></td>
<td></td>
<td>44/44*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mf2</td>
<td></td>
<td>2/2*</td>
<td>15/15*</td>
<td>6/6*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mf4</td>
<td></td>
<td></td>
<td>29/29</td>
<td>(9530)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mf5</td>
<td></td>
<td></td>
<td>21/21</td>
<td>(9704)</td>
<td>11/11</td>
<td>(9764)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mf6</td>
<td>12/12*</td>
<td>7/7*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mf12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22/22</td>
<td>(9677)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Individuals are separated in males (m) and females (f). Numbers in parentheses are the numbers of successful self-allocations of 10 000 multilocus genotypes generated from each pair. Pairs from the parental cohort that did not have any allocated offspring are not shown.

* All 10 000 genotypes were correctly reallocated to the pair.

precision. The extra information bears strictly on the performance of the allocation procedure itself.

The focal set method is a direct application of this same correction technique where parental pairs (the \(P_i\)) stand for populations and family sizes for allocation numbers. Because the number of parental pairs can be very high and therefore the size of the equation system very large, the allocation probabilities have to be expressed in matrix form in order to be solved efficiently by computer solvers.

Output Statistic

In order to assess the efficiency of the focal method relative to family size estimation in comparison with estimates obtained directly from allocations, we used the ratio statistic \(\hat{d}_N / \hat{d}_i \times 100\%\), where

\(\hat{d}_N\): Euclidean distance between the family size array from allocations and that of the exact family size array corresponding to the 18 parental pairs and

\(\hat{d}_i\): Euclidean distance between the family size array estimate after the “focal set correction” and that of the exact family size array.

The comparisons between the 2 family size array estimates, from the focal set method and from the allocations, were performed for the same parameter combinations as given in Allocations: Error Rates and Ambiguity involving inbreeding level, variance of reproductive success, and number of loci.

Case Study

The Driselpoh stream is a tributary of the Soca and Idrijca River basins. The studied part is closed to immigration from downstream. The experimental stretch was emptied by electrofishing each year from 1996 to 1999 to remove unwanted trout from the system. In September 1999, 215 MT and 215 BT reared in a fish farm were released as 1+ individuals of at least 1 year of age, hereafter named the parental cohort. As the existing pure MT populations in the Soca and Idrijca river basin have a low level of genetic variation at microsatellite level (Fumagalli et al. 2002), it was necessary to increase the overall variability to facilitate genetic analyses. This was done by creating a parental stock using males from 1 population and females from another (Lipovscek and Trebuscica; for details, see Meldgaard et al. 2007). The system was sampled by electrofishing in September each following year.

Using PAPA, we were initially able to allocate 411 of the 447 offspring individuals unambiguously (91.9%). Fifteen individuals were allocated to more than 1 couple, and 21 could not be allocated. The average correctness level among the 411 offspring as estimated from the built-in simulations was 0.949. The offspring were allocated to 20 families out of 294 possible pairs. Allocated number of offspring (family size) ranged from 2 to 79 per pair.

In order to use the focal set method, an allocation matrix for the 20 contributing parental pairs was constructed. Allocation of 10 000 generated multilocus genotypes to the 20 families showed large variations in correct allocation proportions between pairs (Table 2). Seven pairs received 10 000 self-allocations, indicating that all offspring from the focal couple would be allocated to the correct pair. Seven pairs received between 9500 and 10 000 self-allocations and the remaining 4 pairs between 6146 and 9500 self-allocations.

Two of the new estimates of family sizes were nonpositive integers (Table 3). These zero or slightly negative sizes occur when pairs are “filled” mainly by offspring that are erroneously allocated to those pairs. When the number of erroneous allocations to a parental pair exceeds the number of correct allocations, the solutions may involve some negatives. In this case, a second round of allocations may be performed without the “empty pairs” followed by a second application of the focal set with the remaining pairs. Here Bf3 × Mm5 and Bf3 × Mm9 were removed from the focal set of parents, and self-allocation was considered only among the remaining 18 pairs. The results are reported in Table 3 where it may be seen that although most family sizes remained unchanged when compared with the first allocation round, some were
substantially modified. These modifications are certainly sufficient to influence the estimates of the variance of reproductive success between pairs and also between individuals within the same sex group.

**Results**

As in the Methods, the results will be reported first for matters of allocation ambiguity and correctness rates and then for the focal set method.

**Allocations: Error Rates and Ambiguity**

*Variables Affecting Allocation Error Rates*

Parental set inbreeding level had a strong effect on allocation error rates (Figure 3). This was especially true when the genetic information contents were low relative to other genetic information contents found in the present study. For instance, with 5 loci, the error rate ranged from 5% to 45% over the unrel, grp88, grp44, and grp22 inbreeding levels. When 9 loci were used, the error rate ran only from 0% to 7% over the same range of inbreeding levels. The variance of reproductive success as defined by the Poisson, aqua, and trout family size arrays had a negligible effect on allocation error rate (data not shown).

**Accuracy of Allocation Error Rates Estimated from PAPA Simulations**

The average discrepancy between the true allocation error and the estimated error from the PAPA simulator decreased as the estimated error also decreased (Figure 4). In other words, high correctness estimates were, on average, better estimates than lower correctness estimates. Under all

![Figure 3](image_url)

**Figure 3.** Relationships between allocation error and number of loci (5, 7, and 9). Each curve represents an inbreeding level (from top to bottom: grp22, grp44, grp88, and unrel).

![Figure 4](image_url)

**Figure 4.** The average discrepancies between exact allocation error and allocation error estimated from the PAPA simulator as a function of the PAPA-estimated correctness. Given any fixed correctness estimate level, the average discrepancies increased with reproductive success variance. Discrepancies for trout, aqua, and Poisson success distributions are represented by triangle, square, and diamond symbols, respectively.

---

Table 3. Allocated (n, before slash) and estimated (N, after slash) family size to each of the 18 remaining pairs of marble (M) trout and brown (B) trout after removal of Bf3 × Mm5 and Bf3 × Mm9 from the focal set of parents and self-allocation only to the remaining 18 pairs.

<table>
<thead>
<tr>
<th></th>
<th>Bm1</th>
<th>Bm3</th>
<th>Bm4</th>
<th>Bm5</th>
<th>Bm6</th>
<th>Bm7</th>
<th>Bm8</th>
<th>Bm10</th>
<th>Mm5</th>
<th>Mm7</th>
<th>Mm8</th>
<th>Mm9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bf1</td>
<td>38/38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8/6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bf2</td>
<td></td>
<td>32/32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bf3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mf1</td>
<td></td>
<td></td>
<td>44/44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mf2</td>
<td></td>
<td></td>
<td></td>
<td>15/15</td>
<td>7/7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mf3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mf4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mf5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mf6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mf12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parents are separated in males (m) and females (f).
simulation conditions, the average discrepancy between true and PAPA estimated error was lower than 4%, 3%, and 2% when the estimates were higher than 85%, 95%, and 99%, respectively.

Reproductive variance had a significant, albeit relatively small, impact on average values of the allocation error discrepancy. As expected, the simulator did best with the Poisson success distribution and better with the aqua than with the trout” distribution. Given nearly identical estimated correctness levels, the impact of reproductive success variance over discrepancy decreased with lower parental inbreeding (Figure 5).

Relationships between Allocation Error and Ambiguity Rates

Allocation error rates dropped more quickly than ambiguity with decreasing inbreeding levels (Figure 6). This was especially true when going from high (grp44) to moderate (grp88) levels of inbreeding. Below the 10% ambiguity rate, the allocation error and ambiguity rates stood close together. In fact, the difference between the averages of these 2 allocation variables never exceeded 3% under all conditions of reproductive variance, number of loci, or parental inbreeding, provided the ambiguity rate stood below the 10% level.

Although ambiguity provided a quick estimate of allocation error that was often doing slightly better or worse than estimates from the PAPA simulator, it was significantly less accurate than simulations within the intermediate inbreeding level range (Table 4). Because the latter condition cannot be easily characterized in specific allocation contexts, it seems safer to rely on the PAPA simulator over the entire range of parental allocation parameters.

Focal Set Method: Family Size Estimates

Under all simulation conditions, there was a noticeable average improvement on the family size array as a result of applying the focal set method (Figure 7). The level of improvement increased with sample size. However, based on simulations involving 10 000 and 20 000 offspring samples, there seems to be an upper bound to the improvement over family size estimates (data not shown). This limit depended on the parameters of the parental system in an apparently unpredictable way but generally stood within the 30–60% interval. Other than size of offspring sample, there were no other clear and systematic effects from variance of reproductive success, inbreeding level of the parental set, or number of loci.

Summary of Main Results

- The higher the inbreeding level of parental sets, the larger the allocation error rate.
- Ambiguity rates below 10% provide a quick and easy estimate of allocation error rate.
- The focal set method improves family size arrays under all parameter combinations.
- Reproductive success variance has a negligible effect on allocation error rates.
- The higher the correctness estimates from PAPA, the more accurate.

Figure 5. The impact of reproductive success variance on the average discrepancies, relative to exact allocation error, of the PAPA error estimate. The black histograms result from simulations run with 5 loci and parental inbreeding level = unrel. The gray histograms result from 7 loci and inbreeding parental level = grp44. With each of the 6 simulation conditions, the correctness estimated from PAPA was 0.95.

Figure 6. Curves represent error (square dots) and ambiguity (diamond dots) as functions of inbreeding level (grp22, grp44, grp88, and unrel). The left to right panels correspond to 5, 7, and 9 loci, respectively. Note that the y axis scale varies between panels.
Discussion

The study of the impact of parental inbreeding level on allocation error rate (see Figure 3) shows that the more inbred the parental set, the larger the allocation error rate. This effect may be large and becomes even more noticeable as the genetic contents are decreased. One immediate consequence is that simulators based on the generation of artificial, therefore unrelated, parents from allelic frequency distributions are likely to output underestimates of allocation error. These overly optimistic estimates may be provided overtly or covertly in the form of some threshold statistical value associated with a predefined allocation confidence level. Therefore, one should try, whenever feasible, to validate parental allocation based on the actual parental genotypes as in parental PAPA simulations as opposed to genotypes generated from sampled allelic frequencies as, for example, in preparental PAPA simulations. In the absence of collected parental genotypes, correctness estimates from artificial parental genotypes are still useful in providing an upper bound for allocation correctness.

It was also found that the higher the correctness level provided by PAPA, the more accurate. As a rule of thumb, estimated correctness rates ≥85% or, equivalently, allocation error rates ≤15% correspond to exact correctness rates reaching at least 80% under most parental allocation conditions. PAPA correctness scores of at least 95% can be trusted to correspond to 90–100% exact correctness with near certainty. Very high correctness estimates (≥99%) may be taken practically for granted. Note that these results refer to the parental PAPA simulations wherein the genotypes of the collected parents are used to generate artificial offspring.

Increasing the reproductive variance also increased the allocation error discrepancy between the PAPA estimates and exact values. This was to be expected because parental allocation simulators assume that all potential parents have the same a priori reproductive capability that leads to a Poisson type of reproductive success distribution. Although success variance increased error discrepancy under all parameter combinations, this effect was found to be larger with more inbred parental sets. This makes sense because the presence of both highly fecund and closely related parents will generally raise the allocation error probability. However, when estimated correctness levels exceeded 85%, reproductive success variance had little impact on estimated error discrepancies (Figure 4). Therefore, success variance can in fact be ignored when estimating allocation error, as it is in parental simulators, provided the estimated allocation correctness scores are high.

A somewhat surprising finding was that ambiguity and allocation error rates are similar as long as ambiguity does not exceed 10%. This interesting statistical property provides a back-of-the-envelope way of estimating allocation error rates in the lower range of ambiguity rates. It is also worth noting that contrary to simulation procedures, ambiguity calculations are assumption free.

Ambiguity rates in the middle range do not decrease as quickly as allocation error rates as a response to extra

<table>
<thead>
<tr>
<th>nb loci</th>
<th>Inbreeding</th>
<th>22</th>
<th>44</th>
<th>88</th>
<th>unrel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ambiguity estimate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.01420</td>
<td>0.06470</td>
<td>0.11902</td>
<td>0.02470</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.02441</td>
<td>0.00861</td>
<td>0.03430</td>
<td>0.00015</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.01010</td>
<td>0.02059</td>
<td>0.00507</td>
<td>0.00012</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PAPA simulator estimate</td>
<td>0.05718</td>
<td>0.05075</td>
<td>0.03097</td>
<td>0.01850</td>
</tr>
<tr>
<td>7</td>
<td>0.02766</td>
<td>0.02656</td>
<td>0.01020</td>
<td>0.00339</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.01804</td>
<td>0.02514</td>
<td>0.00571</td>
<td>0.00084</td>
<td></td>
</tr>
</tbody>
</table>

Figure 7. Reduction (%) of error in family size array from the focal set correction as a function of size of offspring sample. From top to bottom, the curves represent allocation with 5, 7, and 9 loci, respectively. The inbreeding level is grp88 and the offspring sample sizes are 125, 250, 500, 1000, 2500, and 5000.
genetic information. In other words, extra loci are more efficient at eliminating false parents than at differentiating similar candidate parents. A possible explanation is that ambiguous cases sometimes involve inbred parents so that extra loci will also reflect this connection. Consequently, allocation error estimates from the simulator are generally more reliable except when ambiguity is quite low (≤10%).

On average, the focal set correction method provided more accurate family size estimates than did the original allocations under all investigated parameter combinations. The corrections would sometimes reach as high as 70%. Also they depended on the offspring sample size, larger sizes producing more important improvements. However, there always seemed to be an upper, asymptotic limit to the average level of possible correction under each parental context. Clearly, the accuracy of the focal set allocation matrix is enhanced through enlargement of the offspring sample but only up to a point.

A large proportion of the problems raised within parental analysis studies may be tackled by analyzing family sizes associated with parental pairs without specifying the parent-offspring connections. In fact, all questions focusing on the formation of mating pairs fall into this category. A classic example is sexual selection and the variables that are driving it within various species such as size and major histocompatibility complex allele combinations. Reproductive success variance and number of partners for males and females can obviously be calculated through family sizes without explicitly identifying the offspring. In aquaculture settings, parental allocation improves the efficiency of selective breeding programs in several ways (Duchesne and Bernatchez 2007), for instance, by assessing fertility success (Selvamani et al. 2001) for which family size estimates are sufficient. The relationships between various breeding environment factors, such as diet, drug intake, etc., and fitness are readily assessed by compiling and comparing family size estimates. Clearly, all the above applications of parental analysis and several others could benefit from the improved family size estimates brought about by the focal set method.

Unfortunately, the focal set correction has not yet been implemented in PAPA or, to our knowledge, in any of the currently used parental allocation programs. However, given pertinent research goals, performing this correction “manually” can be worth the extra time and effort.

Limitations of This Study

This study does not take allele scoring errors into account although such errors are commonplace (Bonin et al. 2004). However, we believe that most of our findings can carry over to parental systems where scoring error does not exceed usual rates that stand in the vicinity of 1–3% on an allelic basis. The potentially large impact of inbreeding level on allocation error rate would certainly remain. The results bearing on discrepancies between exact and PAPA-estimated allocation error would still hold with allocations run under the most stringent error model wherein no error is accepted. Under such a model, there would certainly be a loss of allocations due to scoring errors leading to some offspring not being allocated but the remaining offspring would behave essentially as in the present study.

However, we can only speculate about the statistical properties of allocation runs allowing for some degree of error scoring. Broadening the allocation error model certainly increases both the allocation error and the ambiguity rates so that the connection between these 2 random variables when ambiguity stands in the low range may be essentially the same. As for the relationship between PAPA error allocation estimates and associated discrepancies, it would probably differ to some extent from the 1 observed in absence of scoring error. However, we believe that PAPA correctness estimates in the high range (≥85%) would generally be associated with low discrepancies (exact vs. estimated allocation error) provided scoring error stays within the usual bracket.

This entire study was run under the assumption that only a small proportion of potential parents have indeed contributed to the offspring sample. Although this prevents the specific results to be generalized sensu stricho to parental allocation with larger proportions of contributors, they certainly are conservative evaluations of the capability of the PAPA simulator to estimate allocation error. Also, there is little doubt that the relationships between ambiguity and allocation error rate would essentially carry over to non-SPC parental systems. In particular, low ambiguity (<10%) should signal similar allocation error rates. Clearly, the use of the focal set method to improve family size estimates is not restricted to low proportions of contributors. However, the size of allocation matrices grows as fast as the square number of contributors with subsequent extra time and effort to build them manually.

Glossary

Allocation error rate: The proportion of incorrect allocations among all allocations (excluding cases of ambiguity). As with correctness rates, it may be exact or estimated. The correctness rate and the allocation error rate are complementary statistics; their sum always equals 1.

Ambiguity: An ambiguous result associated with a specific offspring is one where the highest likelihood is shared by several parental pairs. “Ambiguous” is one of the possible output from the PAPA allocation and simulation procedures.

Ambiguity rate: The proportion of ambiguity cases among offspring processed for parental allocation.

Correct allocation: A correct allocation is one where the pair allocated to the offspring is in fact the true parental pair. When there is no allocated pair, as in the case of ambiguity, the result of the allocation procedure is neither correct nor incorrect. The term “correct” should not be confused with the word “success” as found in
some other parental allocation programs, meaning simply that the offspring has been allocated, correctly or not.

Correctness rate: The proportion of correct allocations among all allocations (excluding cases of ambiguity).

Discrepancy: The absolute (unsigned) value of the difference between exact allocation error rate and some estimate of the allocation error rate.

Exact correctness rate: The true proportion of correct allocations as opposed to estimated correctness rates obtained from running a simulator.

Family size: The number of offspring (among the offspring sample) associated with a specific parental pair. It may be exact or estimated either directly from the allocations or as a result of applying the focal set method.

Family size array: The array (vector) of family sizes of all candidate parental pairs.

Building the Focal Set and the Allocation Family Size Array

First, the real offspring is allocated to the putative parents. The family size allocated to each potential parental pair is then computed. All parental pairs with at least 1 allocated offspring are then collected within a focal set and the corresponding allocation family sizes are recorded within an array, say \( <n_1, n_2, \ldots, n_n, \ldots, n_n> \).

Building the Allocation Matrix

For each of the \( L \) pairs of the focal set, say \( p_i \), a large number (1000) of genotypes are generated as offspring from \( p_i \). These offspring are then allocated to the complete, original set of putative parents. The proportion of allocations of the \( p_i \) artificial offspring allocated to each of the \( L \) focal set pairs is then recorded. Note that allocations to nonfocal pairs are ignored. Those proportions are then divided by their sum so as to add up to unity. Each proportion is now viewed as an estimate of the probability that \( p_i \) offspring are allocated to \( p_1, p_2, \ldots, p_n, \ldots, p_L \), respectively. These estimates are denoted \( Pr_{p_i \rightarrow p_j} \) and put together to form the focal allocation matrix \( \langle M_F \rangle \):

\[
\begin{bmatrix}
Pr_{p_1 \rightarrow p_1} & Pr_{p_2 \rightarrow p_1} & \cdots & Pr_{p_L \rightarrow p_1} \\
Pr_{p_1 \rightarrow p_2} & Pr_{p_2 \rightarrow p_2} & \cdots & Pr_{p_L \rightarrow p_2} \\
\vdots & \vdots & \ddots & \vdots \\
Pr_{p_1 \rightarrow p_L} & Pr_{p_2 \rightarrow p_L} & \cdots & Pr_{p_L \rightarrow p_L}
\end{bmatrix}
\]

Computing a Corrected Family Size Array \( \langle N \rangle \)

Some allocations of the \( p_i \) real offspring may have been wrongly allocated to other pairs of the focal set. The family size of \( p_i \) estimated directly from the allocations \( \langle n_i \rangle \) is in fact the sum of correct allocations of the \( N_i \) true offspring from the \( p_i \) pair to itself and of offspring generated by other pairs but wrongly allocated to \( p_i \). Therefore, one may write

\[
n_i = Pr_{p_i \rightarrow p_1} N_1 + Pr_{p_i \rightarrow p_2} N_2 + \cdots + Pr_{p_i \rightarrow p_L} N_L,
\]

where array \( \langle N \rangle = <N_1, N_2, \ldots, N_n, \ldots, N_n> \) stands for the true numbers of offspring belonging to pairs \( p_1, p_2, \ldots, p_n, \ldots, p_L \) of the focal set.

All specific equations, one for each of \( n_1, n_2, \ldots, n_n, \ldots, n_n \), together form an equation system that may be expressed in matrix form as follows:

\[
\begin{bmatrix}
Pr_{p_1 \rightarrow p_1} & Pr_{p_2 \rightarrow p_1} & \cdots & Pr_{p_L \rightarrow p_1} \\
Pr_{p_1 \rightarrow p_2} & Pr_{p_2 \rightarrow p_2} & \cdots & Pr_{p_L \rightarrow p_2} \\
\vdots & \vdots & \ddots & \vdots \\
Pr_{p_1 \rightarrow p_L} & Pr_{p_2 \rightarrow p_L} & \cdots & Pr_{p_L \rightarrow p_L}
\end{bmatrix}
\begin{bmatrix}
N_1 \\
N_2 \\
\vdots \\
N_L
\end{bmatrix}
= \begin{bmatrix}
n_1 \\
n_2 \\
\vdots \\
n_L
\end{bmatrix}
\]

This can also be written in the more compact format:

\[
M_F N_T = n_T,
\]

where the reader will recollect that \( M_F \) stands for the focal allocation matrix and that the \( T \) superscripts indicate transposition to column vector form.

Solving the above system for the \( \langle N \rangle \) array, one obtains estimates of the true numbers of offspring for each pair of the focal set.

Comments on the Focal Set Method

- The focal allocation matrix \( M_F \) is built from simulating allocations to the whole set (not just the focal set) of potential parents because this better represents the allocation conditions and, therefore, should produce more accurate estimates of the allocation proportions for each parental pair of the focal set.

- The focal allocation matrix \( M_F \) involves allocating artificial offspring only from parental pairs that belong to the focal set. In theory, if the \( M_F \) were extended to all potential parental pairs to become a full allocation matrix \( M_A \), the family size estimations (\( N \)) could be more accurate. However, under the very conditions that we are considering here, where most pairs have zero offspring, complete parental sets will usually be very large compared with corresponding focal sets and the ensuing calculations much more intensive. The only potential drawback associated with the focal set reduction is the possibility that some potential pairs with no allocated offspring have actually contributed offspring all of which wrongly allocated to one or several members of the focal set. Although this is certainly a theoretical possibility, it is an unlikely event. In fact, this would mean that a pair, say \( p_{k_L} \), has bred some offspring and that none of which was allocated to itself but to one or several similar pairs that, in turn, contributed none of their own offspring to the \( p_{k_L} \) allocations. Note that the simulations used to validate the focal set simulation method do not a priori preclude this possibility so that some spillover of offspring generated outside the focal set will automatically be reflected in the results.

- Solving for the corrected family size array \( \langle N \rangle \) involves standard linear algebra procedures. For instance, this can
be easily done by first inverting the focal allocation matrix \( \mathbf{M}_F \) and then multiplying the allocation family size array \( <\mathbf{n}> \) by the inverted matrix because
\[
\mathbf{M}_F^{-1}\mathbf{M}_F\mathbf{n}^T = \mathbf{M}_F^{-1}\mathbf{n}^T \quad \text{and} \quad \mathbf{N}^T = \mathbf{M}_F^{-1}\mathbf{n}^T.
\]
Note that matrix inversion and multiplication are predefined functions within several widespread software like Excel.

- The solution for the family size array \( <\mathbf{N}> \) will usually involve fractional (non-integer) numbers. Fractional numbers should be rounded up to the closest integer. Negative family size values obtained from the focal set method might occur when pairs are filled mainly by offspring that are erroneously allocated to those “empty” pairs. In this case, the empty pairs should be removed and a second round of allocations followed by a second application of the focal method should be run.

**Funding**

French Biodiversity Institute (contract no. 2001/583/P00202); French Bureau of Genetic Resources (contract SRP-04 B/2003—CV03000029); Sansouire Foundation; the Tolmin Angling Association.

**Acknowledgments**

We thank Rasmus Nielsen and Dorte Bekkevold for comments on earlier versions of the manuscript. We would like to thank René Guyomard and Lars-Erik Holm for their generous donation of primer aliquots for testing. This is an ISEM-2007-151 publication.

**References**


Received July 5, 2007
Accepted November 20, 2007

Corresponding Editor: William Modi