Chapter 8
Individual-based Genotype Methods in Aquaculture

Pierre Duchesne and Louis Bernatchez

Introduction

DNA marker technologies have revolutionized the way aquaculture genetics research is being conducted (Liu and Cordes 2004). Early on, most applications of molecular genetics in aquaculture relied on the estimation of demographic parameters of diversity and differentiation that were derived from averaging the genetic composition over populations or stocks. It has been recognized for nearly 25 years, however, that further knowledge of relevance for stock management and production may be obtained from the analysis of individual-based genotypic information (Smouse et al. 1982). The blooming development of new genetic markers over the last decade, namely variable number of tandem repeat loci (especially microsatellites), Amplified Fragment Length Polymorphism (AFLP), and Single Nucleotide Polymorphism (SNP) have revived a major interest in studies based on the definition of individual multilocus genotypes, and opened exciting avenues of research and applications. Basically, studies of relevance for aquaculture and based on the analysis of individual multilocus genotypes can be grouped into three broad categories of applications: parentage (including kinship), group allocation, and hybrid detection.

Parental allocation studies necessitate the assessment of 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 Wilson and Ferguson 2002, Jones and Ardren 2003). Parental allocation improves the efficiency of selective breeding programs in many ways, namely the following:

- establishing selected strains without having to keep families in separate tanks (Wilson and Ferguson 2002)
- investigating parent to offspring transmission of illness or parasitism
- assessing fertilization success (Selvamani et al. 2001)
- measuring reproductive success variance among breeders (Jackson et al. 2003)
- avoiding mating between closely related individuals and thus minimizing inbreeding (Ferguson and Danzmann 1998, Jackson et al. 2003, Norris et al. 2000)
- improving heritability estimates of desirable traits (Ferguson and Danzmann 1998, Vandeputte et al. 2004)
- allowing a higher rate of genetic improvement because it becomes possible to identify the progeny of parents with desirable or undesirable characteristics (Wilson and Ferguson 2002).

Studies of group allocation (also called "assignment methods") typically imply the determination of population membership of single individuals (Manel et al. 2005).

This consists of assigning an individual to the population in which its multilocus genotype has the highest probability of occurring. Such estimation may be relevant to more precisely quantify gene flow and the degree of differentiation between stocks, quantifying the admixture proportion of different stocks in a sample of individuals of unknown origin such as wild versus cultured (Miggiano et al. 2005), or enhancing traceability for trade control purposes in animals and products, and thus allow consumers to obtain information on the origin and the production chain of food products (Liu and Cordes 2004, Hayes et al. 2005).

In aquaculture, genetic group allocation may be used to identify species or strain membership of specimens. Such identifications are useful both at the input and output end of production facilities. For instance, controlling for possible admixture in purebred populations can be done in an objective fashion when based on solid genetic data. Allocation can also reveal proportions of wild versus cultivated specimens in the marketplace or in a natural system undergoing invasion by farmed escapees or deliberately stocked by a nonnative strain. Coarse traceability can also be performed when distinct production organizations are associated with distinct strains.

Hybridization between or within species is both a common natural phenomenon and the consequence of mixing due to human-related activities, including aquaculture, and stocking of domesticated fish (Congiu et al. 2001, Vaha and Primmer 2006). Identification of hybrid individuals is often a necessary first step in the implementation of management strategies, such as breeding or translocation programs for threatened species since, it allows the removal of morphologically indistinguishable hybrid individuals from the wild population or the identification of indigenous individuals for breeding programs (Hansen 2002, Manel et al. 2005, Vaha and Primmer 2006). Early identification of hybrids may help reduce the impact of introgression between cultured and wild fish (Morizot et al. 1991, Young et al. 2001). Also identification of hybrids can impact trade by detecting hybrid production labeled as purebred, for example, sturgeon caviar (Congiu et al. 2001).

Because these issues have been treated in several recent reviews, our intent here is not to address the suitability of various molecular techniques, nor is it meant to review the empirical applications of individual-based genotype analyses. We do not wish to provide an exhaustive guide or detailed treatment to the existing analytical methods or related computer software packages. Instead, our main goal is to explain the basics of statistical principles and applications of specific methods that we have developed and applied in our laboratory over the recent years. In an attempt to render the chapter content easily accessible to the nonstatistician scientists, we have deliberately opted for verbal explanations rather than relying on the treatment of mathematical complexity and equations.

Parental Allocation

Definition and General Principles

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. In some contexts, it is known in advance that the genotypes of all the parents

involved in the generation of the set of offspring are included in the collection of the putative parental genotypes. If that is the case, then the allocation system, comprising parental and offspring genotypes, is said to be closed. When some parental genotypes are missing, the allocation system is said to be open. Despite obvious similarities, the allocation problems for closed and open systems turn out to be quite different, the latter being more complex.

The two main factors affecting the performance of a parental allocation process are the number of potential parental pairs and the genetic contents of the genotypes. Performance decreases with the size of the parental set while it increases with genetic contents. Other important performance factors are the relatedness level of the parental set, accuracy of genotype scoring, and sexing of potential parents. Closely related potential parents tend to be more similar than unrelated parents resulting in a higher probability of misidentification. Whenever possible, it is generally advantageous to sex parents since this reduces by at least one half the number of potential parental pairs to be considered (Wilson and Ferguson 2002).

Markers

In theory, any type of marker can be used for performing parentage allocation. However, microsatellites are currently the most popular because of their potential for high variability even among individuals of the same strain (Liu and Cordes 2004). For instance, using eight highly variable microsatellite markers, Norris and others (2000) correctly allocated 95% of offspring from more than 12,000 potential parental pairs. Generally, codominant markers are best suited for parental allocation since allele transmission from parent to offspring is never masked by allelic dominance. The use of diploid codominant markers will be assumed throughout the following discussion.

Scoring Errors: Effects and Modeling

Here a scoring (transmission) error is defined as the result of mistaking a specific allele for another one. While scoring microsatellites, it is estimated that errors occur at a rate of 0.5 to 3%. Erroneous allocations due to scoring errors are not likely. The main negative effect of erroneous allele scores is possible loss of correct parental allocations. The probability that a genotype contains at least one scoring error increases rapidly with number of loci. Therefore, as one increases the information genetic contents by adding extra loci, one is also increasing the proportion of erroneous genotypes and thus leading to a larger proportion of incorrect allocations. This dilemma can be broken by integrating an appropriate scoring error model within the allocation process.

Within closed allocation systems, the negative effect of scoring errors can be completely neutralized by allowing a small nonzero probability estimate to the scoring of allele X as any distinct allele Y. The uniform error model (see definition below) provides such an error-catching mechanism. The transmission error probability (ε) estimate does not have to be accurate; estimates of say 1%, 2%, and 3% for ε will have the same effect on the allocation output.

The transmission error probability can be distributed in several ways over (erroneous) alleles. However, it is well known that scoring errors usually involve alleles that

are close to the true allele. This information can be fed into error modeling through the following formalization. Suppose the parental allele X is referred to as the focal allele. Then the distance between any allele Y and X can be measured in terms of number of offsets, that is, the difference between Y and X divided by the smallest allelic distance between any two alleles found in the locus (Figure 8.1). For instance, if a locus is of type tetra (nucleotide) then Y = 172 is −2 offsets away from X = 180.

The uniform error model is the simplest error model. It distributes e uniformly over all possible nonfocal alleles. Restricted error models distribute e over close neighbors of the focal allele. The examples of a ±1 offset model and a ±2 offset model are shown in Table 8.1.

Allocation Methods in Closed Systems

Basically, parental allocations can be based either on likelihood or on exclusion.

Likelihood

Given an offspring, the likelihood of a specific parental pair is essentially a measure of the probability that this pair has generated the offspring. There are three possible outputs associated with the allocation of a particular offspring. When only one parental pair has the largest likelihood, the offspring is allocated to the parental pair with the

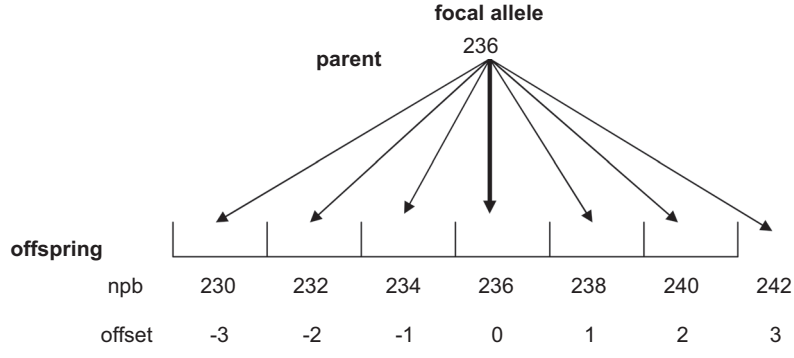

Figure 8.1. Measuring a transmission error in offset units. The distance between any allele Y and X is measured in terms of number of offsets (i.e., the difference between Y and X divided by the unit distance [smallest allelic distance between any two alleles found in the locus]).

Table 8.1. The examples of a *±1 offset model* and a *±2 offset model.*

−2 offsets	*−1 offset*	*0 offset = focal allele*	*+1 offset*	*+2 offsets*
0.002	0.01	0.98	0.01	0.002
	0.008	0.98	0.008	

largest likelihood. When several parental pairs share the largest (nonzero) likelihood, the offspring is not allocated but the output is scored as ambiguous. When all parental pairs have zero likelihood, the offspring is not allocated and the output is scored as null. Although most allocation programs do not distinguish explicitly between ambiguous and null outputs (both are scored as nonallocations), this distinction allows the computation of three system-based allocation statistics: proportions of offspring that have been scored as allocated, ambiguous, and null. These statistics turn out to be very useful in the context of the overall assessment and subsequent improvement of an allocation system. For instance, any proportion of ambiguity, except negligible, is indicative of a lack of resolution (i.e., insufficient genetic contents). In such cases, the only cure is to add one or several loci to the existing set until all ambiguity disappears.

Within closed systems, allocations should usually be performed with the uniform error model since it can absorb all kinds of errors including those generated by null alleles scored at any offset distance from the focal allele. The only drawback of the uniform model is that it may, though not necessarily, increase the proportion of offspring classified as ambiguous. This can be corrected by using a nonuniform error model but more efficiently by adding one or several loci.

Exclusion

Exclusion-based allocation is based on the idea that as information accumulates, only real parents remain after all other potential parents have turned out to be impossible candidates. Exclusion-based allocation should generally not be used in closed systems since it takes far more genetic information to exclude the set of false parents than to find the most likely pair. Unless otherwise stated, we will thereafter refer to likelihood-based allocation. Exclusion will be further discussed in the context of open system allocations.

Breeding Designs (Closed Systems)

Sometimes the offspring from blocks of breeders are put together in a single tank. Block matings generally reduce the total number of potential parental pairs as compared with allowing all adults to breed together. This reduction could translate subsequently into a reduced number of loci necessary to reach a satisfactory level of allocation correctness. Provision has been made in the last version of Package for the Analysis of Parental Allocation (PAPA) software (Duchesne et al. 2002) to allow definition of blocks of breeders reflecting breeding designs in aquaculture settings. Distinct blocks may share specimens and they may be sexed or unsexed.

Validation of Allocations in Closed Systems

Allocation to a parental pair may not always be correct. Ideally one should be able to test the correctness rate (CR), that is, the proportion of correct allocations over all

allocations, not all offspring, by checking the allocations against empirical evidence. However, under most circumstances, establishing parental connections through direct observation even in hatchery fish can prove very difficult and expensive. It is therefore customary to use simulations to estimate correctness rates. Also, simulations are useful when it comes to deciding on a set of loci sufficiently informative to reach a satisfying level of CR.

Preparental and Parental Simulations

Basically there are two types of parental allocation simulation procedures. One procedure (preparental) generates artificial parental genotypes from allelic frequencies (estimated from samples) and then artificial offspring from these parents (Figure 8.2). Another procedure (parental) uses the genotypes of real, collected parents. Preparental simulations are useful to decide on a minimal set of loci to attain the desired correctness rate even before parents and offspring have been collected. Preliminary choice of a sufficient set of loci can save lab work and resources. However, preparental simulations tend to underestimate minimal genetic information contents mainly because it generates sets of totally unrelated parents. Sets of real parents, especially when drawn from a hatchery population, may contain several subsets of highly related specimens. Therefore, it might be safer to add an extra locus to the minimal set found from preparental simulations especially when the targeted correctness level is barely reached.

To estimate correctness rates more precisely, parental simulations should be run when the set of collected parents has been genotyped. Parental simulations are not biased by the relatedness structure of the parental set.

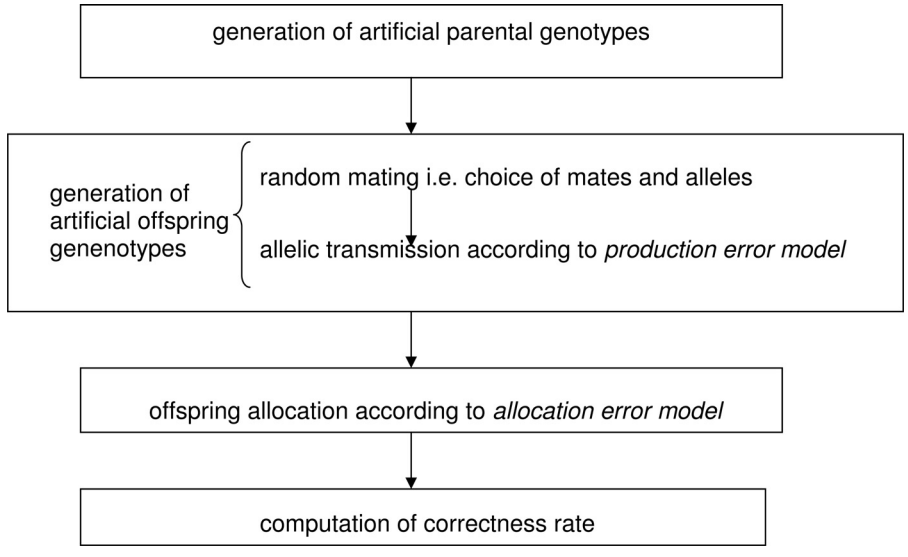

Figure 8.2. Preparental simulator procedure. The preparental procedure generates artificial parental genotypes from allelic frequencies (estimated from samples) and then artificial offspring from these parents. The parental procedure is similar except that it uses the genotypes of real, collected parents.

Production and Allocation Error Models

To estimate correctness rates more realistically, the production of artificial offspring during simulations has to mimic scoring errors. Therefore, there is a need for a production error model. The production error probabilities associated with various numbers of offsets do not have to be very accurate although they do impact on correctness rate estimations. After artificial offspring have been generated, they are processed for allocation. As with true offspring, an allocation error model is used to capture scoring errors. Ideally one should be able to define production and allocation error models separately. Allocation error models in simulations should generally be the same as the one used in allocating real offspring.

Likelihood and Exclusion Methods in Open Systems

Likelihood

Allocation in open systems poses a double problem (i.e., identify true parents that belong to the collected parental set and identify uncollected parents as uncollected). Likelihood-based allocation can be very efficient in solving the collected parent problem but is liable to mistake an uncollected parent for a collected one (i.e., overallocate). Overallocation increases sharply with the proportion of uncollected parents. With more uncollected parents, there is a higher probability that collected specimens are sufficiently similar to uncollected parents to become likely candidates for (erroneous) allocation. This problem is more acute with methods allowing a nonzero probability for any kind of scoring error, which translates into nonzero likelihood for all possible parental-offspring genotype combinations. The overallocation probability can only be assessed when a reasonably accurate estimate of the missing part of the parental set is available (Wilson and Ferguson 2002). Unfortunately, likelihood-based allocation cannot provide such an estimate on the basis of the available genotypes. In short, likelihood methods in open systems tend toward overallocation, the extent of which cannot be safely estimated without a (generally lacking) reliable estimate of the uncollected portion of the parental set.

Exclusion

The drawbacks of likelihood-based methods in open systems have led some researchers to resort to the exclusion allocation method. This method essentially compares the genotype of each potential parent with that of the offspring. Parental genotypes are excluded as soon as both offspring alleles are absent on a single locus of the parental genotype. In addition, no more than two nonexcluded parental genotypes have to remain for the allocation to be performed. The idea is that, given enough loci, nonparental collected specimens will eventually be excluded on at least one locus.

The exclusion method has several drawbacks. It is very costly in terms of genetic information since most excluded candidates would have been discarded on account of

low likelihood based on much less powerful sets of loci. Since scoring errors are more numerous with each additional locus (Jones and Ardren 2003), it is plausible that a substantial number of genotypes will contain at least one error. Such errors are very likely to provoke the loss of one or several allocations especially when parental genotypes are erroneous. Some researchers have suggested tolerance for mismatches not exceeding a predetermined number. However, mismatches may also come from a truly nonparental genotype. Therefore, this less stringent version of exclusion, while it does reduce the probability of erroneous exclusion, also increases the probability of retaining nonparental combinations (i.e., erroneous allocations). This tradeoff between two types of errors cannot be easily assessed in the absence of a sound estimate for the missing proportion of uncollected parents. Therefore, the choice of a number of tolerated mismatches is largely arbitrary.

To alleviate the stringency of the exclusion method resulting in overexclusion, another approach is sometimes used that includes rescoring a nearly perfectly matching genotype. The idea is to see if some scoring error might not be the reason for missing an allocation by so little. Although this method does make some sense, it is prone to self persuasion and is certainly not amenable to correctness analysis. Briefly stated, exclusion methods tend to miss sizable numbers of true parents and do not lend themselves to rigorous evaluations of correctness rates. They would be efficient if based on a very informative set of loci and extremely accurate genotypes. These two conditions are not generally met except in some forensic contexts.

The PASOS Approach (Open Systems)

Likelihood-based methods lean toward overallocation whereas exclusion methods tend to overexclude (i.e., eliminate true parents). The PASOS software (Duchesne et al. 2005) uses a mixed approach by first picking up the most likely parental pair(s) among all potential pairs based on a uniform scoring error model that ensures that at least one most likely pair is listed. When several most likely pairs are found, the first one in the list is retained. Then an extended exclusion method is applied to the two genotypes of the retained most likely parental pair.

Extended Exclusion Method

The extended exclusion method used by PASOS compares each of the locus genotypes of the two putative parents together with that of the offspring. From these three genotypes, a transmission scenario (Figure 8.3A) is built that associates each offspring allele to a parental allele. Such scenarios are built from a set of rules that aims at restoring the most probable allelic transmission pattern, taking the three genotypes together. Once the two most likely parent-to-offspring allele pairs have been determined, the distance in offset units is computed for each pair. Any allelic distance exceeding the maximum offset tolerance (MOT) specified by the user provokes the exclusion of the corresponding putative parent (Figure 8.3B). Therefore, there may be zero, one, or two parents excluded at each locus. It suffices that the offset tolerance be exceeded on a single locus for the putative parent, relative to the offspring currently processed, to be discarded.

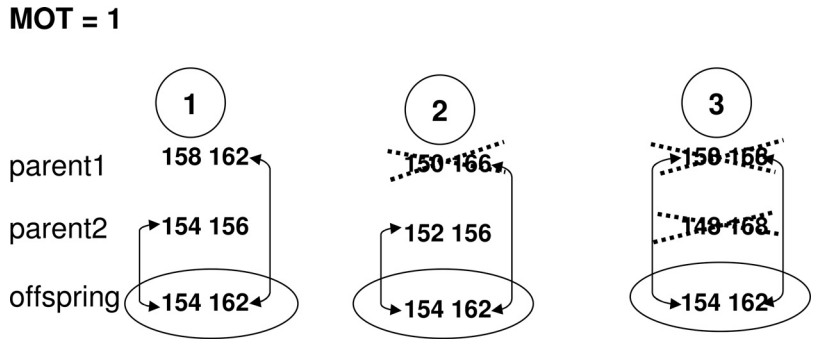

Figure 8.3A. Allelic transmission scenarios: Allelic transmission scenarios are built from a set of rules that aims at restoring the most probable allelic transmission pattern, taking the two parental and the offspring genotypes simultaneously into account.

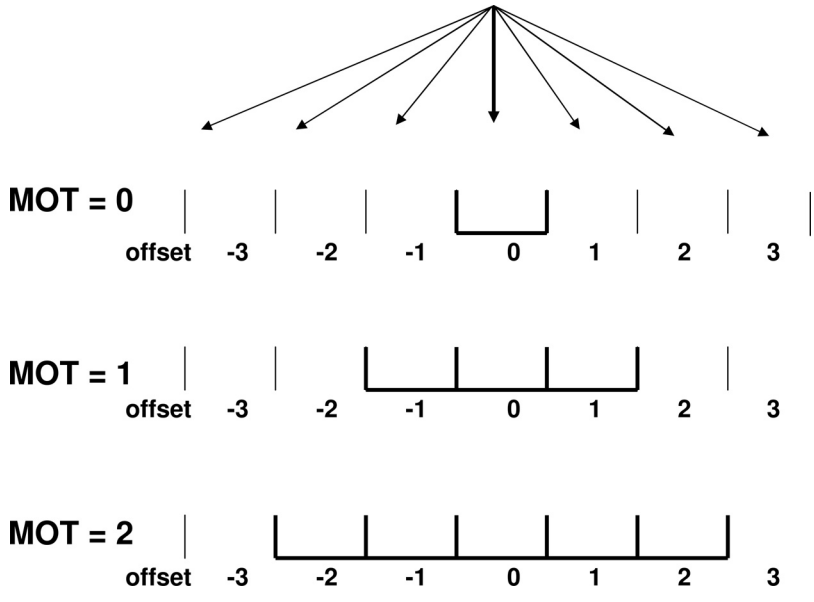

Figure 8.3B. Tolerance net as defined by MOT: Any allelic distance exceeding the maximum offset tolerance (MOT) specified by the user provokes the exclusion of the corresponding putative parent.

Rationale

The two-step allocation approach implemented in PASOS is based on the following rationale. If the two real parents of an offspring belong to the collected set of potential parents, the probability that they will be selected during the likelihood phase will increase with genetic information contents (i.e., with number of loci). If they have been genotyped with scoring errors within the bounds of the maximum offset tolerance,

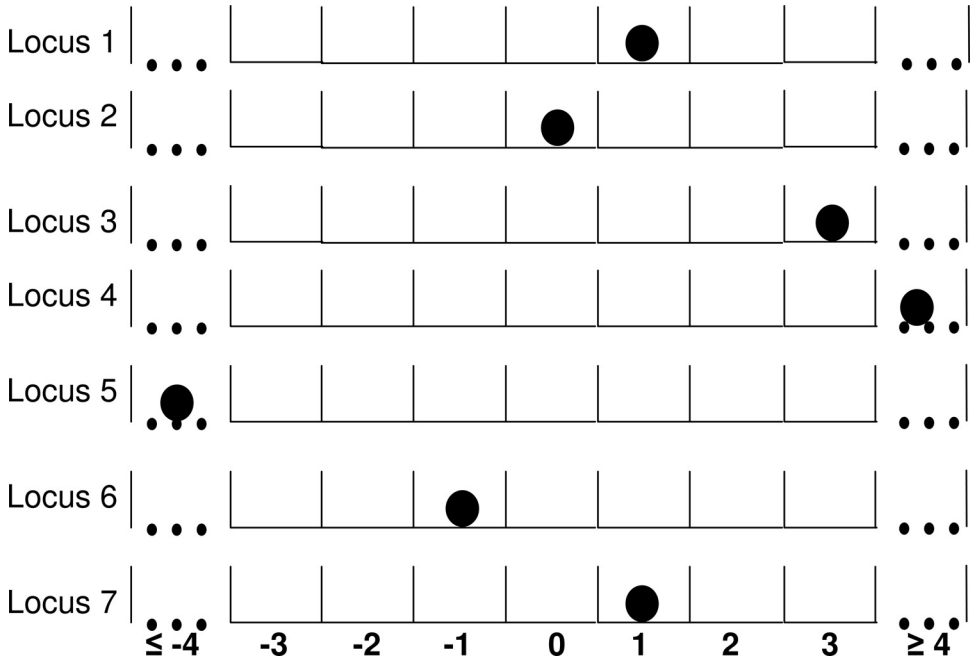

Figure 8.4. Extended exclusion of a false parent. The probability that a nonparental member of the most likely parental pair be eliminated increases with number of loci.

then they will most probably not be discarded during the exclusion phase. If only one parent belongs to the parental set, it will probably be part of each of the most likely pair(s) and thus of the first pair listed. The probability that the nonparental member of the most likely pair be eliminated during the exclusion phase will increase with number of loci (Figure 8.4). If none of the two parents belongs to the set of collected parents, then the most likely pair will contain false parents both of which will eventually be discarded as the number of loci increases.

Sequence Allocation (Allocation) and Proportion of Missing Parents

When PASOS is run sequentially with one, two, three, etc., loci from the allocation set, it makes less and less allocations and eventually reaches a stable or near stable proportion of allocations (Figure 8.5). This happens when false parents have been purged by the extended exclusion procedure. The remaining proportion of allocations may then be taken as an estimate of the proportion of missing parents. The precision of the latter estimate depends on the assumption that the collected parental set comprises specimens that have truly participated, no matter how successfully, in the breeding event at the origin of the offspring sample. If the parental set is inflated with individuals not involved in reproductive events, then the number of missing parents will likely be overestimated. Clearly, precision of the estimate should increase with the size of the offspring sample. The estimated number of missing parents must be fed into simulation runs to obtain estimates of the correction rates.

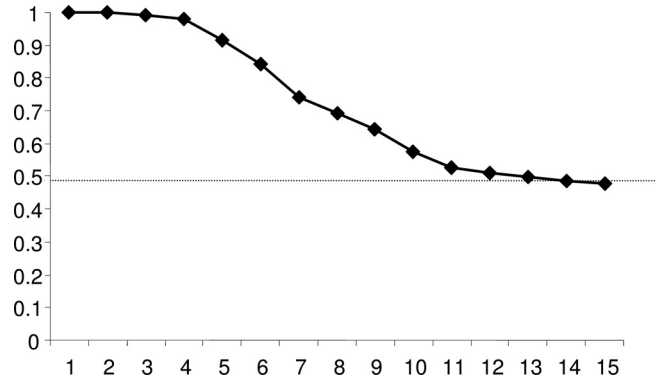

Figure 8.5. Sequence allocation curve. When PASOS is run sequentially with one, two, three. . . loci from the allocation set, it makes less and less allocations and eventually reaches a stable or near stable proportion of allocations which the user may then use to estimate the number of uncollected parents.

Automatic sequence allocation (i.e., with one, two, three or more loci) is implemented in PASOS.

Due to its use of restricted error modeling, PASOS should only be used when scoring is of good quality (i.e., does not generally exceed two offsets from focal alleles). Also, the set of loci should be tested for the presence of null alleles and all loci suspected of containing null alleles should be dropped.

Validation of Allocations in Open Systems

The estimation of the correctness rate within any open allocation system depends heavily on the estimated number of missing parents. In fact, the larger the set of missing parents, the higher the probability that some of their offspring will be mistaken for offspring from the collected parental set. Unfortunately, the number of missing parents often is difficult to estimate under most settings and so estimates have typically been guessed in the past.

However, recent developments in allocation techniques that combine likelihood with exclusion approaches (PASOS) now make it possible to obtain reliable estimates of the missing part of the parental set. Once the sequence allocation of the sample of real offspring has produced a (nearly) stable allocation rate curve, an estimate of the proportion of missing parents is available. The latter can then be fed into parental simulations for obtaining a sound estimate of the correctness rate associated with the specific allocation system.

Preparental simulations should be run whenever possible to find minimal sets of loci. Since the missing part of the parental set cannot be estimated genetically prior to parent collecting, care should be taken to use both optimistic and pessimistic scenarios corresponding to lower and higher proportions of missing parents, respectively. Again, minimal sets of loci should preferably be complemented by an extra locus in case the real parental set comprises highly related specimens.

Features to Look for in Parentage Allocation Programs

In closed as well as open allocation systems, programs should provide simulation facilities. Simulations are usually the only way to obtain a sound estimate of the correctness rate or accuracy of the system (i.e., the proportion of correct allocations among all allocations). In addition, one should be able to run the simulator on both preparental and parental modes. One should be able to run programs either with sexed or unsexed parental sets since sexing in fish cannot always be done easily and reliably.

Closed Systems

In closed systems, programs should provide distinct statistics for ambiguous and null outputs. The proportion of ambiguous outputs is a direct measure of the capacity of the set of loci to perform the allocation task under way. An error model that provides nonzero probability for any possible scoring error such as the uniform error model should suffice under most circumstances. However, with reliable scoring and absence of null alleles, the use of a restricted error model allowing for a limited number of error offsets could save on the number of loci without significantly reducing the number of allocations. A mechanism for defining blocks of breeders reflecting breeding designs in aquaculture settings is desirable. Block definition can increase resolution power of a set of loci and reduce the probability of incorrect allocations.

Open Systems

In open systems, uniform error modeling can lead to overallocation since parent-offspring mismatches can also originate from an incorrect allocation. On the other hand, zero error tolerance is very likely to provoke losses of allocations especially as the number of loci is increased. Restricted error modeling is a means to distinguish between scoring errors and erroneous allocations without dropping a significant proportion of true parents. Restricted error modeling is currently implemented in PASOS. The most important features for parental allocation programs are described in Figure 8.6.

Some Available Programs

Some of the currently available programs with respective allocation methods follow:

- CERVUS (Marshall et al. 1998) (likelihood)
- FAMOZ (Gerber et al. 2002) (likelihood)
- KINSHIP (Goodnight and Queller (1999) (exclusion), (Danzmann 1997) (exclusion)
- NEWPAT (Wilmer et al. 1999) (exclusion)
- PAPA (Duchesne et al. 2002) (likelihood/closed systems)
- PARENTE (Cercueil et al. 2002) (likelihood)
- PASOS (Duchesne et al. 2005) (likelihood + extended exclusion/open systems)

All of these freely available programs can be downloaded at http://www.bio.ulaval.ca/louisbernatchez/links.htm.

General

both pre-parental and parental simulations are available

simulations and allocations may be run with sexed or unsexed parental sets

Closed systems

distinction is made between *null* and *ambiguous* non-allocation statistics

scoring error may be distributed over all non-focal alleles e.g. uniformly

parental files can be structured according to block mating designs

restricted error models may be user-defined

Open systems

restricted error models are available and user-defined

a means to estimate the number of uncollected parents is provided

Figure 8.6. A list of most important features for parental allocation programs.

Group Allocation (Species, Population, or Strain Identification)

Definition and General Principles

Species, population, or strain identification of individuals on the basis of genetic data is technically the same and will thereafter be referred to as group allocation. Only those allocation situations will be considered where each purebred group has been sampled so that fairly accurate estimates of allelic frequencies for each genotyped locus and each purebred group are available (baseline samples).

Some recent developments aim at allocating individuals from mixed samples without prior sampling of group purebreds. Those so-called clustering techniques essentially tend to partition a given mixed sample into subsamples to minimize (or maximize) some statistic associated with population structuring (e.g., linkage disequilibrium). Allocation from good baseline samples produces verifiable results within a small fraction of the computation time required from clustering methods. Moreover, currently used clustering methods tend to perform poorly when group differentiation is weak (Waples and Gaggiotti 2006), a very serious handicap when it comes to strain identification. Finally, they do not provide ad hoc means to estimate the accuracy of their allocations and involve considerable uncertainty (Manel et al. 2005). Given the above drawbacks of clustering methods, they will not be discussed any further since baseline samples are available for most group allocation tasks within aquaculture settings.

The idea underlying group allocation of an individual genotype (G) is rather simple. In its simplest version, the probability (likelihood) that G could be found within a group is computed for each possible group and then G is allocated to the group with highest probability. Since such probabilities are often very small, they are usually expressed in log10 format and comparisons between two populations as log-likelihood ratios. For example, if G is 1,000 times more likely to be found within

population A than it is within population B, the log-likelihood ratio of A relative to B is equal to three.

Within a given allocation task, a minimal log-likelihood ratio (threshold) between the most likely and the next most likely group is defined. If the threshold is not reached for G, it is simply not allocated and classified as nonallocated. For instance, a log-likelihood threshold of two would mean that no individual genotype should be allocated if it is not at least 100 times more probable within the most probable group. The log-likelihood threshold turns out to be an important allocation parameter. Generally, raising the threshold increases the probability of allocating correctly (accuracy) but decreases the number of genotypes being allocated (allocation rate). Care should be taken to choose an appropriate threshold for the task under way.

Another important aspect of group allocation is the question of ghost groups (i.e., groups that have not been sampled as purebreds since they have not yet been identified but which may be represented within the sample of individuals to be allocated). Ghost groups are much more likely when allocations involve wild populations. When it is suspected that ghost groups might exist, one should test whether G might not belong to such an external yet undefined group. This can be done through an exclusion procedure based on membership P values computed from simulations. (See the Simulations section.)

Markers

As in parentage allocation, any type of marker (RFLP, RAPD, AFLP, microsatellite) can be used for performing group allocations. However, very high polymorphism (number of alleles/ locus > 10) does not add substantial allocation resolution when compared to less variable loci. Here, the most important characteristic of a set of loci is sheer number (Ferguson and Danzmann 1998, Bernatchez and Duchesne 2000, Hayes et al. 2005). Therefore, when it comes to distinguishing several weakly differentiated groups (e.g., strains), markers available in virtually unlimited numbers are the best candidates even when each locus has low information content. For such heavy allocation tasks, AFLP markers are currently the most appropriate choice except when a sufficient set of microsatellites already exists (Campbell et al. 2003).

Scoring and Sampling Errors

With dominant markers such as AFLP, allele should be taken as an equivalent for presence/absence in the following discussion. Generally speaking, scoring errors within their usual range (0.5 to 3%) have little impact on group allocation. However, special care should be taken when scoring purebred samples especially when small (>20). As a rule of thumb, purebred samples should contain at least 20, but preferably 30, specimens to obtain reasonably accurate frequency estimates (Ruzzante 1998). Smaller samples might still be used especially when dealing with highly differentiated groups. When using highly polymorphic microsatellite loci with large numbers (>15) of low frequency alleles, sample sizes should be increased accordingly (e.g., to 50 specimens). Note that the low frequency of an allele can suddenly double following sampling of a

single extra copy (Roques et al. 1999). To obtain truly representative purebred samples, sampling should be done as randomly as possible. In particular, overrepresentation of specific families should be avoided.

A special sampling problem arises when some allele is totally absent from one or several purebred samples while present in other purebred samples or the (mixed) sample to be allocated. Customarily, the frequency of a missing allele within a purebred sample was estimated at 1/(N+1) (N = number of scored alleles within sample). This amounts to the expectation that the next allele would be the missing one (maybe-next-allele formula). Another approach consists of fixing the missing allele frequency at some user-defined low value (e.g., 0.01). Practically, missing allele frequency estimates have little impact on the result of an allocation task. If one favors the fixed low value approach, then this value may be seen as an allocation parameter and its value may be chosen to maximize the correct reallocation rate.

Validation of Group Allocations

The accuracy of group allocations, that is, the estimated proportion of correct allocations over all allocations (excluding non-allocated specimens), can be assessed through reallocation and simulation procedures (Figure 8.7).

Reallocation

The reallocation procedure allocates the purebred specimens among the candidate groups as if their group membership were unknown. The latter condition means that each time a purebred specimen is (re-)allocated, the allelic frequencies of its group are recalculated as if it did not belong. This precaution aims at eliminating the bias resulting from the specimen actually weighing on frequency estimates and, as a consequence, artificially increasing the probability of being allocated to its proper group. These as if frequency recalculations are usually referred to as the leave-one-out procedure.

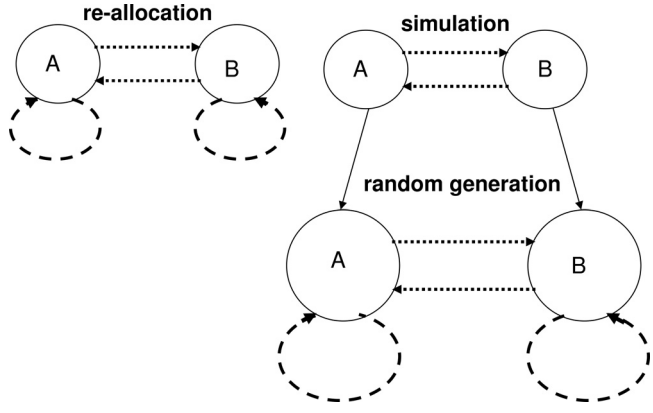

Figure 8.7. Validation procedures in group allocation. The reallocation procedure allocates the purebred specimens among the candidate groups as if their group membership were unknown. To estimate accuracy from simulations, artificial specimens are generated randomly, based on the allelic frequencies derived from purebred samples.

Reallocation of purebreds is usually a very reliable way of estimating accuracy. One important advantage of reallocation oversimulations is that it takes scoring errors automatically into account. On the other hand, accuracy estimates from reallocation may be biased upward when purebred samples include highly inbred specimens (e.g., full and half-siblings). Thus, the quality of accuracy estimates from reallocation is somewhat sensitive to the quality of purebred group samples. Low reallocation rates may result from very poor scoring, a lack of resolution due to poor genetic content relative to group differentiation, or even from an absence of real differentiation (i.e., from samples not actually representing distinct biological entities).

Simulations

Estimations of accuracy can also be obtained from simulations. Artificial specimens are generated randomly, based on the allelic frequencies derived from purebred samples. The simulators currently built into population (group) allocation programs do not allow mimicking of scoring errors. Consequently, accuracy may sometimes be slightly overestimated from simulations since scoring errors do increase the probability of misallocating real genotypes. One important advantage of simulations over reallocation is their potential for spanning a very large range (e.g., tens of thousands of possible genotypes from each group). Therefore, genotypes from prospective mixed samples get a more complete coverage by simulations than they do from reallocation.

Besides accuracy estimations, simulations are sometimes used to obtain likelihood distributions from each purebred sample. Each group likelihood distribution is obtained by producing a large number of artificial genotypes, based on the group allelic distributions, and then the likelihoods associated with the genotypes. Thereafter, the group-specific likelihood distributions may be used to produce a group membership P value for each genotype of a mixed sample. Some allocation programs actually use group membership P values by excluding each candidate groups with membership P value lower than a predefined threshold. When the allocation procedure is based on likelihood ratios, membership P values can still be useful to detect ghost groups: when membership P values are very low (e.g., >0.001) for all potential groups considered, the presence of at least one ghost group may be suspected.

Another usage of simulations is the adjustment of the likelihood ratio allocation threshold. Sometimes a proportion of artificial genotypes are misallocated indicating that there is a nonnegligible probability that real genotypes may also be misallocated. This problem can be solved to a large extent by raising the likelihood ratio allocation threshold until misallocation of simulated genotypes vanishes. Note however that this will generally be associated with a rise in the proportion of nonallocated real and simulated genotypes.

Reallocation Versus Simulation Accuracy Estimates

Accuracy estimates from reallocation and simulations should be close. However, if the estimated accuracy from reallocation is substantially lower than that from simulations, it is probably due to unusually numerous scoring errors. On the other hand, higher accuracy estimates from reallocation could reflect highly inbred portions of samples (e.g., families).

Features to Look for in Group Allocation Programs

Reallocation of purebred genotypes, allocation of mixed samples, and simulations are the three basic procedures that should be provided by group allocation programs. The leave-one-out procedure should be used in reallocating purebred samples.

The log-likelihood ratio allocation threshold should be user defined. Calculation of membership P values for each genotype should be possible even when the allocation procedure is based on likelihood ratio values (i.e., not on low P value exclusion). Membership P values are especially important when there are grounds to believe that some members of the mixed sample may come from a ghost group. Group log-likelihoods for each real genotype should be made available to the user rather than just the allocation or nonallocation decision. Preferably, the user should be able to choose missing allele frequency values either as constants or as the classical maybe-next-allele formula.

Some Available Programs

Currently the three most widely used programs for group allocation based on purebred genotype samples are GENECLASS2 (Piry et al. 2004), WHICHRUN (Banks and Eichert 2000) for codominant markers (microsatellites), and AFLPOP (Duchesne and Bernatchez 2001) for dominant markers (AFLP). These freely available programs can be downloaded at http://www.bio.ulaval.ca/louisbernatchez/links.htm.

Specifics of Hybrid Identification

Definition and General Principles

Hybrids may involve two distinct species, two strains, or two populations within a single species. Genetic identification of either type of hybrids is technically the same problem. However, intraspecific hybrids are typically more difficult to detect due to less genetic differentiation and therefore require considerably more information (i.e., more genotyped loci). Given two source breeds/species, a diagnostic allele (presence/absence) is one that has 100% frequency within one breed and 0% frequency in the other breed/species. Historically, genetic identification of hybrids was associated with the simultaneous presence of diagnostic alleles (presence/absence) of both source breeds/species within a single genotype (Morizot et al. 1991). Indeed genotypes with diagnostic alleles of mixed origin are easily observable and, without any calculation, can be safely attributed to hybridization assuming no other breed/species has contributed to the purported hybrid's genotype. The 100% versus 0% frequency diagnostic criterion has been somewhat relaxed in recent literature and loci with an allele differing by >99% have sometimes been considered diagnostic (Young et al. 2002). However, there has been an increasing awareness that all loci showing a frequency difference beyond sampling error could contribute to distinguish between purebreds and hybrids (Bjornstad and Roed 2002). Even though loci with 10%

frequency differential, for example, have far less hybrid detection power than diagnostic loci, they can still be cumulated to attain any power level.

Hybrid Identification as Group Identification: the Virtual-Hybrid-Group Method

Thus, hybrid identification is technically the same problem as group identification except that preidentified samples of hybrids are usually not available as one of the potential allocation groups. However, F1 hybrid allelic frequency distributions can be directly computed from purebred frequencies, say f1 and f2. For codominant loci such as microsatellites, a straightforward estimate of any hybrid allelic frequency fh is simply the average (f1 + f2)/2 of the two purebred frequencies. For dominant markers (e.g., aflp), fh = 1 − sqrt(1 − f1)*sqrt (1 − f2). This means that purebred samples are sufficient for allocation tasks including purebred and F1 hybrid groups. Again, sets of nondiagnostic loci can be used successfully for hybrid detection. Following the same idea, purebred samples also suffice to identify second-generation hybrids (F2 and backcrosses).

Special Sampling Care

Although hybrid identification is technically the same as any other type of group allocation, it requires special sampling care for two reasons. First, differentiation is weaker between F1 hybrids and purebreds than between two distinct purebreds. Second, since allelic frequencies are computed from the two purebred frequency estimates, sampling errors in the latter will be passed along to the hybrid estimates. Consequently, when hybridization is suspected, sample sizes should be increased (>30), sampling performed as randomly as possible and alleles (or presence/absence in case of AFLP) scored with extra precaution. Clearly, all of the above is even more important when second generation hybrids are considered (Epifanio and Phillipp 1997).

Efficiency and Accuracy in Hybrid Identification

There are two ways to look at the performance of a hybrid identification procedure. One important measure is the probability that, given a specimen classified as hybrid, this specimen is in fact a hybrid. Another important measure is the probability that, given a true hybrid, it was classified (allocated) as a hybrid. Following Vähä and Primmer (2006), we use the words accuracy and efficiency to denote the first and second of these two measures, respectively. The product of these two measures can be seen as the overall performance of the hybrid identification procedure.

If the likelihood distributions for purebreds and hybrids are not (nearly) perfectly disjoint, then there is an unavoidable tradeoff between accuracy and

Low accuracy and high efficiency hybrid id

	number of	specimens	among
allocated to	popA	popB	popA X popB
popA	38	0	0
popB	0	35	0
popA X popB	12	15	50
None	0	0	0

High accuracy and low efficiency hybrid id

	number of	specimens	among
allocated to	popA	popB	popA X popB
popA	35	0	0
popB	0	38	0
popA X popB	0	0	11
None	15	12	39

LOD THRESHOLD

Figure 8.8. The accuracy versus efficiency tradeoff in hybrid identification. One way to strike the desired balance between accuracy and efficiency is to fix the log-likelihood allocation threshold by running allocation simulations. Raising the LOD threshold generally decreases efficiency while increasing accuracy.

efficiency (Figure 8.8). Some users will prefer to make sure that any possible hybrid be identified (i.e., to raise the efficiency component). For instance, when there exists independent data bearing on intermediate morphological traits, uncertain hybrid genetic classification may be used in a cross-validation fashion. On the other hand, in the absence of any control data and especially when there is only a suspicion that hybrid specimens might exist, it is preferable to obtain highly confident hybrid detection (i.e., enhance the accuracy component of performance). One way to strike the desired balance between accuracy and efficiency is to fix the log-likelihood allocation threshold by running allocation simulations. For instance, raising the threshold sufficiently will virtually eliminate false hybrid classification (i.e., accuracy will become close to 100%). Of course, this will be at the expense of a higher rate of nonallocations of both purebreds and hybrids.

So far, we have discussed hybrid identification based on purebred samples. However, as with general group allocation procedure, there exist clustering methods for

hybrid identification. Two such methods have been implemented in STRUCTURE (Pritchard et al. 2000) and NEWHYBRIDS (Anderson and Thompson 2002) and have been recently assessed by Vh and Primmer (2006). It was found that both programs, unless run with very large numbers (n = 48) of codominant loci, showed high rates of misclassification of purebred as F1 hybrids even with moderately high Fst (0.12). Also backcrosses were often misclassified as purebred. Briefly, there are accuracy and efficiency problems with currently available programs performing hybrid allocation without baseline samples. Unfortunately, these methods do not provide any inbuilt mechanism, such as simulation tools, to assess the accuracy and efficiency levels associated with the user's own specific data. Therefore, it is usually much safer in hybrid studies to rely on good quality samples of purebred groups.

Markers

In principle, any type of marker (RFLP, RAPD, AFLP, microsatellite, SNP) can be used for performing hybrid identification. However, correct detection of hybrids takes more genetic information and so, roughly speaking, more loci than allocation of purebred specimens. This is especially true when purebred individuals belong to distinct, but weakly differentiated, strains. Detection of intraspecific hybrids necessitates large numbers of loci and so AFLP markers should be considered until SNP markers can be obtained in large numbers and analyzed at low cost in nonmodel species.

Available Programs

The virtual-hybrid-group method based on purebred samples has been implemented in AFLPOP (Duchesne and Bernatchez 2002) for dominant markers (AFLP). NEWHYBRIDS (Anderson and Thompson 2002) and STRUCTURE (Pritchard et al. 2000) are additional software that provides posterior distribution that individuals fall into different hybrid categories between populations using dominant or codominant markers. These programs can be downloaded at http://www.bio.ulaval.ca/louisbernatchez/links.htm.

Conclusion

The current context in the applications of molecular genetic techniques, particularly as pertaining to individual-based genotype analyses, is extremely positive. There is a wealth of powerful genetic markers that are being developed for an increasing number of cultured species, both vertebrates and invertebrates, and many efficient analytical tools are readily accessible, free of charge for the most part. It is our hope that we have provided a better understanding of the principles underlying some of the most versatile methods currently available for performing parentage, strain/population assignment, and hybrid analyses, as well as useful guidelines for choosing proper efficient analytical software.

Acknowledgments

We thank Dr. John Liu for kindly inviting us to contribute to this book. L.B.'s research is financially supported by FQRNT (Québec), NSERC (Canada), and a Canadian Research Chair in Genomics and Conservation of Aquatic resources.

References

Anderson EC and EA Thompson. 2002. A model-based method for identifying species hybrids using multilocus genetic data Genetics, 160, pp. 1217–1229.

Banks MA and W Eichert. 2000. WHICHRUN (version 3.2): A computer program for population allocation of individuals based on multilocus genotype data. J Hered, 91, pp. 87–89.

Bernatchez L and P Duchesne. 2000. Individual-based genotype analysis in studies of parentage and population allocation: how many loci, how many alleles? Can J Fish Aquat Sc, 57, pp. 1–12.

Bjornstad G and KH Roed. 2002. Evaluation of factors affecting individual allocation precision using microsatellite data from horse breeds and simulated breed crosses, An Gen, 33, pp. 264–270.

Campbell D, P Duchesne, and L Bernatchez. 2003. AFLP utility for population allocation studies: analytical investigation and empirical comparison with microsatellites. Mol Ecol, 12, pp. 1979–1992.

Cercueil A, E Bellemain, and S Manel. 2002. PARENTE: Computer program for parentage analysis. J Hered, 93, pp. 458–459.

Congiu L, I Dupanloup, T Patarnello, F Fontana, R Rossi, G Arlati, and L Zane. 2001. Identification of interspecific hybrids by amplified fragment length polymorphism: the case of sturgeon. Mol Ecol, 10, pp. 2355–2359.

Danzmann RG. 1997. PROBMAX: A computer program for allocating unknown parentage in pedigree analysis from known genotypic pools of parents and progeny, J Hered, 88, pp. 333–333.

Duchesne P and L Bernatchez. 2002. AFLPOP: a computer program for simulated and real population allocation, based on AFLP data, Mol Ecol Notes, 2, pp. 380–383.

Duchesne P, MH Godbout, and L Bernatchez. 2002. PAPA (package for the analysis of parental allocation): a computer program for simulated and real parental allocation Mol Ecol Notes, 2, pp. 191–193.

Duchesne P, T Castric, and L Bernatchez. 2005. PASOS (parental allocation of singles in open systems): a computer program for individual parental allocation with missing parents. Mol Ecol Notes, 5, pp. 701–704.

Epifanio JM and D Phillipp. 1997. Sources for misclassifying genealogical origins in mixed hybrid populations. J Hered, 88, pp. 62–65.

Ferguson MM and RG Danzmann. 1998. Role of genetic markers in fisheries and aquaculture: useful tools or stamp collecting? Can J Fish Aquat Sc, 55, pp. 1553–1563.

Gerber S, S Mariette, R Streiff, C Bodénès, and A Kremer. 2000. Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis, Mol Ecol, 9, pp. 1037–1048.

Goodnight KF and DC Queller. 1999. Computer software for performing likelihood tests of pedigree relationship using genetic markers, Mol Ecol, 8, pp. 1231–1234.

Hansen MM. 2002. Estimating the long-term effects of stocking domesticated trout into wild brown trout (*Salmo trutta*) populations: an approach using microsatellite DNA analysis of historical and contemporary samples. Mol Ecol, 11, pp. 1003–1015.

Hayes B, AK Sonesson, and B Gjerde. 2005. Evaluation of three strategies using DNA markers for traceability in aquaculture species, Aquaculture, 250, pp. 70–81.

Jackson TR, DJ Martin-Robichaud, and ME Reith. 2003. Application of DNA markers to the management of Atlantic halibut (*Hippoglossus hippoglossus*) broodstock, Aquaculture, 220, pp. 245–259.

Jones AG and WR Ardren. 2003. Methods of parentage analysis in natural populations. Mol Ecol, 12, pp. 2511–2523.

Liu ZJ and JF Cordes. 2004. DNA marker technologies and their applications in aquaculture genetics, Aquaculture, 238,, pp. 1–37.

Manel S, OE Gaggiotti, and RS Waples. 2005. Allocation methods: matching biological questions with appropriate techniques TREE, 20, pp. 136–142.

Marshall TC, J Slate, LEB Kruuk, and JM Pemberton. 1998. Statistical confidence for likelihood-based paternity inference in natural populations, Mol Ecol, 7, pp. 639–655.

Miggiano et al. 2005. AFLP and microsatellites as genetic tags to identify cultured gilthead seabream escapees: data from a simulated floating cage breaking event, Aqua Intern, 13, pp. 137–148.

Morizot DC, SW Calhoun, LL Clepper, and ME Schmidt. 1991. Multispecies hybridization among native and introduced centrarchid basses in central Texas, Trans Am Fish Soc, 120, pp. 283–289.

Norris AT, DG Bradley, and EP Cunningham. 2000. Parentage and relatedness determination in farmed Atlantic salmon (*Salmo salar*) using microsatellite markers. Aquaculture, 182, pp. 73–83.

Piry S, A Alapetite, JM Cornuet, D Paetkau, L Baudouin, and A Estoup. 2004. GENECLASS2: A software for genetic allocation and first-generation migrant detection. J Hered, 95, pp. 536–539.

Pritchard JK, M Stephens, and P Donnelly. 2000. Inference of population structure using multilocus genotype data, Genetics, 155, pp. 945–959.

Roques S, P Duchesne, and L Bernatchez. 1999. Potential of microsatellites for individual assignment: the North Atlantic redfish (genus *Sebastes*) species complex as a case study. Mol Ecol, 8, pp. 1703–1717.

Selvamani MJP, A Sandie, and M Degnan. 2001. Microsatellite Genotyping of Individual Abalone Larvae: Parentage Assignment in Aquaculture Mar Biotech, 3, pp. 478–485.

Smouse PE, RS Spielman, and MH Park. 1982. Multiple-locus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. Am Nat, 119, pp. 445–463.

Vähä JP and CR Primmer. 2006. Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. Mol Ecol, 15, pp. 63–72.

Vandeputte M, M Kocour, S Mauger, M Dupont-Nivet, D De Guerry, M Rodina, D Gela, D Vallod, B Chevassus, and O Linhart. 2004. Heritability estimates for growth-related traits using microsatellite parentage assignment in juvenile common carp (*Cyprinus carpio L.*), Aquaculture, 235, pp. 223–236.

Waples RS and O Gaggiotti. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity, 15, pp. 1419–1439.

Wilmer JW, PJ Allen, PP Pomeroy, SD Twiss, and W Amos. 1999. Where have all the fathers gone? An extensive microsatellite analysis of paternity in the grey seal (*Halichoerus grypus*). Mol Ecol, 8, pp. 1417–1429.

Wilson AJ and MM Ferguson. 2002. Molecular pedigree analysis in natural populations of fishes: approaches, applications, and practical considerations. Can J Fish Aquat Sc, 59, pp. 1696–1707.

Young WP, CO Ostberg, P Keim, and GH Thorgaard. 2001. Genetic characterization of hybridization and introgression between anadromous rainbow trout (*Oncorhynchus mykiss irideus*) and coastal cutthroat trout (*O-clarki clarki*), Mol Ecol, 10, pp. 921–930.