Molecular Ecology Notes (2005) **5**, 701–704 doi: 10.1111/j.1471-8286.2005.01001.x

PROGRAM NOTE

PASOS (parental allocation of singles in open systems): a computer program for individual parental allocation with missing parents

PIERRE DUCHESNE, TANGUY CASTRIC and LOUIS BERNATCHEZ
Département de Biologie, Université Laval, Ste-Foy, Québec, Canada, G1K 7P4

Abstract

PASOS is a parental allocation program designed to identify collected parents based on individual multilocus genotypes while detecting missing parents when a proportion of them have not been collected. It makes use of restricted error tolerance in order to distinguish between a partially incorrect genotype from a false parent's genotype. PASOS also introduces the technique of sequence allocation allowing the user to obtain estimates of the proportion of missing parents and of allocation correctness. The PASOS interface is very similar to the one found in PAPA, its closed system counterpart (Duchesne *et al.* 2002). A help file thoroughly describes all technical terms such as error modelling, parameters and procedures. PASOS can be downloaded free of charge from: http://www.bio.ulaval.ca/louisbernatchez/.

Keywords: computer software, error model, microsatellites, parentage analysis, parental allocation

Received 17 December 2004; revision received 02 February 2005; accepted 18 March 2005

In studies aiming at documenting parentage in natural populations, sampling of candidate parents is often incomplete (Wilson & Ferguson 2002) (i.e. the allocation system is open). Using CERVUS 2.0 (Marshall *et al*. 1998), Wilson & Ferguson (2002) have shown that allocation success declines dramatically as the proportion of candidate parents sampled drops. One potentially important problem in such situations is overallocation (i.e. allocating offspring bred from uncollected parents (UC) to collected ones). Up until now, no parental allocation program has provided an estimate of the proportion of collected parents directly computed from specific parental and offspring genotypes. Hence, the overallocation effect could not be simulated from real data sets.

If genotyping were perfect, overallocation could be avoided by increasing the number of loci. However, scoring errors are commonplace (O' Reilly *et al*. 1998), and their probability of occurrence increases with the number of loci (Jones & Ardren 2003). Hence, if scoring errors were not tolerated, one could lose significant portions of correct allocations (SanCristobal & Chevalet 1997; Gerber *et al*. 2000). On the other hand, tolerating any type of discrepancy between parental and offspring alleles would provoke a tendency to overallocate. A solution to the above dilemma is to apply a restricted tolerance criterion, which consists in tolerating no more than a certain level of discrepancy between parental and offspring alleles. This has the potential to make all correct allocations to collected parents while detecting the presence of UC. PASOS makes use of restricted error tolerance to estimate the proportion of collected parents and to perform parental allocation in open systems.

We first provide definitions for the main technical terms. A *cumulative sequence* of sets of loci (CSL) is of the following form: L1, L1 + L2, L1 + L2 + L3 … , where each Li represents an individual locus. The main feature of a CSL is that each set adds a new locus to the previous set. An *offset* is the minimal distance in number of mutational steps between two alleles of microsatellite loci. For instance, within a dinucleotide locus, alleles 254 and 258 are two offsets apart but only one offset within a tetranucleotide locus. The maximum offset tolerance (MOT) refers to the maximum number of offsets between a parental and an offspring allele that PASOS accepts as possibly due to a scoring error. The MOT is a user-defined parameter. When a parental allele, scored X, is transmitted as an offspring allele, scored Y, it may be the same or different from Y. The

Correspondance: L. Bernatchez, Université Laval, Département de Biologie, Ste-Foy, Québec, Canada, G1K 7P4. Fax: +1418-656-2043; E-mail: louis.bernatchez@bio.ulaval.ca

probability that X becomes Y will be referred to as the *transmission probability* from X to Y. The *error sum* is the sum of all transmission probabilities whenever Y is different from X. In PASOS, there are two types of allocations, [i.e. to a specific collected parent and to the group of uncollected parents (UC)]. The *allocation rate* equals the number of allocations to collected parents divided by the total number of possible allocations (i.e. twice the number of offspring). The *correctness rate* is the proportion of correct allocations among all allocations to collected parents.

Error models are used to produce artificial offspring in simulation procedures. They are defined by providing transmission probabilities from parental allele X to each offspring allele Y of the following form: X – 2 offsets, X – 1 offset, X, X + 1 offset, X + 2 offsets. The sum of all transmission probabilities must equal one as in the model: 0.002, 0.008, 0.980, 0.008, 0.002. These do not have to be totally accurate and it is generally acknowledged that these decrease dramatically with distance from the parental allele X and that the error sum usually stands within the 0.01 … 0.05 range (e.g. O'Reilly *et al*. 1998). Note that error modelling in PASOS assumes that the scores of the parental allele X and its offspring copy Y do not differ by more than two offsets. If the actual scoring of genotypes generally complies with this assumption, then very few true parents will be missed when using the default value (2) for the MOT parameter.

PASOS combines an approach based on parental pair likelihoods with a subsequent filtering procedure. The allocation of an offspring in PASOS starts with the search for the most likely pair among all potential pairs of collected parents. The likelihoods are computed according to a fixed error model, wherein the transmission probability from X to X equals 0.98 and the remaining 0.02 is evenly distributed over all remaining offspring alleles for any given locus. This first step ensures that at least one most likely pair is obtained. Pairs of collected true parents will generally turn out to be the most likely pair. However, when some true parents are in fact missing from the collected parental set, some pairs of most likely parents will contain one or two false parents. These false parents have to be filtered out. To identify UC, PASOS builds a most likely allelic transmission scenario from parents to offspring. For instance, given the most likely pair of parental genotypes at a dinucleotide locus (242, 244) (234, 238) and offspring genotype (234, 242), the most likely transmissions would be 242 → 242 and 234 → 234. However, if the offspring genotype was (230, 242), then the most likely transmissions would be 242 → 242 and 234 → 230. After rebuilding transmissions, PASOS computes the distances between each transmitted parental allele and its presumed offspring counterpart. In the first example, both distances are zero but in the second example, one transmission shows a two offsets distance. The latter may be interpreted in two ways: either the parent is false or there has been a two offset scoring error or mutation between parental and offspring alleles. If PASOS is run with a maximum offset tolerance (MOT) of two, then the scoring error interpretation is chosen and so, based on the current locus, the first and the second putative parents are kept. However, if MOT had been set to one, then the first putative parent would have been kept while the second would have been rejected. For a putative parent from the most likely pair to be kept, all transmissions for this parent have to stay within the MOT for each locus (i.e. a single distance larger than the current MOT suffices to reject the candidate parent). Rejection of a candidate parent is equivalent to allocation to the group of the uncollected parents (UC).

A parental allocation algorithm in open systems should ideally be capable of correct allocation whenever it allocates to a collected parent (100% correctness rate). Also, its rate of allocation should be equal to the proportion of collected parents (i.e. it should never miss a true collected parent). These two goals should be reached whenever there is enough available genetic information and scoring errors do not exceed the current MOT parameter value. We also expect from such a program that the genetic content it needs for high confidence allocations does not increase importantly as the proportion of uncollected parents increases.

We ran the PASOS simulator with the *collected_parents* file (*Demo folder*) as a sample from which to generate artificial parents. The MOT was set at two and the error production model was 0.002, 0.008, 0.980, 0.008, 0.002 (i.e. a 0.02 error sum was distributed within two offsets from the parental allele). Figure 1 shows five allocation rate CSL curves with 500 nonsexed parents, among which, 500, 400, 300, 200, 100 were collected, the rest uncollected (0, 100, 200, 300, 400) along with horizontal lines marking the corresponding exact proportions of collected parents (1.0, 0.8, 0.6, 0.4, 0.2). Figure 2 shows correctness rate CSL curves with the same five proportions of collected parents. Each point on the curves was obtained by allocating 5 × 1000 simulated offspring. Note that each allocation curve 'converges' towards its corresponding proportion of collected parents and that increasing proportions of missing parents do not require much more genetic information in order to reach high correctness rates.

Allocation in PASOS is performed in three steps: sequence allocation of the offspring, sequence simulations and allocation proper. We now explain each step in detail.

The first step consists in allocating the offspring by choosing the sequence allocation option. The program will output directly on an EXCEL sheet an allocation rate for each set of locus of the CSL. Once drawn as a function of sets of loci, these rates produce a very informative curve. Such curves typically start at very high levels, undergo a sharp drop and then reach a break point where they begin to

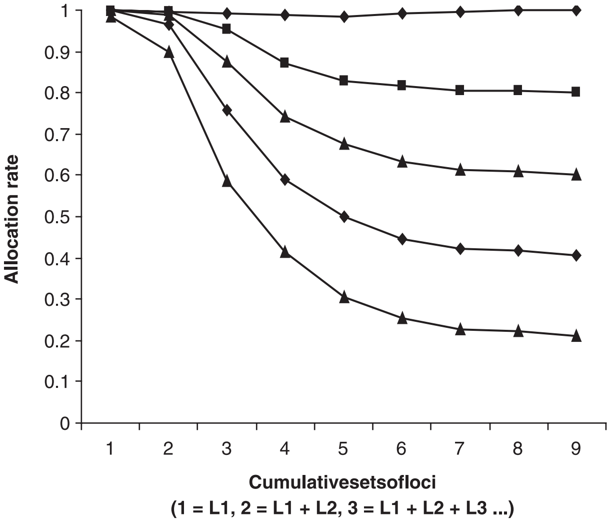

Fig. 1 Allocation rate CSL curves with 500 nonsexed parents, among which, 500, 400, 300, 200, 100 collected (from the top to the lower curve), the rest uncollected (0, 100, 200, 300, 400). The maximum offset tolerance (MOT) was set at two and the error production model was 0.002, 0.008, 0.980, 0.008, 0.002. The tick marks (1, 2, 3 …) on the *x*-axis correspond to sets of loci of the cumulative sequence: L1, L1 + L2, L1 + L2 + L3, L1 + L2 + L3 + L4 …

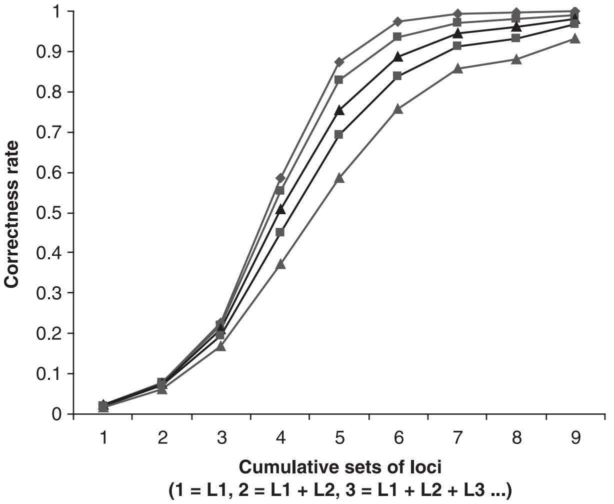

Fig. 2 Correctness rate CSL curves with 500 nonsexed parents among which 500, 400, 300, 200, 100 collected (from the top to the lower curve), the rest uncollected (0, 100, 200, 300, 400). The maximum offset tolerance (MOT) was set at two and the error production model was 0.002, 0.008, 0.980, 0.008, 0.002. The tick marks (1, 2, 3, 4 …) on the *x*-axis correspond to sets of loci of the cumulative sequence: L1, L1 + L2, L1 + L2 + L3, L1 + L2 + L3 + L4 …

stabilize (see Fig. 1). Such a break point may not be apparent in situations of uinsufficient genetic resolution or seriously flawed scoring.

The break point of the allocation rate CSL curve can be taken as a first estimate of the proportion of collected parents hence of the number of uncollected parents. For instance, in Fig. 1, all curves but the top one show a break point region located around the seven loci subset. Using the latter estimate and the same CSL produce a simulated allocation curve to be compared with the sequence allocation from the true offspring. Some refining of both the error model as well as the number of missing parents may be necessary in order to fit the simulated allocation curve to the true offspring allocation curve. Fitting of the simulated curve should not be considered in an overly strict manner. In fact, perfect fits will generally prove impossible, largely because the error model does not distinguish between loci, whereas true error does not distribute evenly across loci. However, numerous simulations (data not shown) revealed that even relatively loose fits produce very satisfactory estimates of the allocation correctness rate curve. Once fitting parameters have been selected, the corresponding correctness rate CSL curve is drawn in order to decide on the most convenient set of loci to be retained for allocation. There is usually a trade-off between rate of allocation and rate of correctness, as more loci may mean less allocations but a larger proportion of correct allocations. In contrast, increasing correctness rates slightly may sometimes lead to a significant loss of allocations. Consequently, in situation where correctness certainty is not crucial, one could fix a minimum correctness rate and select a set of loci none larger than the one providing such a minimum. If the estimated correctness rate is close to an acceptable minimum, one may increase confidence by adding an extra locus to the selected set.

The last step simply consists in allocating the real offspring with the selected set of loci. Following this three-step procedure, the user thus obtained parental allocations for each progeny, an estimate of the overall allocation correctness rate, as well as an estimate of the proportion of parents that contributed to reproduction but were uncollected.

PASOS is written in C++ and its interface is very similar to the one found in PAPA 1.1 (Duchesne *et al*. 2002). The main differences are the error modelling format, based on offset units in PASOS but not in PAPA 1.1, the addition in PASOS of a file that summarizes the main simulation results and of a sequence allocation option both in allocation and simulation procedures. PASOS accepts the same genotype file formats as PAPA 1.1.

Acknowledgements

This work was supported by a research grant from Science and Engineering Research Canada (NSERC strategic program) to L.B. and J.J. Dodson, and funds from the Canadian Research Chair in

Conservation Genetics of Aquatic Resources (L.B.). This study is a contribution to the research programs of CIRSA (Centre Inter-universitaire de Recherche sur le Saumon Atlantique) and Québec-Océan.

References

Duchesne P, Godbout MH, Bernatchez L (2002) PAPA (Package for the analysis of parental allocation): a computer program for simulated and real parental allocation. *Molecular Ecology Notes*, **2**, 191–193.

Gerber S, Mariette S, Streiff R, Bodenes C, Kremer A (2000) Comparison of microsatellites and amplified fragment length polymorphisms markers for parentage analysis. *Molecular Ecology*, **9**, 1037–1048.

Jones GJ, Ardren WR (2003) Methods of parentage analysis in natural populations. *Molecular Ecology*, **12**, 2511–2523.

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, **9**, 639–655.

O'reilly P, Herbinger C, Wright J (1998) Analysis of parentage determination in Atlantic salmon (*Salmo salar*) using microsatellites. *Animal Genetics*, **29**, 363–370.

SanCristobal M, Chevalet C (1997) Error tolerant parent identification from a finite set of individuals. *Genetical Research*, **70**, 53–62.

Wilson AJ, Ferguson MM (2002) Molecular pedigree analysis of fishes: approaches, applications, and practical considerations. *Canadian Journal of Fisheries and Aquatic Science*, **59**, 1696–1707.