Molecular Ecology Notes (2002) **2**, 191–193

PROGRAM NOTE

PAPA (package for the analysis of parental allocation): a computer program for simulated and real parental allocation

PIERRE DUCHESNE, MARIE-HÉLÈNE GODBOUT and LOUIS BERNATCHEZ
Département de biologie, Université Laval, Ste-Foy, Québec, Canada, G1K 7P4

Abstract

PAPA is a parental pair allocation and simulator program. The allocation method is based on the likelihood of a parental pair producing the multilocus genotype found in the offspring being tested, which will be referred to as the *breeding likelihood*. Estimated level and structure of allele transmission errors in offspring are parameters fed into the allocation procedure. The embodied Monte-Carlo simulator also allows modelling of many allocation conditions, including transmission error and the estimated proportion of missing parents. Simulations may be run prior to the collection of real parents in order to define the minimal set of loci that is necessary to reach a desired level of allocation success. Post-collection simulations aim at statistically assessing the reliability of nonsimulated allocations. Simulations output values for several random variables.

Keywords: individual relationships, likelihood, microsatellite, multi-locus genotype, parentage, simulation

Received 3 September 2001; revision received 25 October 2001; accepted 9 November 2001

The assessment of precise parental relationships within populations through parentage allocation allows researchers to define social structure (Amos *et al*. 1993), mating patterns (Clapham & Palsbøll 1997; Hughes 1998), kinship (Blouin *et al*. 1996), and quantify reproductive success (DeWoody *et al*. 1998; Coltman *et al*. 1999; Garant *et al*. 2001). Such analyses may also contribute to improve the efficiency of selective breeding programmes in domesticated populations (Herbinger *et al*. 1997; Estoup *et al*. 1998; Ferguson & Danzmann 1998).

PAPA is a computer program that performs parental allocation based on breeding likelihood methods, and also comprises simulators that allow statistical assessments of allocation accuracy. It was written in C++ and is available as executable code to run on IBM PC compatible machines under Windows 95 or higher. It can be downloaded free of charge at http://www.bio.ulaval.ca on L. Bernatchez's personal web page.

The parentage allocation method used in PAPA is based on breeding likelihood (Sancristobal & Chevalet 1997). Given an offspring genotype, the likelihood of a parental pair of genotypes is defined as the probability of this pair breeding the offspring genotype among all of its possible descents. Contrary to exclusion methods, likelihood-based parentage allocation methods allow for some degree of transmission errors due to genotype misreading or mutation (e.g. Sancristobal & Chevalet 1997; Marshall *et al*. 1998). Moreover, for allocation to take place under an exclusion-based method, all likelihoods, save one, must equal zero, whereas the condition for allocation under likelihood methods is simply that of the highest computed likelihood belonging to a single allocation unit. The relaxed nature of the latter condition means that likelihood methods generally call for much less extensive genetic information than exclusion-based ones. PAPA complements other available parental allocation programs in several ways. Unlike PAPA, PROBMAX (Danzmann 1997) is based on the number of loci compatible with a parent–offspring relationship. However, the latter statistic, contrary to breeding probability proper, does not make full use of the combinatorics involved in allele transmission processes, resulting in a loss of power. Also, PROBMAX does not provide any systematic treatment of transmission errors. PAPA also differs from paternity/maternity allocation programs such as CERVUS (Marshall *et al*. 1998) in

Correspondence: L. Bernatchez. Fax: + 1418-656-2043; E-mail: Louis.Bernatchez@bio.ulaval.ca

the sense that it is basically a parental pair allocation program. Paternity/maternity allocation usually involves choosing among several candidate parents of the same sex, the other sex being represented by a single known individual. In that case, individual parent and parental pair allocation procedures are equivalent. However, allocation on a parental pair basis is the logical choice whenever there are several putative parents of both sexes and all or most of their genotypes are known.

PAPA provides a Monte-Carlo simulator that may be used to obtain empirical distributions of many relevant random variables such as rates of successful allocations and allocation failures. The embodied simulator allows modelling of all allocation conditions (allele transmission error, missing parents, etc.). Simulations may also be run prior to collection of parents, on the basis of a subsample of genotypes from the same population as the parents and offspring to be collected. Such preparental simulations may be particularly useful in order to decide on a minimal set of loci to reach a desired level of allocation success, thus avoiding unnecessary postcollection genotyping. On the other hand, parental simulations, which involve already known parental genotypes, produce empirical distributions wherefrom the accuracy of the allocation of real offspring may be assessed.

User interface and genotype file formats. The choice of options is made by clicking buttons. Input files may be selected either by filling a text box or by selection from a list box. All parameter values are fed to PAPA through text boxes. A Help button provides an index of technical words and their definition.

All genotype input files should be formatted in text files as in the following example:

Ne18–1229243 188188 301313 111135 188204
Ne18–2235255 172184 305313 000000 196198

The first item in each row refers to the name of the individual. Each locus is described with six digits, three to each allele. The 000000 string represents a nongenotyped locus.

Allocation method. To allocate an offspring, its breeding likelihood is computed for each potential parental pair. If the highest likelihood belongs to a single parental pair, then the offspring is allocated to the latter. When all parental pairs show zero likelihood, this is referred to as a null-likelihood. A situation where more than one parental pair get the highest (non-zero) likelihood score is referred to as an ambiguity. In the advent of a null-likelihood or an ambiguity, the offspring is not allocated and the procedure is said to have failed.

Allocation parameters. Users have control over three allocation parameters: (i) choice of loci; (ii) global level of transmission error; and (iii) distribution of transmission error over alleles.

Any combination of loci may be chosen from the full set enumerated in the names of loci file provided by the user. Given a specific locus, the global level of error $\in$ is the sum of all probabilities that an allele **a** transmitted to an offspring through reproduction becomes **b** (**a** $\neq$ **b**) as a consequence of mutation or genotype misreading. In other words, there is a probability $1-\in$ that allele **a** remains **a** in an offspring after transmission from a parent and subsequent genotyping. Once chosen, the value of the parameter $\in$ applies to all loci. As PAPA is primarily (although not restricted to) designed for the use of microsatellite data, it is assumed that allelic identification refers to allelic size in base pairs. The *distribution* of error is user-controlled via parameter **S**, the speed of loss of error probability over neighbouring alleles. Given parental allele **a**, the error probability attributed to allele **b** is proportional to $\mathbf{1/D^S}$. where $\mathbf{D = |a - b|}$, i.e. the distance between **a** and **b**.

Allocation output formats. PAPA provides two types of output for allocation procedures. One is a text file enumerating offspring names and multilocus genotypes, each followed by all pairs of parents' names and multilocus genotypes showing highest non-zero likelihood. A second allocation output is available in spreadsheet format (Excel). Names of offspring and allocated parents are enumerated, but genotypes are omitted.

Simulation conditions. Simulations may be run under two distinct conditions: (i) preparental; and (ii) parental. The preparental condition is one where parents have not yet been genotyped but a sample of genotypes from the same population is available. In the parental condition, a set of real potential parents is provided. We now consider the procedures associated with these two conditions.

The main purpose of preparental simulations is to provide a statistical basis for the choice of loci to be eventually genotyped for allocation purposes. The user provides a sample file of multilocus genotypes to generate pseudo-collected and uncollected parents. Users are requested to provide values for the number and proportion of pseudo-collected parents, global level and distribution of transmission error. Each iteration is a 3-step process. First, pseudo-collected and uncollected parental genotypes are generated. Second, the two sets of parental genotypes are combined to breed pseudo-offspring. Third, pseudo-offspring are allocated to parental pairs belonging to the pseudo-collected parents only.

The parental simulation procedure is nearly identical to the preparental one except for the use of real collected parental genotypes. Its main purpose is to build empirical distributions of random variables, given the real collected parental genotypes. Such distributions may be used to assess the accuracy level of a specific allocation process.

Simulation output random variables

Non-sexed parents condition

Four random variables are considered:

correct_2_par proportion of offspring with two parents correctly allocated

correct_1_par proportion of offspring with one parent correctly allocated

correct_0_par proportion of offspring with no parent correctly allocated

failed proportion of offspring with failed allocation (null-likelihood or ambiguity)

Sexed parents condition

Seven random variables are considered:

correct_pair proportion of offspring with correctly allocated parental pair

incorrect_pair proportion of offspring with incorrectly allocated parental pair

correct_male proportion of offspring with correctly allocated male

incorrect_male proportion of offspring with incorrectly allocated male

correct_female proportion of offspring with correctly allocated female

incorrect_female proportion of offspring with incorrectly allocated female

failed proportion of offspring with failed allocation (null-likelihood or ambiguity)

Data for all random variables are outputted in a spreadsheet (Excel). The user may run all option procedures of the program using demo files which can also serve as format models.

Acknowledgements

The authors are grateful to Julie Turgeon for her suggestion of the program's acronym, and D. Fraser for his constructive comments. The research program of L.B. on northern fishes is supported by NSERC (Canada) research grants.

References

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

Blouin MS, Parsons M, Lacaille V, Lotz S (1996) Use of microsatellite loci to classify individuals by relatedness. *Molecular Ecology*, **5**, 393–401.

Clapham PJ & Palsbøll PJ (1997) Molecular analysis of paternity shows promiscuous mating in female humpback whales (*Megaptera novaeangliae*, Borowski). *Proceedings of the Royal Society of London B*, **264**, 95–98.

Coltman DW, Bancroft DR, Robertson A, Smith JA, Clutton-Brock TH, Pemberton JM (1999) Male reproductive success in a promiscuous mammal: behavioural estimates compared with genetic paternity. *Molecular Ecology*, **8**, 1199–1209.

Danzmann RG (1997) PROBMAX: a computer program for assigning unknown parentage in pedigree analysis from known genotypic pools of parents and progeny. *Journal of Heredity*, **88**, 333.

DeWoody JA, Fletcher DE, Wilkins SD, Nelson WS, Avise JC (1998) Molecular genetic dissection of spawning, parentage, and reproductive tactics in a population of redbreast sunfish, *Lepomis auritus*. *Evolution*, **52**, 1802–1810.

Estoup A, Gharbi K, SanCristobal M, Chevalet C, Haffrey P, Guyomard R (1998) Parentage assignment using microsatellites in turbot (*Scophtalmus maximus*) and rainbow trout (*Oncorhynchus mykiss*) hatchery populations. *Canadian Journal of Fisheries and Aquatic Sciences*, **55**, 715–725.

Ferguson MM & Danzmann RG (1998) Role of genetic markers in fisheries and aquaculture: useful tools or stamp collecting? *Canadian Journal of Fisheries and Aquatic Sciences*, **55**, 1553–1563.

Garant D, Dodson JJ, Bernatchez L (2001) A genetic evaluation of mating system and determinants of individual reproductive success in Atlantic salmon (*Salmo salar* L.). *Journal of Heredity*, **92**, 137–145.

Herbinger CM, Doyle RW, Taggart CT *et al.* (1997) Family relationships and effective population size in a natural cohort of cod larvae. *Canadian Journal of Fisheries and Aquatic Sciences*, **54**, 11–18.

Hughes C (1998) Integrating molecular techniques with field methods in studies of social behavior: a revolution results. *Ecology*, **79**, 283–299.

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, **7**, 639–655.

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