PROGRAM NOTE

AFLPOP: a computer program for simulated and real population allocation, based on AFLP data

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

Abstract

AFLPOP is a population allocation and simulator program based on amplified fragment length polymorphism markers. The allocation method is an adaptation of Paetkau's method for co-dominant alleles. Besides population allocation of specimens of unknown origin, re-allocation of sample genotypes, as well as allocation of artificial (Monte Carlo) specimens, may be run to estimate expected rates of correct allocations. Thanks to its embodied simulator, AFLPOP can provide information on the rates and types of incorrect allocations and on empirical distributions of likelihood statistics. A filtering procedure within AFLPOP allows the selection of loci according to user-defined criteria.

Keywords: AFLP, likelihood, multi-locus-genotype, population assignment, software, simulation

Received 17 April 2002; revision received 9 May 2002; accepted 9 May 2002

General outline

The development of hypervariable genetic markers over the last decade, and particularly of microsatellite DNA loci, has revived a major interest in such studies using individual-based genotypic information and has opened exciting avenues of research and applications (Davies *et al*. 1999). Namely, studies of individual-based population allocation (*sensu* Paetkau *et al*. 1995) may be relevant in both animals and plants to quantify the gene flow and dispersal, to assess the genetic distinctiveness of populations and to establish relationships among individuals within and among populations or higher taxonomic groupings (reviewed in Waser & Strobeck 1998; Hansen *et al*. 2001) with more precision. An extension of these approaches is to detect an admixture of populations in a sample of individuals of unknown origin (Potvin and Bernatchez 2001). Most empirical studies using population allocation methods have relied on the use of microsatellite DNA loci with high allelic diversity. However, an important constraint faced by the use of microsatellite loci in any type of individual-based population allocation is the lack of statistical power in situations of weak population differentiation (marine species, large terrestrial plant populations, etc.) among putative source populations (Hansen *et al*. 2001). This is not only a result of the logistical limitation in largely increasing the number of loci used, but also of the limited informative content of low-frequency allelic diversity (Roques *et al*. 1999). Indeed, Bernatchez & Duchesne (2000) showed analytically that, in contrast to studies of parentage, increasing the number of loci used is more critical than increasing allelic diversity per locus in studies of population allocation. Among all methods available, the analysis of amplified fragment length polymorphism (AFLP) is probably the one offering the best potential for increasing the number of loci screened, and thereby potentially increasing power in studies of population allocation. However, currently available software for performing such analyses and related methods (Cornuet *et al*. 1999; Banks *et al*. 2000; Pritchard *et al*. 2000) are not designed to accommodate easily the use of AFLP data.

AFLPOP is a computer program that performs population allocation by answering the following question: given an individual AFLP genotype and a set of AFLP data for putative source populations, what population is the genotype most likely to belong to? It also comprises simulators that allow statistical assessments of allocation efficiency. In addition, AFLPOP offers a locus filter that performs easy and fast selection of loci according to user-defined logical criteria. This automatic selection mechanism is especially convenient, given the sometimes enormous number of AFLP loci. AFLPOP was written in VBA (Visual Basic for Applications) and runs automatically as soon as it is opened

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

© 2002 Blackwell Science Ltd

as an EXCEL workbook. Currently, it runs on any 97 or later version of PC EXCEL. Most results are produced both in numerical and graphical form. AFLPOP can be downloaded free of charge at http://www.bio.ulaval.ca/index-alt.html on L. Bernatchez's personal web page.

User interface

Choice of options is made by clicking buttons. Input files may be selected from a list box. All parameter values are fed to AFLPOP through text boxes. Help in the form of a ControlTipText can be obtained for each item (parameter or file name) within any current dialog box.

Input file format

The input files have to be in EXCEL spreadsheet format (.xls). The user will find examples of input files in the Demo Files AFLPOP folder. All files are formatted in exactly the same fashion. The first two columns are devoted to locus identification and the first row is devoted to specimen identification. Since each dominant band is represented by its presence (1) or absence (0), the remainder of the sheet is just a matrix filled with binary digits (zeros and ones). Note that the loci are doubly identified (by number and by name).

Allocation method

Suppose G is the genotype of the specimen to be allocated. First the likelihood is computed that G be found in each of the candidate populations based on their respective dominant band frequencies. G is then allocated to the population showing the highest likelihood for G. The likelihood of genotype G in population X, among all possible genotypes in X, is the product:

$$L_X = \prod_{i=1}^{n_1} f_{i,X} \prod_{j=1}^{n_0} (1 - f_{i,X})$$

where $f_{i,X}$ is the frequency of dominant band in locus i in population X; n_1 is the number of loci with presence of dominant bands in genotype G; and n_0 is the number of loci with absence of dominant bands in genotype G.

However, for computational reasons, it is more convenient to work with the log of the likelihood:

$$\log L_X = \sum_i^{n_1} \log f_{i,X} + \sum_i^{n_0} \log (1 - f_{i,X})$$

In practice, this means that each 1 in G is replaced by the log of the frequency of 1 s within the same locus in population X and each 0 is replaced by the log of the frequency of 0 s. Since log(0) is not defined, it is replaced by log(ε) with the value of ε being very small and chosen by the user (e.g. 0.001). Then the logs are added together to obtain the log-likelihood of genotype G in X. Likelihoods are computed for each population. If there is a single largest log-likelihood, then G is allocated to the corresponding population, otherwise it is not allocated and the procedure is said to have failed.

This method is an adaptation of Paetkau's method (Paetkau *et al*. 1995) for co-dominant markers. In the latter, however, each one-locus-likelihood is obtained by multiplication of the frequencies of its two component alleles. In the context of AFLP (dominant) markers, the only two possible one-locus-genotypes are 0 and 1 and so their likelihoods are simply their respective frequencies within the locus. Strictly speaking, the above formulae are based on the assumptions that the $f_{i,X}$ are accurate and that the loci are statistically independent (no significant linkage disequilibrium).

Available procedures

AFLPOP offers five procedures.

Allocation (origin unknown)

Allocates specimens of unknown origin to their most probable population.

Re-allocation (source populations)

Re-allocate specimens of the source populations. The main purpose of this procedure is to estimate the expected allocation success rate for each of the source populations. Note that each specimen to be allocated is withdrawn from its population and allelic frequencies for this population are computed anew.

Simulation: one iteration (likelihood statistics shown)

A number (decided by the user) of random genotypes are generated from sample population files and are allocated. This procedure outputs empirical distributions of various likelihood statistics.

Simulation: many iterations

At each iteration random specimens are generated and allocated. This procedure serves to estimate distributions of rates of nonallocation and rates of allocations to the right and wrong populations. The number of genotypes and the number of iterations are under user control.

Locus filtering

AFLP loci can be very abundant but their allocation power may range from null to very large. The *Locus filtering* procedure allows the user to delete loci according to various criteria applied to a chosen set of AFLP files. Once loci have been deleted the original files are not modified but new files are created which include all but the deleted loci. Each new file bears the same name as its original counterpart except for an extension which briefly describes the filtering procedure. Also the user may specify other files from which to delete the same loci as in the original set of genotype files. This possibility can be used prior to an allocation procedure so that once the sample files have been filtered the remaining set of loci are the same both in sample and allocation files. We briefly describe the three types of locus filtering available in AFLPOP.

Deletion of loci with max–min frequencies less than or equal to a value to be specified Given a specific locus, let us call the proportion of presences (1 s) within a given sample a *frequency*. The deletion procedure eliminates all loci with maximum frequency minus minimum frequency (max–min) smaller than or equal to some user-defined level. The user specifies the interval for which the max–min criteria are to apply. For instance, to delete identical frequency loci it suffices to use the criteria specification: [0, 1]; 0. The interval specification is motivated by the observation that near tips (0 and 1) frequencies (*T-loci*) have more allocation power than loci with middle range frequencies (*M-loci*).

Identification of clusters of redundant (potentially linked) loci Loci that show identical scores among all specimens across all sample files are potentially linked and therefore are said to be redundant in information content. When a cluster of such 'redundant loci' is found by the filtering procedure all loci but one are eliminated.

Selection of prefixed loci The user may ask for the selection of all loci whose names start with some specified string of characters, a 'prefix'. This is especially useful to make a locus selection on the basis of primer grouping. The names of the loci should be prefixed accordingly.

Allocation parameters

Users have control over two allocation process parameters: choice of zero replacement value and minimal log likelihoood difference. These two parameters are available in allocation and simulation procedures.

Choice of zero frequency replacement value

If individual X shows a dominant band on a locus L for which population P has a frequency of 0 for dominant bands, then, strictly speaking, X should not be allocated to P, whatever the remainder of its genotype. But because sampling errors cannot be avoided, we can never be sure that the true frequency of P is indeed 0. Therefore, it would be ill-advised to throw away large amounts of information on this basis. Hence, it is reasonable to replace zero frequencies by a (very) small frequency, thereby preserving the available information. There are many ways to choose zero replacement values. Some people use formulae dependent on the number of specimens. Others prefer a constant replacement value. At the time of publication, AFLPOP accepts one formula, i.e. 1/(number of specimens + 1) as well as any user-defined constant value. Once a zero replacement value ε has been chosen, the frequencies of ones (all dominant bands) will automatically be replaced by $1 - \varepsilon$.

Minimal log likelihoood difference (MLD)

Specimens are allocated on the basis of log-likelihoods. The user is invited to decide on the minimal difference (between the most likely and the second most likely populations) necessary to allocate specimens. For instance an MLD of 3 means that a genotype has to be 10^3 times more likely to be found in population X than in any other population to be allocated to X. A higher MLD will generally lead to a lesser rate of wrong allocations but to a higher rate of nonallocations. A good way to optimize the minimal log difference is to run the *Simulation: many iterations* procedure with various MLD values until the smallest value is found which practically guarantees the rate of correct allocations desired by the user.

Output files

The results are outputted as an EXCEL workbook file. All procedures produce a sheet describing the information provided by the user (choice of files and parameter values) and a sheet showing the locus frequencies. In all procedures except *Locus filtering*, likelihoods and allocations are presented on separate sheets both in numerical and graphical form. Besides a result file proper, the *Locus filtering* procedure outputs new files after possible deletion of loci according to a user-defined criteria. These new files bear the same name as the original ones plus an extension which briefly describes the filtering criteria and they are formatted exactly as input files. The location of all output files is the same as that of the input files.

Demo files

In addition to the main program file, the AFLPOP package contains a repertoire of short demo files. The user may run all procedures of the program using these files which can also serve as format models.

Acknowledgements

The authors are grateful to David Campbell and Sean Rogers for having provided the most precious of all programming tools: sensible requests. The research program of L. B. on the evolution and conservation of northern fishes is supported by NSERC (Canada) research grants and a Canadian Research Chair.

References

Banks MA, Eichert W (2000) WHICHRUN (Version 3.2): a computer program for population assignment of individuals based on multilocus genotype data. *Journal of Heredity*, **91**, 87–89.

Bernatchez L, Duchesne P (2000) Individual-based genotype analysis in studies of parentage and population assignment: how many loci, how many alleles? *Canadian Journal of Fisheries and Aquatic Sciences*, **57**, 1–12.

Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M (1999) Comparison of methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*, **153**, 1989–2000.

Davies N, Villablanca FX, Roderick GK (1999) Determining the source of individuals: multilocus genotyping in nonequilibrium population genetics. *Trends in Ecology and Evolution*, **14**, 17–21.

Hansen MM, Kenchington E, Nielsen EE (2001) Assigning individual fish to populations using microsatellite DNA markers. *Fish and Fisheries*, **2**, 93–112.

Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology*, **4**, 347–354.

Potvin C, Bernatchez L (2001) Lacustrine spatial distribution of landlocked Atlantic salmon populations assessed across generations by multi-locus individual assignment and mixed-stock analyses. *Molecular Ecology*, **10**, 2375–2388.

Pritchard JM, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

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

Waser PM, Strobeck C (1998) Genetic signatures of interpopulation dispersal. *Trends in Ecology and Evolution*, **13**, 43–44.