Journal of Heredity 2004:95(6):536–539
doi:10.1093/jhered/esh074

Computer Note

GENECLASS2: A Software for Genetic Assignment and First-Generation Migrant Detection

S. PIRY, A. ALAPETITE, J.-M. CORNUET, D. PAETKAU, L. BAUDOUIN, AND A. ESTOUP

From the Centre de Biologie et de Gestion des Populations, INRA, Campus International de Baillarguet CS 30 016, F34988, Monferrier-sur-Lez CEDEX, France (Piry, Alapetite, Cornuet, and Estoup); Wildlife Genetics International, Box 274, Nelson, BC V1L 5P9, Canada (Paetkau); and CIRAD, Département Cultures Pérennes, Avenue Agropolis, F34098 Montpellier CEDEX 5, France (Baudouin).

Address correspondence to Sylvain Piry at the address above, or e-mail: piry@ensam.inra.fr.

GENECLASS2 is a software that computes various genetic assignment criteria to assign or exclude reference populations as the origin of diploid or haploid individuals, as well as of groups of individuals, on the basis of multilocus genotype data. In addition to traditional assignment aims, the program allows the specific task of first-generation migrant detection. It includes several Monte Carlo resampling algorithms that compute for each individual its probability of belonging to each reference population or to be a resident (i.e., not a first-generation migrant) in the population where it was sampled. A user-friendly interface facilitates the treatment of large datasets.

The general aim of genetic assignment methods is to assign or exclude reference populations as possible origins of individuals on the basis of multilocus genotypes. Faster and cheaper development of highly polymorphic genetic markers, such as microsatellites, has increased the efficiency of such methods (see Estoup et al. [1998] for an empirical comparison of assignment results using microsatellite and protein loci). Genetic assignment methods are useful in addressing issues such as relationships, structure, and classification at the individual level (reviewed in Estoup and Angers [1998]). Because assignment methods allow us to draw inferences about where individuals were or were not born, they also have the potential to provide direct estimates of real-time dispersal through the detection of immigrant individuals (Paetkau et al., 2004; Rannala and Mountain 1997).

Several methods based on different assignment criteria computed for likelihood estimation have been developed to reach these goals (Cornuet et al. 1999; Paetkau et al. 1995; Rannala and Mountain 1997). The question of individual assignment to population samples also prompted the development of statistical methods distinguishing between resident individuals that are "misassigned" (have a genotype that is most likely to occur in a population other than the one in which the individual was sampled) by error from real immigrant individuals (i.e., type I error). Monte Carlo resampling methods have been proposed to identify a statistical threshold beyond which individuals are likely to be excluded from a given reference population sample (Cornuet et al. 1999; Paetkau et al., 2004; Rannala and Mountain 1997). The principle behind these resampling methods is to approximate the distribution of genotype likelihoods in a reference population sample and then compare the likelihood computed for the to-be-assigned individual to that distribution. Paetkau et al. (2004) have shown that the Monte Carlo resampling methods of Cornuet et al. (1999) and Rannala and Mountain (1997) generally result in an excess of resident individuals being excluded. In fact, the identification of accurate critical values required resampling methods to preserve the linkage disequilibrium deriving from recent generations of immigrants (i.e., admixture linkage disequilibrium) (Stephens et al. 1994) and to reflect the sampling variance inherent in the limited size of reference datasets. Paetkau et al. (2004) proposed a new Monte Carlo resampling method taking into account those aspects and that better control type I error rates. In particular, this resampling method was found to perform better than other ones for the detection of first-generation migrants (Paetkau et al. 2004).

Most available computer programs have been developed for assigning individual diploid genotypes (e.g., GENECLASS [Cornuet et al. 1999] and Immanc [Rannala and Mountain 1997]). GENECLASS2 provides an efficient and user-friendly tool that (1) computes various genetic assignment criteria used for likelihood estimation, (2) treats datasets with diploid or haploid data, (3) assigns or excludes individuals as well as groups of individuals to reference populations, and (4) computes probabilities that each individual belongs to each reference population or is a resident (i.e., not a first-generation immigrant) in the population where the individual has been sampled (cf. the first-generation migrant detection task). For the computations performed in (4), different Monte Carlo resampling algorithms, including that of Paetkau et al. (2004), have been implemented.

Statistical Criteria

Three types of criteria used for likelihood estimation have been implemented in GENECLASS2: genetic distance-based criteria, a criterion directly based on allele frequencies,

and Bayesian criteria (see Cornuet et al. [1999] for a comparative study).

Distance Criteria

The assignment criterion is a genetic distance computed between the individual or group of individuals to be assigned and each reference population (Cornuet et al. 1999). The following distances have been implemented: Nei's standard genetic distance (Nei 1972), Nei's minimum genetic distance (Nei 1973 in Nei 1987), Nei's *Da* distance (Nei et al. 1983), Chord distance (Cavalli-Sforza and Edwards 1967), and a distance taking into account the allele size of microsatellite markers (Goldstein et al. 1995). For a review of a mathematical description of these distances, see Takezaki and Nei (1996).

Frequency Criterion

Paetkau et al.'s (1995) formula used to compute the assignment criterion was generalized for any level of ploidy or for groups of individuals. A multinomial distribution was used to compute the likelihood that an individual (group of individuals) originates from a given population sample. This likelihood, L, of the sample follows a multinomial distribution:

$$L = \frac{m!}{m_1!m_2!\ldots m_k!}p_1^{m_1}p_2^{m_2}\ldots p_k^{m_k}. \quad (1)$$

Since a gamma function $\Gamma(x+1) = x!$, the log-likelihood can be written as follows:

$$Log(L) = Log(\Gamma(m+1)) - \sum_{i=1}^{k} Log(\Gamma(m_i+1)) + \sum_{i=1}^{k} m_i Log(p_i), \quad (2)$$

with m, the total number of genes to be assigned, or m_i, the number of copies of allele i in the to-be-assigned sample, and p_i being the frequency of allele i in the reference population sample. When an allele is found in the to-be-assigned sample but not in the reference dataset, its frequency is set to an adjustable default value (e.g., 0.01). Paetkau et al. (2004) have shown that the value of this default frequency for missing alleles has little effect on assignment results.

Bayesian Criteria

A derived method (Rannala and Mountain 1997) has been implemented for treating genomes with any level of ploidy or groups of individuals. The prior distribution for allele frequency at a given locus is a Dirichlet distribution with parameter $\alpha_i = \alpha/k$, where i is the current allele and k is the total number of different allelic states at this locus over all reference populations. In Rannala and Mountain (1997) $\alpha = 1$, while $\alpha = k$ in Baudouin and Lebrun (2000), the latter case corresponding to a uniform prior distribution of allele frequencies. The likelihood (L) that an individual (group of individuals) originates from a given population sample is

$$\begin{aligned} Log(L) &= Log(\Pr(m/n) \\ &= Log(\Gamma(m+1)) + Log(\Gamma(n+\alpha)) \\ &\quad + \sum_{i=1}^{k} Log\left(\Gamma\left(m_i + n_i + \frac{\alpha}{k}\right)\right) \\ &\quad - \sum_{i=1}^{k} Log(\Gamma(m_i+1)) \\ &\quad - \sum_{i=1}^{k} Log\left(\Gamma\left(m_i + \frac{\alpha}{k}\right)\right) \\ &\quad - Log(\Gamma(m+n+\alpha)), \end{aligned} \quad (3)$$

with Γ being a gamma function with parameters m, the total number of genes to be assigned, or m_i, the number of copies of allele i in the to-be-assigned sample, and n, the total number of genes in the reference population sample, or n_i, the number of copies of allele i in the reference population sample.

Self-Assignment and Detection of Migrants

These procedures are based on the computation of the above statistical criteria for all individuals included in a given population dataset. When the population considered is that where the individual has been sampled (i.e., self-assignment task), individuals are excluded from their population during computation (leave-one-out procedure; Efron 1983).

For the specific task of first-generation migrant detection, the statistical criterion computed for likelihood estimation can be one of three types: (1) the likelihood of the individual genotype within the population where the individual has been sampled (L_home), (2) the ratio of L_home to the highest likelihood value among all available population samples including the population where the individual was sampled (L_max) (Paetkau et al. 2004), and (3) the ratio of L_home to the highest likelihood value among all population samples excluding the population where the individual was sampled (L_max_not_home). Note that the likelihood ratios L_home/L_max and L_home/L_max_not_home have more power than the L_home statistics (cf. Paetkau et al. [2004] for L_home/L_max versus L_home, and unpublished results for L_home/Lmax_not_home versus L_home). Such likelihood ratios are appropriate when all source populations for immigrants are thought to be sampled. However, if some source populations are clearly missing, it becomes more appropriate to use L_home as the test statistic for the detection of first-generation migrants.

Probability Computation

The program computes the probability of the multilocus genotype of each individual to be encountered in a given population. Monte Carlo methods allow computing a random sample of multilocus genotypes for a large number of

individuals (e.g., 1000 or 10,000). The assignment criterion values of the simulated individuals are then computed, stored, and sorted, so that the probability of an observed multilocus genotype can be estimated as the rank of its corresponding criterion value within the distribution of simulated criterion values (Cornuet et al. 1999; Rannala and Mountain 1997).

Historically, probabilities were computed by simulating a single set of a large number of individuals by the random drawing of alleles using allele frequencies directly estimated from the reference population samples (e.g., the programs GENECLASS [Cornuet et al. 1999] or IMMANC [Rannala and Mountain 1997]). Paetkau et al. (2004) show that these Monte Carlo resampling methods introduce a bias that leads to overrejection of resident individuals. To correct for this bias, Paetkau et al. (2004) propose a new Monte Carlo resampling method which has been implemented in GENECLASS2, in addition to those used in Cornuet et al. (1999) and Rannala and Mountain (1997). This new simulation algorithm generates population samples of the same size as the reference population sample. The assignment criterion is then computed for each individual of the newly simulated population minus itself (leave-one-out procedure). The program iterates until the total number of simulated assignment criterion values is reached (e.g., 1000 or 10,000). Because this method takes into account the sample size of the reference population, it better reflects the sampling variance associated with the analyzed dataset than the resampling procedure of Cornuet et al. (1999) and Rannala and Mountain (1997).

The second important feature of the resampling method of Paetkau et al. (2004) is that multilocus genotypes of the simulated individuals are generated by the random drawing of multilocus gametes using the following procedure. In the case of sexually reproducing diploid individuals, one individual (i.e., a potential "parent") is randomly drawn from the reference population sample, and one gene and corresponding allelic state is then randomly drawn among the two gene copies for each locus. A second gamete is designed the same way from a second individual ("parent") and both gametes are associated to give a simulated multilocus diploid genotype. This method was also adapted to haploid individuals with a diploid reproduction phase. The random generation of gametes as a basis for constructing simulated individual genotypes has the advantage of preserving the potential admixture linkage disequilibrium deriving from recent generations of immigrants (Paetkau et al. 2004).

It is worth noting that, while the computation of the assignment criteria was generalized for any level of ploidy or for groups of individuals, the three resampling methods implemented in GENECLASS2 only apply to haploid or diploid biological organisms with a sexual reproduction phase.

The computation of the probability of belonging is relatively time consuming. Typically computation without using the probabilities of belonging option takes a few seconds to a few minutes to run, while the computation of probabilities takes a few minutes to several hours depending on the sizes of the analyzed and reference datasets, the number of simulated individuals, and the speed of the processor. Note also that the computations based on the algorithm of Paetkau et al. (2004) are more time consuming than those based on the algorithm of Cornuet et al. (1999) or Rannala and Mountain (1997).

Management of Missing Data

Missing data were treated as follows. A locus is excluded from all computations when one or more reference population samples have no observation (i.e., genotypes) at this locus. On the other hand, when the to-be-assigned entity (individual or group of individuals) was genotyped for several loci but not for a locus *l*, then computations are done using all loci except locus *l*. A list of used and unused loci is given for each individual computation in the output file. It is worth mentioning that criterion values are not comparable among individuals when based on a different number of loci.

Program Features

Input Data Files

For the simple computation of criteria, a data file containing a mixture of diploid or haploid data is accepted, and the level of ploidy is taken into account during computations. However, probabilities based on Monte Carlo resampling methods are computable only for data files containing diploid or haploid data and not a mixture of both. The file formats accepted by GENECLASS2 are those used by the following population genetics software programs: GENEPOP (Raymond and Rousset 1995), GENETIX (Belkhir et al. 1996–2001), and FSTAT (Goudet 1995). GENECLASS2 also converts input data files into any of the three above file formats.

Datasets of virtually any size can be treated, provided the computer has enough memory to load the data and make the computations. Two files are needed for the assignment of individuals not included in the reference population sample. One file contains the reference dataset, the other the individuals or groups of individuals to be assigned. Only a single file is needed when the to-be-assigned individuals are included in the reference population sample (cf. self-assignment or migrants detection).

Output Data Files

The criterion $-\log_{10}(L)$ or the genetic distance for the distance-based method, as well as probabilities, are displayed in a grid with one row per individual and one column per population. When the probability computation option is not used, an adjustable number of best-matching populations are sorted for each individual and a score is given for each population. In a reference file with P populations, the score of an individual i in a population T is computed as follows:

$$Score_{i,T} = \frac{L_{i,T}}{\sum_{j=1}^{P} L_{j,T}}, \qquad (4)$$

with $L_{i,T}$ the likelihood value of the individual i in the population T. In the case of self-assignment, a "quality index" is computed as the mean value of the scores of each individual in the population it belongs to.

All results can be printed, or saved in a CSV format (i.e., values are written by rows, fields separated with semicolons) for further treatment in a spreadsheet (e.g., Microsoft Excel). A tool also included in GENECLASS2 displays basic summary statistics of the analyzed datasets: the number of alleles and genes per population and locus, allelic frequencies, heterozygotes proportions, and Nei's gene diversity (i.e., expected heterozygosity) (Nei 1987).

Running Environment

GENECLASS2 was developed in the Pascal object programming language and compiled with Borland Delphi6 and Kylix2. Therefore the software can be run on a Microsoft Windows or a Linux platform. An easy-to-use graphics interface has been designed to guide the user in the assignment process: choice of task (assign/exclude population as origin of individuals or detection of migrants), choice of statistical criterion for likelihood estimation, computation of probabilities, etc. The package includes a user-friendly help file with graphical interfaces that explain how to run the program to perform the two above tasks.

Program Availability

GENECLASS2 is freely available in English or French at http://www.montpellier.inra.fr/CBGP/softwares. A self-extracting setup executable leads the user through installation of the software on Windows-based machines. An RPM file containing the program file and the libraries allows the package to be installed on Mandrake and Red Hat Linux platforms. A .tar.Z archive allows manual installation of the binaries on other kinds of Linux platforms. A registration form allows users to be kept informed of new releases.

Acknowledgments

This work was supported by grants from the CIRAD/INRA on "Approches biomathématiques et biotechnologiques pour l'identification génétique et la gestion adaptée des populations animales et végétales" (to S.P., J.-M.C., and L.B.) and from the INRA SPE department on methods associated with real-time population genetics (to J.-M.C. and A.E.).

References

Baudouin L and Lebrun P, 2000. An operational bayesian approach for the identification of sexually reproduced cross-fertilized populations using molecular markers. In: Proceedings of the International Symposium on Molecular Markers for Characterizing Genotypes and Identifying Cultivars in Horticulture, Montpellier, France, March 6–9, 2000 (Doré C, Dosba F and Baril C, eds). Leuven, Belgium: International Society for Horticultural Science; 81–93.

Belkhir K, Borsa P, Chikhi L, Raufaste N, and Bonhomme F, 1996–2001GENETIX 4.02, logiciel sous Windows TM pour la génétique des populations, Laboratoire Génome, Populations, Interactions; CNRS UMR 5000; Université Montpellier II, Montpellier, France.

Cavalli-Sforza LL and Edwards AWF, 1967. Phylogenetic analysis: models and estimation procedures. Am J Hum Genet 19:233–257.

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

Efron B, 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc 78:316–331.

Estoup A and Angers B, 1998. Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations. In: Advances in molecular ecology (Carvalho G, ed). Amsterdam: IOS Press; 55–86.

Estoup A, Rousset F, Michalakis Y, Cornuet JM, Adriamanga M, and Guyomard R, 1998. Comparative analysis of microsatellite and allozyme markers: a case study investigating microgeographic differentiation in brown trout (*Salmo trutta*). Mol Ecol 7:339–353.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, and Feldman MW, 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci USA 92:6723–6727.

Goudet J, 1995. FSTAT version 1.2: a computer program to calculate Fstatistics. J Hered 86:485–486.

Nei M, 1972. Genetic distance between populations. Am Nat 106:283–291.

Nei M, 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Nei M, Tajima F, and Tateno Y, 1983. Accuracy of estimated phylogenetic trees from molecular data. J Mol Evol 19:153–170.

Paetkau D, Calvert W, Stirling I, and Strobeck C, 1995. Microsatellite analysis of population structure in Canadian polar bears. Mol Ecol 4: 347–354.

Paetkau D, Slade R, Burden M, and Estoup A, 2004. Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. Mol Ecol 13:55–65.

Rannala B and Mountain JL, 1997. Detecting immigration by using multilocus genotypes. Proc Natl Acad Sci USA 94:9197–9201.

Raymond M and Rousset F, 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J Hered 86:248–249.

Stephens JC, Briscoe D, and O'Brien SJ, 1994. Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. Am J Hum Genet 55:809–824.

Takezaki N and Nei M, 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics 144:389–399.

Received December 15, 2003
Accepted May 20, 2004

Corresponding Author: Sudhir Kumar