**Program Note**

**IDENTIX, a software to test for relatedness in a population using permutation methods**

**KHALID BELKHIR,† VINCENT CASTRIC† and FRANÇOIS BONHOMME‡**  
*Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université Montpellier II, Case Courrier 63, Place E. Bataillon, 34095 Montpellier CEDEX 5 France, Département de biologie, Université Laval, Ste-Foy, Québec, Canada, G1K 7P4*

**Abstract**

The computer program **IDENTIX** estimates relatedness in natural populations using multilocus genotypic data. Queller & Goodnight’s (1989) and Lynch & Ritland’s (1999) estimators of pairwise relatedness are implemented, as well as the identity index of Mathieu et al. (1990). Estimates of the confidence intervals around these pairwise values are also provided. The null hypothesis of no relatedness (multilocus genotypes are independent draws from a panmictic population) is tested using a permutation method that compares the observed distribution of the moments of pairwise relatedness coefficients to that expected in unstructured population.

**Keywords**: bootstrap, computer program, jackknife, natural populations, permutation test, relatedness

Received 5 March 2002; revision received 27 May 2002; accepted 18 June 2002

Measures of relatedness between pairs of individuals have helped to address key issues across a wide spectrum of evolutionary and conservation biology studies. For example, the heritability of quantitative traits can be estimated in natural populations, provided that accurate estimates of the true — but unknown — genealogical relationships are available (Ritland 2000). Measures of genetic relatedness are also fundamental to the study of social systems, including the evolution of altruistic behaviour (Hamilton 1964), social structure (Blouin et al. 1996) and mating systems (Landry et al. 2001). In conservation programmes for endangered species there is also a need to identify close relatives in order to avoid consanguineous mating and to reduce deleterious consequences of inbreeding and the loss of genetic variation (Avise 1995).

One can proceed by attempting to infer the relationships among each pair of individuals in a sample and test their reliability by a jackknife or a bootstrap method, as implemented here following Stone & Björklund (2001) and Queller & Goodnight (1989). Alternatively, it will often suffice to test for the existence of some level of relatedness among the members of a given population sample against the null hypothesis of complete unrelatedness as expected in a random sample from a panmictic population. This latter approach is implemented in the present computer program by a permutation resampling test.

Several relatedness estimators have been proposed in the literature and these have contrasting behaviours in different contexts (Van de Casteel et al. 2001). **IDENTIX** implements three estimators, leaving the user the choice of the estimator to be used in any particular context. Relatedness can be defined using two alternative definitions, both highlighting different aspects of sexual reproduction processes.

First, because two closely related individuals are more likely to share identical alleles by descent than are non-related individuals, one can use $r_{xy} = 2p_x p_y$, where $p_x$ and $p_y$ are the frequencies of an allele at a locus in two individuals. This coefficient may be estimated from the observed identity by state of the alleles carried by two individuals. For a given locus, Lynch & Ritland (1999) defined their estimator as:

$$r_{xy} = \frac{p_x (S_x + S_y) + p_y (S_x + S_y) - 4p_x p_y}{(1 + S_x) + (1 + S_y) - 4p_x p_y},$$

where $S_x$ and $S_y$ are the number of unique alleles at a locus in two individuals. For full-sib relationships, $r_{xy} = 0.5$ for full-sibs and $r_{xy} = 0$ for unrelated individuals in an infinitely large panmictic population.
where \( S_{xy} = 1 \) if individual \( x \) is homozygous and 0 otherwise. \( S_{xy} = 1 \) if allele \( a \) from \( x \) is the same as the allele \( c \) from individual \( y \), \( p_i \), and \( p_j \) are the frequencies of alleles \( a \) and \( b \) in the sample.

A multilocus estimate is obtained by weighting each contributing locus by:

\[
W_{xy} = \frac{(1 + S_{xy})(p_x + p_y) - 4p_x p_y}{2p_x p_y}
\]

A symmetrical estimate is obtained by averaging the two reciprocal estimates \( W_{xy} \) and \( W_{yx} \) as \( 0.5 \times (W_{xy} + W_{yx}) \).

Under the same definition of relatedness another estimator was derived previously by Queller & Goodnight (1989) and has been expressed using the same notation as in Lynch & Ritland (1999):

\[
r_{xy} = \frac{0.5(S_{xy} + S_{yx} + S_{xy} + S_{yx}) - p_x - p_y}{1 + S_{xy} - p_x - p_y}
\]

However, this formula is undefined (the denominator = 0) for heterozygous individuals \( (S_{xy} = 0) \) at a biallelic locus \( (p_x + p_y = 1) \) so IDENTIX implements the original multilocus formula given in Queller & Goodnight (1989) for pairwise comparisons.

Second, because closely related individuals are also more likely to produce homozygous offspring, one can alternatively measure the effects of relatedness as the expected proportion of loci that are homozygous in the offspring of the chosen pair of individuals (Mathieu et al. 1990), a quantity which would estimate well the consanguinity of the offspring in cases where identical alleles are likely to be identical by descent, something especially relevant when interested in the fitness consequences of consanguineous matings. At a given locus, \( I \) = identity is classically defined as:

\[
I_{xy} = \frac{n_{yx}}{\sqrt{n_x n_y}}, \quad \text{where} \quad n_{xy} = \frac{\sum n_i n_{ij}}{2}
\]

where \( n_i \) is the number of copies of allele \( j \) in individual \( x \).

This index can be rewritten following the Lynch & Ritland (1999) formulation as:

\[
I_{xy} = \frac{(S_{xy} + S_{yx} + S_{xy} + S_{yx})}{2} - 1 + S_{xy}^{-1/2} + S_{xy}
\]

The multilocus estimate is obtained by weighting loci by \( 1/\Sigma(p_i)^2 \) where \( p_i \) is the sample frequency of allele \( j \) at locus \( l \).

In order to compare the sampling variances for the three estimators, we used Monte Carlo simulations to generate multilocus genotypes with varying degrees of relationships.

To investigate the sensitivity of each estimator to the number of loci, number of alleles at each locus and to different types of allelic distributions, a factorial design was used consisting of three levels for the number of loci \( (1, 5 \) and \( 10) \), a range between 3 and 20 for the number of alleles and three types of allelic distributions (equifrequent, random and ‘triangular’). For a locus with \( n \) alleles, equifrequency is obtained by assigning a proportion of \( 1/n \) to each allele. Triangular distribution gives the relative weights of 1, 2, …, \( n \) to the alleles, while for the random distribution, allele frequencies are randomly generated from a uniform distribution.

For each combination of the number of loci \( \times \) number of alleles \( \times \) allelic distribution we generated multilocus genotypes randomly for 1000 pairs of unrelated individuals, 1000 pairs of full-sibs and 1000 pairs of half-sibs, successively assuming a random population mating and unlinked loci.

The variance over the 1000 values for each estimator was obtained for each set of 1000 pairs of genotypes. This process was repeated 100 times for each combination of parameters.

Figure 1 reveals that for a single locus, sampling variance of Queller & Goodnight and Lynch & Ritland relatedness coefficients declines with increasing allele number for the various relatedness levels as well as allele-frequency distributions. On the other hand, the identity index resulted in smaller sampling variance in almost every situation and over a wide range of allele numbers. Those patterns were reproduced for simulations among five and 10 loci.

Once estimates of pairwise relatedness have been obtained, there still remains a need to test for their significance. IDENTIX fulfils this need at two different levels by applying different resampling techniques. First, for each pairwise estimate a confidence interval is obtained by a bootstrap or a jackknife over loci. Second, to determine globally whether individuals in the given sample are genetically more related than expected, given that their parents had mated randomly, IDENTIX implements a test of the null hypothesis of no relatedness by comparing the distribution of the moment of pairwise relatedness coefficients in a population sample with its null expectation.

This null distribution is obtained by a conventional Monte Carlo resampling procedure (Guo & Thompson 1992), which proceeds by randomly selecting either 2 \( N \) alleles or \( N \) genotypes without replacement, independently for each locus, assigning them at random to the \( N \) individuals, then recalculating the statistic. In the case of a population which departs from Hardy–Weinberg equilibrium (as can be created by consanguineous mating systems) resampling must be carried out at the genotypic rather than at the allelic level. Ho is rejected with a significance level of 5% if the observed value of the statistic is above the 95% level of the resampled statistics.
The arithmetic mean of the $N^2(N-1)/2$ pairwise relatedness coefficients from a sample can be compared directly to the proportion of resampled means to determine whether individuals within this sample are genetically more related than expected under the null hypothesis. Whenever the mean of the distribution does not differ from its null expectation, the possibility remains that distributions may differ with respect to their variance. Significantly higher variance in the observed pairwise relatedness coefficients could indicate that several independent groups of related individuals were sampled. In this case, pairwise comparisons involve either related or unrelated individuals, a pattern that would increase the variance of the relatedness distribution.

As an example of how permutation tests can be used to detect relatedness, we considered an application from a natural population of a lacustrine fish. Thirty-four adult brook charr (*Salvelinus fontinalis*) originating from Clish Pond (85 ha), located in northern Maine (USA), were genotyped at six polymorphic microsatellite loci (Castric et al. 2002). Significant departures from Hardy–Weinberg proportions were found, with a strong heterozygote deficit ($F_{IS} = 0.157$) that was also highly variable across loci. One possible explanation for the deficit may include sampling biases, whereby individuals collected originated from very few families. If so, fish should be genetically more related than expected at random. The mean pairwise Identity in the population was 0.4668, a value only reached with a probability of 2.4% under the assumption of no relatedness (Fig. 2). We therefore concluded that fish within this population sample are genetically more related than expected in a randomly mating population, a result consistent with the hypothesis that sampled fish were not a random draw from a large panmictic population but rather a subset from a limited number of families only.

IDENTITY input files are fully compatible with the package GENETIX (Belkhir et al. 2001), a complete software package...
for population genetic data analyses running under Windows™ that will also import/export files from/to other commonly used packages, and is available at http://www.Univ-montp2.fr/~genetix/genetix.htm. This stand-alone program was written in delphi 4.0 and runs very swiftly on Windows platforms. Several options are available. For the computation of pairwise estimates and confidence limits, the user is allowed to specify the estimator to be used, the level of precision of computations, and to choose the samples and loci that are to be included in the analysis. For the permutation test, the user can specify the estimator, the type of randomization to perform (permutation across alleles or across genotypes), the number of randomized pseudo-samples to generate, and whether the mean or variance should be compared to their respective simulated distributions. The resulting distributions can be visualized using the included graphic option. The program is available at http://www.Univ-montp2.fr/~genetix/identix01.zip.

Acknowledgements

We thank N. Raufaste for valuable help in homogenizing estimators’ notations. We are indebted to Sean Rogers for providing editorial advice. This is a contribution to the research programme of GIROQ.

References