

IDENTIX, A SOFTWARE TO TEST FOR RELATEDNESS IN A POPULATION USING PERMUTATION METHODS.

Khalid Belkhir ¹

Vincent Castric ²

François Bonhomme ¹

¹ *Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université Montpellier II, Case Courrier 63, Place E. Bataillon, 34095 Montpellier CEDEX 5 France*

² *Département de biologie, Université Laval, Ste-Foy, Québec, Canada, G1K 7P4*

Keywords : Relatedness, permutation test, Bootstrap, Jackknife, computer program, natural populations.

Correspondance: Khalid Belkhir *Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université Montpellier II, Case Courrier 63, Place E. Bataillon, 34095 Montpellier CEDEX 5 France. Fax : 04 67 14 45 54. mail : belkhir@univ-montp2.fr*

Abstract.

Identix is a computer program to detect 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 Identity (the expected proportion of loci that are homozygous in the offspring of a pair of individuals). Estimate of the confidence intervals around the observed 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 pairwise relatedness coefficients to that expected in unstructured populations.

Measures of relatedness between pairs of individuals have helped 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 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 programs 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 all pairs of individuals in a sample; or 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. This is an alternative to the jackknife or bootstrap methods implemented either by Stone and Björklund (2001) or by Queller and Goodnight (1989) to give the confidence interval for relatedness values.

Three estimators

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, we use r_{xy} = twice the probability that a random gene taken from individual x is identical by descent with a random gene taken from individual y at a diploid autosomal locus. (e.g., $r_{xy} = 0.5$ for full-sib relationships, $r_{xy} = 0.25$ for half sibs and $r_{xy} = 0$ for unrelated individuals in an infinitely large panmictic population). 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 :

$$r_{xy} = \frac{p_a(S_{bc} + S_{bd}) + p_b(S_{ac} + S_{ad}) - 4p_a p_b}{(1 + S_{ab}) + (p_a + p_b) - 4p_a p_b}$$

where $S_{ab} = 1$ if individual x is homozygous and 0 otherwise. $S_{ac} = 1$ if allele a from x is the same as the allele c from individual y . p_a and p_b are the frequencies of alleles a and b in the sample.

A multilocus estimate is obtained by weighting each contributing locus by

$$W_{xy} = \frac{(1+S_{ab})(p_a+p_b)-4p_a p_b}{2p_a p_b}$$

A symmetric estimate is obtained by averaging the two reciprocal estimates r_{xy} and r_{yx} as $0.5.(r_{xy} + r_{yx})$.

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

$$r_{xy} = \frac{0.5(S_{ab}+S_{ad}+S_{bc}+S_{bd})-p_a-p_b}{1+S_{ab}-p_a-p_b}$$

However, this formula is undefined (the denominator = 0) for heterozygous individuals ($S_{ab} = 0$), for biallelic loci or equifrequent alleles ($p_a + p_b = 1$) so Identix implements a pairwise adaptation of the original multilocus formula given in Q&G (1989).

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 well estimate the consanguinity of the offspring in cases where identical alleles are likely identical by descent, something especially relevant when interested in the fitness consequences of consanguineous matings. At a given locus, Identity is classically defined as

$$I_{xy} = \frac{i_{xy}}{\sqrt{i_{xx}i_{yy}}}, \text{ where } i_{xy} = \frac{\sum_j n_{jx}n_{jy}}{2}$$

Where n_{jx} is the number of copies of allele j in individual x .

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

$$I_{xy} = \frac{(S_{ab}+S_{ad}+S_{bc}+S_{bd})}{2*\sqrt{1+S_{ab}}*\sqrt{1+S_{cd}}}$$

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

A statistical test of pairwise relatedness coefficients

Once an estimate of pairwise relatedness has been obtained, there still remains a need to test for its significance, i.e. determine whether individuals in a given cohort are genetically more related than expected given 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 with its null expectation. This null distribution is obtained by a resampling procedure, which proceeds by randomly selecting either $2N$ 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 done at the genotypic rather than at the allelic level. H_0 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 pairwise relatedness coefficients in a sample can be directly compared to that in random draws from a large panmictic population 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.

Using Identix on simulated data

We used Monte Carlo simulations to generate multilocus genotypes with varying degrees of relationships and to compare the sampling variances for the three estimators.

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 $1, 2, \dots, n$ 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 randomly generated multilocus genotypes 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, the Queller & Goodnight and Lynch & Ritland relatedness coefficients have declines in sampling variance with increasing alleles 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 5 and 10 loci.

Using Identix on empirical data

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. submitted). 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 (Figure 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.

Identix input files are fully compatible with the package Genetix, a complete software package for population genetic data analyses running under WindowsTM that will also Import/Export files from/to other commonly used packages and is available at www.univ-montp2.fr/~genetix/genetix.htm. This standalone program was written in DELPHI 4.0 and runs very fast 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 randomisation to perform (permutation across alleles or across genotypes), the number of randomised pseudo-samples to generate, and whether the mean or variance should be compared to their respective simulated distributions. The resulting distributions can be visualised using the included graphic option. The program is available at <ftp://162.38.181.25/pub/identix.zip>.

Acknowledgements

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

Literature cited

- Belkhir K, Borsa P, Chikhi L, Raufaste N & Bonhomme F(2001). GENETIX 4.02, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier (France).
- Blouin MS, Parsons M, Lacaille V and Lotz. S (1996). Use of microsatellite loci to classify individuals by relatedness. *Molecular Ecology*. 5, 393-401.
- Castric V, Belkhir K, Bernatchez L, Bonhomme F (2002). Heterozygote deficiencies in lacustrine brook charr (pisces, salmoninae) : A test of alternative hypotheses. *in press Heredity*.
- Hamilton WD (1964). The genetical evolution of social behaviour. I & II. *Journal of Theoretical Biology*. 7, 1-52.
- Landry C, Garant D, Duchesne P, Bernatchez L (2001). "Good genes as heterozygosity": MHC and mate choice in Atlantic salmon (*Salmo salar*). *Proc R Soc Lond B Biol Sci* **268**(1473): 1279-85.
- Lynch M., Ritland K (1999). Estimation of pairwise relatedness with molecular markers. *Genetics* 152, 1753-1766.
- Mathieu E., Autem M, Roux M, Bonhomme F (1990). Épreuves de validation dans l'analyse de structures génétiques multivariées : comment tester l'équilibre panmictique ? *Revue de Statistique Appliquée* 38(1), 47-66.
- Queller DC., Goodnight KF (1989). Estimating relatedness using genetic markers. *Evolution* 43(2), 258-275.
- Ritland K (1996). Estimators for pairwise relatedness and inbreeding coefficients. *Genetical Research* 67, 175-186.
- Ritland K (2000). Marker-inferred relatedness as a tool for detecting heritability in nature. *Molecular Ecology* 9, 1195-1204.
- Stone J, Björklund M (2001). DELRIOUS: a computer program designed to analyse molecular marker data and calculate delta and relatedness estimates with confidence. *Molecular Ecology Notes* 1, 209-212.
- Thompson EA (1975). The estimation of pairwise relatedness. *Annals of Human Genetics*. 39, 173-188.
- Van De Casteele T, Galbusera P, Matthysen E (2001). A comparison of microsatellite-based pairwise relatedness estimators. *Molecular Ecology* 10, 1539-1549.

28/03/2002

Figures

Figure 1. Monolocus sampling variance for the three estimators. Solid lines, large dashes and dot dashed lines denote estimates for Identity, Queller & Goodnight and Lynch & Ritland, respectively.

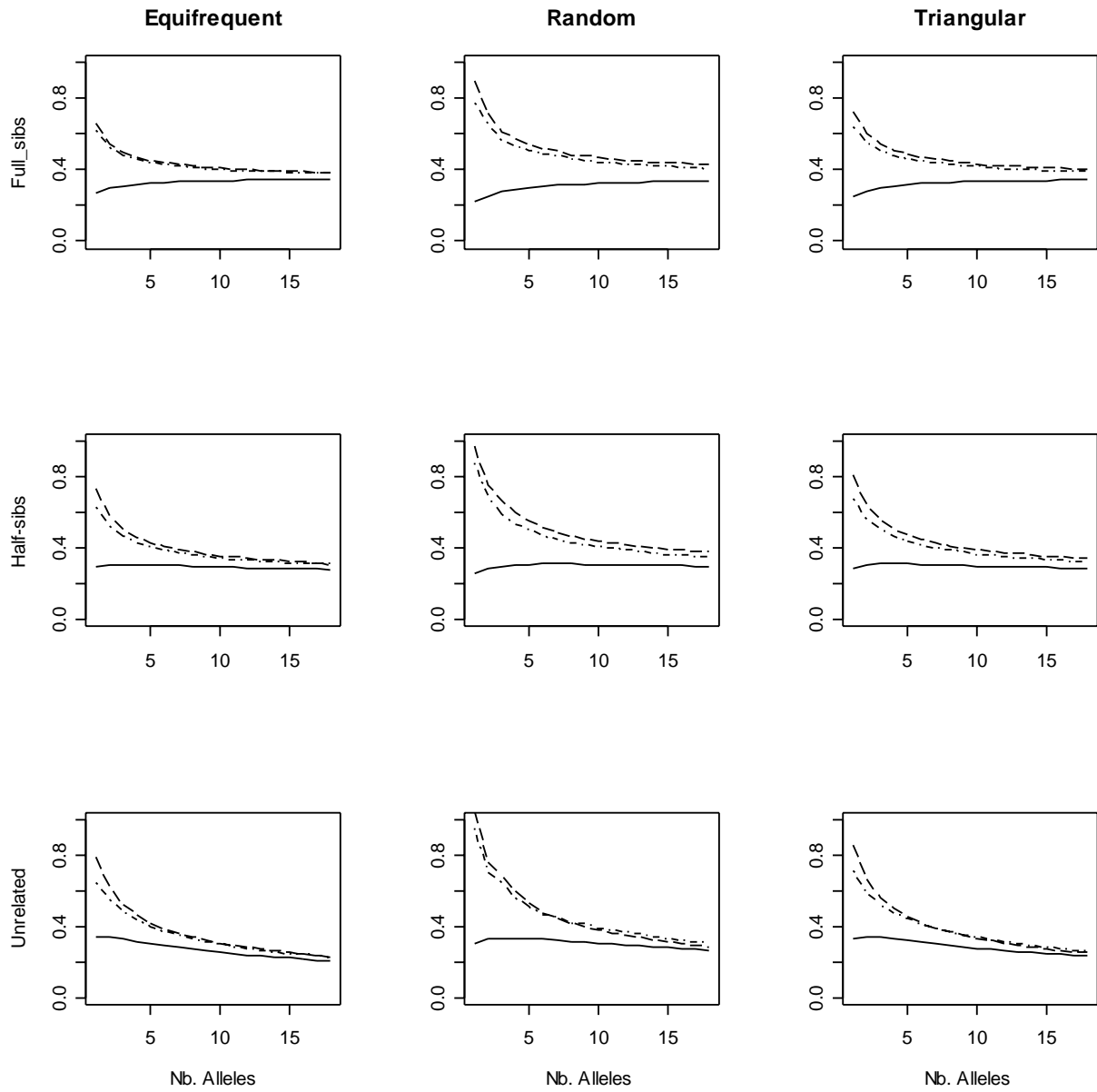


Figure 2. An example of the permutation test in a natural population of a freshwater fish *Salvelinus fontinalis* Mitchill. The arrow's position indicates that the observed mean of identity departs from its expectation in randomised populations.

Fig 2

