Molecular Ecology Notes (2006) **6**, 576–579 doi: 10.1111/j.1471-8286.2006.01256.x

PROGRAM NOTE

ML-RELATE: a computer program for maximum likelihood estimation of relatedness and relationship

STEVEN T. KALINOWSKI, AARON P. WAGNER and MARK L. TAPER
Department of Ecology, Montana State University, Bozeman, Montana 59717, USA

Abstract

Genetic data are useful for estimating the genealogical relationship or relatedness between individuals of unknown ancestry. We present a computer program, ML-RELATE that calculates maximum likelihood estimates of relatedness and relationship. ML-RELATE is designed for microsatellite data and can accommodate null alleles. It uses simulation to determine which relationships are consistent with genotype data and to compare putative relationships with alternatives. ML-RELATE runs on the Microsoft Windows operating system and is available from www.montana.edu/kalinowski.

Keywords: maximum likelihood, null allele, relatedness, relationship, software

Received 13 September 2005; revision received 27 October 2005; accepted 22 November 2005

Genetic data are frequently used to estimate the genealogical relationship or relatedness between individuals of unknown ancestry (see Blouin 2003 for review). Ecological studies estimating relatedness can be classified according to the number of individuals in the pedigree being estimated. The simplest case is estimating the relationship (or relatedness) between two individuals. This is the case we discuss in this paper, and the case analysed by the computer program, ML-RELATE, that we describe below. Other statistical methods will be more useful for other applications (e.g. paternity tests when the maternal genotype is available).

There are several computer programs available to estimate relatedness and relationship (See Blouin 2003 Table 1 for a comparison of programs) — enough that choosing a program can be daunting. We wrote ML-RELATE because we wanted a software that met three criteria. First, we wanted a stand-alone program that runs on the Microsoft Windows operating system. Second, we wanted to calculate maximum likelihood estimates of relatedness. Milligan (2003) recently showed that maximum likelihood estimates of relatedness usually have a lower root mean squared error than other estimators. Third, we wanted a program that could accommodate null alleles. Null alleles are common in microsatellite data, and can easily lead to errors in estimating relatedness or relationship if their potential presence is not accommodated in relatedness or relationship calculations (Dakin & Avise 2004; Wagner *et al*. 2006).

Correspondence: S. T. Kalinowski, Fax: (406) 994 3190; E-mail: skalinowski@montana.edu

ML-RELATE is described in detail in a user's manual (available at www.montana.edu/kalinowski). Here, we focus on the statistical analysis performed by the program.

Genealogical relationships between individuals are conveniently represented mathematically as probabilities that genotypes in the individuals share zero, one or two alleles identical by descent (see Lynch & Walsh 1998; Blouin 2003; or Buckleton *et al*. 2005 for reviews). A few examples illustrate this translation. Let k_0, k_1 and k_2 represent the probabilities that two individuals share zero, one or two alleles (respectively) at a locus. If two individuals are parent–offspring, k_0 will equal 1, k_1 will equal 0 and k_2 will equal 0. If two individuals are full-siblings, k_0, k_1 and k_2, will equal 0.25, 0.5 and 0.25, respectively (see Table 1 for other relationships commonly of interest). If individuals are inbred, additional coefficients are needed to represent the genealogical relationship between the individuals (e.g. Milligan 2003). Here we make the typical (e.g. Thompson 1991; Milligan 2003) assumption that neither of the two individuals being compared is inbred, so three *k*-coefficients are sufficient. We note, however, that the consequences of violating this assumption have not been explored. We also assume a closed population (i.e. no migrants entering population).

The *k*-coefficients representing the genealogical relationship between two individuals can be plotted as a point on a graph (Fig. 1). The three *k*-coefficients must sum to one, so only two need to be plotted. If the individuals are noninbred (as we are assuming), there is the additional restriction that $k_1^2 \geq 4k_0k_2$ (Thompson 1991). When these restrictions are considered, the range of possible values for *k*-coefficients,

Table 1 A list of *k*-coefficients for common relationship categories

Relationship	k_0	k_1	k_2
Parent–Offspring	0	1	0
Full-siblings	0.25	0.50	0.25
Half-siblings grandchild–grandparent Niece/Nephew–Uncle/Aunt	0.50	0.50	0
First cousin	0.75	0.25	0
Unrelated	1	0	0

k_m represents the probability that two individuals share m alleles IBD under a given relationship.

the '*k*-space', is roughly triangular, except that one side of the triangle is curved concave inward. A plot of common relationships shows that three relationships define the extremes of the *k*-space: unrelated (U), parent/offspring (PO) and monozygotic twin (M). In general, relationships that are close to each other in *k*-space are relatively difficult to differentiate. For example, if sufficient genetic data are not collected, first cousins (1C), may be mistakenly be identified as unrelated (U) or half-sibs (H), but are less likely to be identified as parent/offspring (PO) (Fig. 1).

There is a convenient relationship between *k*-coefficients and the coefficient of relatedness, *r*, between two individuals. If the individuals are not inbred, *r* is equal to

$$r = \frac{1}{2}k_1 + k_2. \quad (1)$$

The relationship between two individuals can be estimated from genetic data by evaluating the likelihood of points in *k*-space (see Lynch & Walsh 1998; Blouin 2003; Buckleton *et al*. 2005; for reviews). For notational simplicity, let the vector **K** represent three *k*-coefficients $\mathbf{K} = \{k_0, k_1, k_2\}$. By definition, the likelihood of **K** is equal to the probability of observing the genetic data present in two individuals having relationship **K**. Let $L(\mathbf{K})$ represent this likelihood. Several authors have given formulae for $L(\mathbf{K})$ (e.g. Thompson 1991; Milligan 2003; Wagner *et al*. 2006); and we will not review them here. The maximum likelihood estimate of *r* between two individuals is found by searching the entire parameter space of **K**, finding the values that maximize the likelihood, and then inserting these values into Equation 1. The simplex optimization routine is useful for searching for the maximum likelihood value of **K** (e.g. Press *et al*. 1992). The likelihood surface can have multiple 'peaks' of varying heights, so starting the search from multiple starting points is often necessary to ensure that the highest peak is located (S. Kalinowski, unpublished results). The maximum likelihood relationship between a pair of individuals is found by inserting different values of **K** into the likelihood equation (each corresponding to a relationship of interest) and determining which yields the highest likelihood.

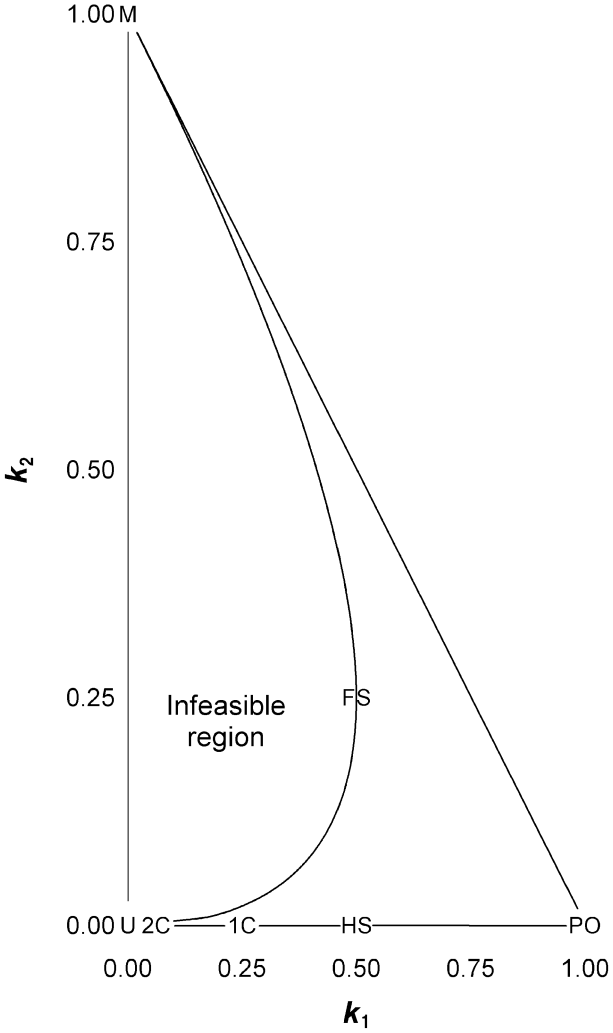

Fig. 1 The *k*-coefficients of several common genealogical relationships (U, unrelated; HS, half-siblings; 1C, first cousins; 2C, second cousins; FS, full-siblings; PO, parent–offspring; M, monozygotic twin). The figure is constructed so that relatedness increases towards the right and towards the top of the figure.

Unless large numbers of loci are scored (e.g. 30–40 microsatellite loci), estimates of relatedness or relationship are strongly affected by sampling error — in particular, interlocus variance of gene identity. Therefore, statistical methods are needed to assess the uncertainty surrounding estimates of relationship or relatedness. We present two methods for doing this: a statistical approach for testing between two a priori hypotheses, and a method for determining what relationships are consistent with genetic data (i.e. a method for constructing confidence sets for the relationship between two individuals). These approaches are either identical or similar to methods commonly used for pedigree analysis (e.g. Marshall *et al*. 1998; Goodnight & Queller 1999; McPeek & Sun 2000).

The statistical test to evaluate two competing a priori hypotheses is illustrated with an example. A researcher

observes two adult female hyenas in the proximity of a den and suspects that the hyenas are sisters but would like to exclude the possibility that they are unrelated. Genetic data are collected and the relationship full-siblings has the highest likelihood. However, the likelihood for unrelated is not much lower. Therefore, the researcher might wonder if the hyenas are actually unrelated, but appeared genetically to be siblings by chance. This question can be addressed with the following statistical test. Let $\mathbf{K}_{Putative}$ represent the *k*-coefficients for the putative relationship between the hyenas (in this case, full siblings), and let $\mathbf{K}_{Alternative}$ represent the *k*-coefficients for a alternative hypothesis of interest. We define the test statistic, Λ , equal to

$$\Lambda = \mathrm{Ln}\left[\frac{L(\mathbf{K}_{Putative})}{L(\mathbf{K}_{Alternative})}\right]. \quad (2)$$

The sampling distribution of Λ is determined through simulation. Genotypes are simulated for the alterative hypothesis (in this case, unrelated individuals) and Λ is calculated for each simulated pair. (These simulations assume that the allele frequencies in the population are equal to the allele frequencies in the sample. Genotypes are simulated in two steps. First, the number of alleles identical by descent is chosen from **K**, and then genotypes are chosen given **K**.) This is performed a large number of times, and the proportion of times that the simulated Λ is greater or equal to the observed Λ is recorded. This is the *P* value for the hypothesis test. If this *P* value is small, the alternative hypothesis is rejected. If the *P* value is large, both the putative and alternative relationships are consistent with the data.

In many circumstances, researchers will not have a putative relationship to test against an alternative. In such cases, a list of all plausible relationships between two individuals is useful. Such a list can be constructed by testing several relationships and determining which are consistent with the observed data. Again, we use a likelihood ratio for a test statistic and simulation to estimate its distribution. Let $\mathbf{K}_{Null}$ represent a possible relationship between two individuals to be tested for plausibility. Let $\mathbf{K}_{ML}$ represent the maximum likelihood estimate of **K** for the two individuals. and let the test statistic Λ' equal

$$\Lambda' = \mathrm{Ln}\left[\frac{L(\mathbf{K}_{ML})}{L(\mathbf{K}_{Null})}\right]. \quad (3)$$

The significance of Λ' observed between two individuals is established via simulation. Genotype pairs are simulated for the null hypothesis and Λ' is calculated for each of the simulated pair of individuals. This is performed a large number of times, and the proportion of times that the simulated Λ' is greater or equal to the observed Λ' is recorded. This is the *P* value for the hypothesis test. If this *P* value is small, the null hypothesis can be rejected. If the *P* value is large, the null hypothesis is not rejected and that relationship is included in the list of relationships that are considered consistent with the observed data. This process is repeated for all relationships of interest.

Estimates of relatedness and relationship can be biased by null alleles (e.g. Dakin & Avis 2004; Wagner *et al*. 2006). The problem is most severe for relationship estimation. Consider as an example a cross between a dam with genotype *ii* and a sire with genotype *jn* (where *n* is a null allele and *i* and *j* are non-null alleles). With this pair of parents, there is a 50% chance that an offspring will have the genotype *in*. In this circumstance, the apparent genotype of each individual will be: dam (*ii*), sire (*jj*) and offspring (*ii*). If the potential presence of null alleles is not accounted for, these genotypes will exclude the actual sire of the juvenile from being a parent. Likelihood equations, however, are easily modified to account for genotypes for null alleles (Wagner *et al*. 2006).

ML-RELATE is a computer program for estimating relatedness and relationship from codominant genetic data (such as microsatellites). In addition to calculating maximum likelihood estimates of relatedness and relationship, it performs the two hypothesis tests described above. Null alleles can be accommodated in all calculations. ML-RELATE is available from www.montana.edu/kalinowski, runs on the Microsoft Windows operating system, and reads data files in the GENEPOP format (Raymond & Rousset 1995). Most calculations performed by ML-RELATE require less than a second of computation time for a typical desktop computer, but can become more lengthy if relatedness is estimated for all pairs of individuals in a data set. For example, calculating a matrix of pairwise relatedness estimates for 59 individuals from eight microsatellite loci required approximately 15 s of computation.

Several steps were taken to check for errors in the calculations performed by ML-RELATE. Hardy–Weinberg tests for the presence of null alleles were compared to results produced by GENEPOP (Raymond & Rousset 1995). Relatedness estimates were checked with calculations performed using Microsoft Excel. In addition, we replicated some of the simulations by Milligan (2003). Last, we used the program to analyse a hyena microsatellite data set (A. Wagner, unpublished).

References

Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology & Evolution*, **18**, 503–511.

Buckleton J, Triggs CM, Walsh SJ (2005) *Forensic DNA Evidence Interpretation*. CRC Press, Boca Raton, Florida.

Dakin EE, Avise JC (2004) Microsatellite null alleles in parentage analysis. *Heredity*, **93**, 504–509.

Goodnight KF, Queller DC (1999) Computer software for performing likelihood tests of pedigree relationship using genetic markers. *Molecular Ecology*, **8**, 1231–1234.

Lynch M, Walsh B (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, Massachusetts.

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

McPeek MS, Sun L (2000) Statistical tests for detection of misspecified relationships by use of genome screen data. *American Journal of Human Genetics*, **66**, 1076–1094.

Milligan BG (2003) Maximum-likelihood estimation of relatedness. *Genetics*, **163**, 1153–1167.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in C. The Art of Scientific Computing*, 2nd edn. Cambridge University Press, Cambridge.

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

Thompson EA (1991) Estimation of relationships from genetic data. In: *Handbook of Statistics* (eds Rao CR, Chakraborty R), Vol. 8, pp. 255–269. Elsevier Science Publishers, Amsterdam.

Wagner AP, Creel S, Kalinowski ST (2006) Estimating relatedness and relationships using microsatellite loci with null alleles. *Heredity*, (Accepted pending revision).