Molecular Ecology (1999) **8**, 1231–1234

SHORT COMMUNICATION

Computer software for performing likelihood tests of pedigree relationship using genetic markers

K. F. GOODNIGHT and D. C. QUELLER
Department of Ecology & Evolutionary Biology MS-170, Rice University, 6100 Main St., Houston, TX 77005–1892, USA

Abstract

Molecular techniques are making ever more genetic markers available for use in parentage assignment, and measures of relatedness. We present a program, *Kinship*, designed to use likelihood techniques to test for any non-inbred pedigree relationship between pairs of individuals, using single-locus codominant genetic markers. *Kinship* calculates the likelihood that each pair of individuals in a data set are related by a given pedigree hypothesis, and likelihood ratios for any pair of hypotheses. The program also uses a simulation routine to attach statistical significance to its results.

Keywords: kinship, likelihood, parentage, pedigree, relatedness, software

Received 30 November 1998; revision received 4 February 1999; accepted 11 February 1999

Introduction

The use of genetic markers to investigate pedigree relationships has become a common technique in studies of animal populations. Parentage assignment is the most common application of this sort, yielding information on mating systems (Soukup & Thompson 1998), sexual selection (Dufour & Weatherhead 1998) and social organization (Conrad *et al*. 1998). Using genetic markers for such assignments can yield information that would be difficult or impossible to obtain from behavioural observations, e.g. in cases where postmating mechanisms affect paternity (Fitzsimmons 1998) or where mating itself is difficult or impossible to observe (Coltman *et al*. 1998). This note introduces a new program designed to assign parentage and other pedigree relationships using genetic data.

Parentage can be simply excluded or accepted based on the genotypes of the test individuals, because parents must share at least one allele with their offspring (barring mutation). The simplest and most common case for such exclusions is when one parent, usually the mother, is known, and the father must be identified from among several candidates. Knowing the genotype of one parent makes exclusions easier for the other, although exclusions can still be performed when neither parent is known.

Correspondence: K. F. Goodnight. Fax: +1-713-285-5232; E-mail: keithg@rice.edu

Exclusions can be performed using either multilocus fingerprint data, or single-locus markers such as microsatellites. With molecular techniques continually making greater numbers of variable single-locus, codominant markers available, these markers have considerable advantages over multilocus data: they can be consistently scored across multiple gels, and they can be used for unbiased estimates of useful genetic parameters such as relatedness (Queller *et al*. 1993).

Single-locus markers also lend themselves to the use of likelihood methods (Edwards 1972) in identifying parentage and other relationships. Simple exclusion discards information for nonexcluded pairs. To exploit all available data, likelihood calculations have been used both to categorically assign individuals to their parents (Coltman *et al*. 1998; Marshall *et al*. 1998) and to fractionally assign parentage to multiple candidates based on their likelihood (Smouse & Meagher 1994). Marshall *et al*. (1998) present a program, CERVUS, to carry out the assignments.

These studies focus on the case of paternity assignment when the mother is known, but likelihood methods could usefully be applied to other relationships. For example, in many paternity studies using exclusions, a large fraction of offspring cannot be assigned to any genotyped male (e.g. Kempenaers *et al*. 1997, 20–30% of young unassigned in different years; Martinez *et al*. 1998, 20% unassigned). If more than one such unassigned offspring is found in a nest, a test to distinguish full-siblings

from half-siblings could determine whether the unassigned offspring represent one or more additional males. A test to distinguish half-siblings from unrelated individuals could provide the same information for uncollected males in the population as a whole, and even reveal the variation in the absent males' reproductive success. Likelihood tests could provide information on family structure of individuals for which no family data are otherwise available (e.g. individuals already adult when a study begins).

We present a computer program that uses likelihood methods applied to codominant genotype data to test hypothesized relationships among individuals. The program, *Kinship*, is written for Apple Macintosh computers. *Kinship* codes each type of pedigree relationship in terms of the probability of candidate individuals sharing an allele identical by descent from the maternal or paternal line. For example, a mother–offspring pair must share an allele by maternal 'descent' but do not share an allele by paternal descent (barring inbreeding). *Kinship* can perform its calculations on both diploid and haploid individuals and can test haplodiploid relationships as well as diploid.

Table 1 Sample pedigree relationships as coded for use by *Kinship*. R_p represents the probability of individuals so related sharing an allele identical by paternal descent, and R_m the probability of their sharing an allele identical by maternal descent

	R_p	R_m
Mother–offspring	0.0	1.0
Father–offspring	1.0	0.0
Full siblings	0.5	0.5
Full sisters (haplodiploid)	1.0	0.5
Half siblings (maternal)	0.0	0.5
Cousins (maternal)	0.0	0.25
Unrelated	0.0	0.0

Primary hypothesis (true relationship)

Calculations and features

First, take the simplest case of two haploids. Call them X and Y. The population frequencies of the two alleles are designated as P_x and P_y. R specifies the probability that they are identical by descent. If X and Y are identical, the likelihood must be the probability of drawing the first, P_x, times the probability that the second allele is identical, either by descent or by state, $R + (1 - R)P_x$. If X differs from Y, it is the probability of drawing the first allele, times the probability that the second is not identical by descent, times the probability of drawing the second allele by chance: $P_x(1 - R)P_y$. The overall likelihood is the product over all loci (we assume independence of loci).

The case of two diploids uses the same principles, but is more complicated. We still have individuals X and Y, but now each has two alleles, one labelled the maternal allele (X_m, Y_m) and one labelled the paternal allele (X_p, Y_p). We also have two R values: R_m representing the probability that the maternal alleles of X and Y are identical by descent, and R_p representing the same for the paternal alleles. These two R values can be specified for any non-inbred relationship (some examples are given in Table 1). Assume first that we know which alleles of X and Y are maternal and which paternal. Then, because we assume no inbreeding, the likelihood is the product of two independent terms, one for the maternal alleles, and one for the paternal. As in the haploid case, these terms depend on whether the two alleles are identical, and therefore the likelihood calculation follows one of four paths as shown in Table 2.

Although kinship includes some capability for specifying the maternal or paternal alleles when they are known, the more usual case will be that the user does not know which alleles are maternal and which are paternal. When this is true, kinship takes the average of four calculations, based on the four possible assumptions about which alleles are maternal and which paternal (note: these are not the four entries in Table 2; rather Table 2 is invoked four times, once for each possible configuration of maternal and paternal alleles). Again, the total likelihood is the product over all loci of the calculation just described.

Similar principles are used in the case where one individual is diploid and the other haploid.

The resulting likelihood for the hypothesis described

Table 2 Likelihood expressions used by *Kinship* in its calculations at one locus for two diploid individuals X and Y. R_m and R_p define the hypothesized relationship as in Table 1. The alleles are defined as maternal (X_m and Y_m) or paternal (X_p and Y_p). P_{xm}, P_{ym}, P_{xp} and P_{yp} are the population frequencies of the alleles. One of the four expressions in the table is called depending on the patterm of matches/mismatches among the four alleles. If maternal and parernal alleles cannot be discriminated (*Kinship* uses parentage information in the data set to make the discrimination, if such data are included) the table is invoked multiple times, once with each possible permutation of alleles, and the average of the different values is used

	$X_p = Y_p$	$X_p \neq Y_p$
$X_m = Y_m$	$P_{xm}(R_m + (1 - R_m)P_{xm}) \times P_{xp}(R_p + (1 - R_p)P_{xp})$	$P_{xm}(R_m + (1 - R_m)P_{xm}) \times P_{xp}(1 - R_p)P_{yp}$
$X_m \neq Y_m$	$P_{xm}(1 - R_m)P_{ym} \times P_{xp}(R_p + (1 - R_p)P_{xp})$	$P_{xm}(1 - R_m)P_{ym} \times P_{xp}(1 - R_p)P_{yp}$

Null hypothesis	Primary hypothesis (true relationship)				
	Parent	Full-sib	Half-sib	Cousin	Unrelated
Parent	—	3	2	1	1
Full-sib	5	—	7	5	2
Half-sib	5	4	—	19	6
Cousin	4	3	16	—	17
Unrelated	2	2	4	13	—

Table 3 Number of loci required to correctly accept 50% of pairs related by the primary hypothesis at a significance level of $P < 0.05$, for the primary/null comparison shown. Results were obtained using simulated data sets with 20 equally frequent alleles per locus. One-thousand simulated data sets were generated for each comparison

by R_m and R_p is the probability of obtaining the observed genotypes if the hypothesis is true. A single likelihood is not very meaningful; it gets lower as more information is added. However, the likelihood ratios for two hypotheses give a measure for the relative support for the two hypotheses. *Kinship* reports likelihood ratios for a primary hypothesis vs. a null hypothesis, both specified by the user. A high value of the ratio supports the primary hypothesis, a low value does not. For each pair of individuals, *Kinship* can report the likelihood ratio, the log of the ratios, or a measure of statistical significance.

It might be assumed that a ratio of 20 or higher represents statistical significance in accepting the primary hypothesis (Brookfield & Parkin 1993). However, in fact the likelihood ratio does not directly represent significance and the ratio which provides statistical significance depends on the amount of information available. With more genetic data, lower ratios are significant (as pairs not related by the primary hypothesis consistently have lower values still). Assigning *P*-values for significance to the likelihood ratio analytically is a difficult statistical problem and *Kinship* uses a simulation routine to determine significance empirically for each case.

The simulation proceeds by first drawing an individual genotype at random, using the allele frequencies of the data set in memory to define the probability of drawing a given allele. This is the X individual. The simulation then draws the Y individual using the values of R_m and R_p that define the pair's relationship. For Y's first allele at a given locus, the simulation either copies X's first allele (with probability R_m) or else draws at random using the population allele frequencies. The second allele is chosen in the same way, using R_p. The program draws a large number of such pairs (the user may specify the number) which are related according to the null hypothesis. The value of the ratio which excludes 95% of these null-related pairs corresponds to the $P = 0.05$ significance level when it is returned for an actual pair in the data set.

The power of *Kinship's* likelihood method depends on the two relationships being compared. Some comparisons require a large number of loci to perform a powerful test, others are easily distinguished with only a few loci. Table 3 shows the number of loci required for some common comparisons. Comparing this table with the results from Brookfield (1993) shows that using $P = 0.05$ to determine significance yields higher power (i.e. fewer loci required to obtain results) than using a ratio of 20:1. Power can be increased further if one or both parents of an individual are specified. To calculate and report the power of its calculations, *Kinship* uses its simulation routine a second time, generating a series of simulated pairs related according to the primary hypothesis. The fraction of these pairs accepted at the significance level desired estimates the fraction of related pairs in the data set that will be accepted.

As an additional feature, if provided with any set of allele frequencies, *Kinship* can use its simulation routines to check the power of calculations independently of testing actual individuals. This can be useful in planning a study. *Kinship* can also perform pairwise calculations of genetic relatedness on the individuals in its data sets, using the method of Queller & Goodnight (1989).

Conclusions

Even in a study for which simple exclusions provide all the information needed, *Kinship* brings the advantages of automating the process, saving considerable time and effort and reducing the chance of errors in performing exclusions manually. When more information is required, *Kinship*'s likelihood methods offer a generality and ease of use not matched by any other available tool.

Kinship can be downloaded at: http://www.bioc.rice.edu/~kfg/GSoft.html, or contact the author at keithg@rice.edu.

References

Brookfield JFY, Parkin DT (1993) Use of single-locus DNA probes in the establishment of relatedness in wild populations. *Heredity*, **70**, 660–663.

Coltman DW, Bowen WD, Wright JM (1998) Male mating success in an aquatically mating pinniped, the harbour seal (*Phoca vitulina*), assessed by microsatellite DNA markers. *Molecular Ecology*, **7**, 627–638.

Conrad KF, Clarke MF, Roberston RJ, Boag PT (1998) Paternity and the relatedness of helpers in the cooperatively breeding bell miner. *Condor*, **100**, 343–349.

Dufour KW, Weatherhead PJ (1998) Reproductive consequences of bilateral asymmetry for individual male red-winged blackbirds. *Behavioral Ecology*, **9**, 232–242.

Edwards AWF (1972) *Likelihood*. Cambridge University Press, Cambridge.

Fitzsimmons NN (1998) Single paternity of clutches and sperm storage in the promiscuous green turtle (*Chelonia mydas*). *Molecular Ecology*, **7**, 575–584.

Kempenaers B, Verheyen GR, Dhondt AA (1997) Extrapair paternity in the blue tit (*Parus caeruleus*): female choice, male characteristics, and offspring quality. *Behavioral Ecology*, **8**, 481–492.

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

Martinez JG, Burke T, Dawson D, Soler JJ, Soler M, Muller PP (1998) Microsatellite typing reveals mating patterns in the brood parasitic great spotted cuckoo (*Clamator glandarius*). *Molecular Ecology*, **7**, 289–297.

Queller DC, Goodnight KF (1989) Estimating relatedness using genetic markers. *Evolution*, **43**, 258–275.

Queller DC, Strassmann JE, Hughes CR (1993) Microsatellites and kinship. *Trends in Ecology and Evolution*, **8**, 285–288.

Smouse PE, Meagher TR (1994) Genetic analysis of male reproductive contributions in *Chamaelium luteum* (L.) Gray (Liliaceae). *Genetics*, **136**, 313–322.

Soukup SS, Thompson CF (1998) Social mating system and reproductive success in house wrens. *Behavioral Ecology*, **9**, 43–48.

This program was written as part of K. F. Goodnight's ongoing work to develop software tools for research and education in population genetics, evolution and ecology. D. C. Queller works on relatedness and kin selection in social insects.
