Molecular Ecology Notes (2003) **3**, 479–481 doi: 10.1046/j.1471-8286.2003.00439.x

PROGRAM NOTE

FAMOZ: a software for parentage analysis using dominant, codominant and uniparentally inherited markers

S. GERBER,* P. CHABRIER† and A. KREMER*

**INRA, UMR BIOGECO, 69 route d'Arcachon, 33612 Cestas cedex, France, †INRA, Unité de biométrie et intelligence artificielle, Chemin de Borde-Rouge, Auzeville, BP 27, F-31326 Castanet-Tolosan cedex, France*

Abstract

FAMOZ (an acronym for father/mother) is a software useful in reconstructing parentage for dominant, codominant and uniparentally inherited markers. It is written in C and TclTk languages and is available for Unix, Linux and Windows systems at http://www.pierroton.inra.fr/genetics/labo/Software/Famoz/index.html. Parameters and assumptions used in the calculations are few and simple. Exclusion and identity probabilities, log-likelihoods of any genetic relationship, potential father and parent or parent pair, half- and full-sibship are calculated based on real or simulated data. Error rates for genotypic mistyping can be introduced. Simulations can be done to build statistical tests for parentage assignment.

Keywords: LOD scores, open code software, paternity, simulation

Received 6 March 2002; revision received 8 May 2002; accepted 10 March 2003

There is a growing interest in the study of gene flow. Most of the studies use codominant markers, and especially microsatellites, to infer gene flow by parentage reconstruction. However, we recently demonstrated that multiple dominant markers could be used for the same purpose (Gerber *et al*. 2000).

When several loci and many potential parents are available, parentage assignment is based on statistical tests, especially when both parents have to be identified. Meagher & Thompson (1986) proposed likelihood methods to infer parentage with codominant markers. We extended their approach more recently to dominant markers (Gerber *et al*. 2000). The existing software packages deal mostly with paternity inference (Marshall *et al*. 1998); our aim was to build a more general frame for male and female assignment including calculations based on codominant, dominant and also cytoplasmic markers. As cytoplasmic markers are uniparentally inherited, they may be helpful in inferring the male or female origin once parentage analysis has identified the parent pair. We provide an open source computer software called FAMOZ, an acronym for father/mother (http://www.pierroton.inra.fr/genetics/labo/Software/Famoz/index.html). Source codes are sent on request. The parameters and hypotheses used are few and basic. The ability of a marker system to be used in parentage analysis can be characterized with exclusion and identity probabilities, expected according to allele frequencies or observed in the sample. 'Log of the odds ratio' (LOD) scores are calculated for any parentage relationship (potential father, single parent or parent pairs). An error rate for genotypic mistyping can be introduced and the package also tolerates missing data. Simulations can be done by generating offspring from the genotyped parents or simply from allele frequencies in the population. The simulations can be used to build statistical tests for parentage assignment. The test can also be applied to simulated samples to evaluate true, apparent and cryptic gene flow.

Correspondence: S. Gerber. Fax: (33) 5 57 12 28 81; E-mail: gerber@pierroton.inra.fr

Computer languages and tools

FAMOZ is a program written in TclTk (ToolCommandLanguage/ToolKit; http://dev.scriptics.com/software/tcltk/; Welch 1999), an easy to use, noncompiled scripting language associating C programs. It provides the user with a window and menus environment to handle files and procedures involved in parentage analysis. Results are given in a simple text format. Blt (http://sourceforge.net/projects/blt/) is used for graphical purposes. Both TclTk and Blt are freely distributed and run on platforms including Unix, Linux, Macintosh and Windows.

Calculation codes (C language) were derived with extensive modifications from programs written by E.

Thompson, available at ftp://ftp.u.washington.edu/pub/user-supported/pangaea/PANGAEA/BOREL/estirel_96.Z. These programs can be compiled with the GCC compilers present on Unix and Linux platforms and freely available for Windows (http://www.delorie.com/djgpp/).

Random numbers for simulations were generated according to the method of Knuth (Press *et al*. 1992).

Hypotheses, parameters and data sets

The parentage analysis available in FAMOZ was originally developed for forest trees and is thus based on an open population for which only a subset of individuals is genotyped (the trees of the study stand). The data from 'inside the study stand' correspond to the potential parents that have been genotyped at the marker loci. The parents from 'outside the study stand' are unknown. The relationships that can be tested with the software include father/offspring, father/mother/offspring and half- or full-sibship (only for codominant markers).

The basic hypotheses used in the calculations are the following:

1 loci are independently inherited according to Mendelian laws;
2 cytoplasmic markers are considered fully uniparentally inherited (either paternal or maternal);
3 allele frequency estimates are considered to be the true allele frequencies;
4 the population is in panmixia; and
5 allelic frequencies 'outside the study stand' are assumed to be the same as 'inside the study stand'.

In addition, the user can introduce the following restrictions:

1 The proportion of typing errors both in simulation and in LOD score calculation. As mistyping is very likely to occur, it is important that it be considered.
2 The departure from Hardy–Weinberg proportions, Wright's fixation index F, in the range [0,1]. When the user provides a non-null value in simulations, homozygous genotypes at any locus are more likely to be obtained than by chance.

Exclusion probabilities

Exclusion probabilities are computed in FAMOZ for codominant (Jamieson & Taylor 1997) and dominant markers (Gerber *et al*. 2000).

Expected or observed identity probabilities represent the probability that two individuals drawn at random or observed from a population will have the same genotype at multiple loci (Waits *et al*. 2001).

'Log of the odds ratio' scores

Likelihood ratio or LOD scores are calculated for any potential parentage relationship (father/parent/parent pair) with a value greater than zero (Gerber *et al*. 2000). The higher the LOD score for a given parent/offspring pair, the higher the likelihood that the parentage relationship is true. Missing data are considered by the software. For cytoplasmic data, the probability of an individual being the true parent of a given offspring is simply 1 if they share the same allele at the cytoplasmic marker (and 0 if they do not share it), divided by the probability that this allele was transmitted by another individual of the population (frequency of this allele).

Since statistical laws associated to the likelihood ratio test are unknown, we used simulations to calculate the threshold values of the LOD score for parentage assignment.

Simulations

In FAMOZ, the simulations for paternity or parentage analysis are designed for either 'inside' or 'outside the study stand' and are built as follows:

1 randomly sampling a mother and father, generating a gamete from the mother and father or randomly generating one or two gametes from the allele frequencies of the data;
2 associating both gametes to generate the offspring; and
3 identifying the most likely father/parent/parent pair of this offspring among the genotyped parents, recording its LOD score value.

For each simulation the user chooses:

1 the number of simulated offspring;
2 the simulation error rate, i.e. the proportion of times a mistake will be made in simulations, replacing the allele from the genotyped parents by a randomly selected one; and
3 the error rate used in LOD score calculation (Gerber *et al*. 2000).

After the simulations, the two distributions of the LOD score values ('inside'/'outside the study stand') can be compared. The intersection of the distributions can be used to determine a threshold with known first and second type errors. Since this rule can be too severe in some experimental cases, the threshold can also be based solely on the distributions of LOD scores corresponding to 'inside the stand'.

Simulating individuals

FAMOZ can simulate individuals for adding fictitious observations to data sets or for simulating additional loci. If *n*

loci are available in the data set and the user wants $m \leq n$ new loci, the software randomly simulates two gametes for each new individual, drawing alleles according to the frequencies of the n existing loci, starting from the first one to the number m. The simulated genotypes are eventually provided to the user.

Testing data

Single parent and parent pair thresholds are provided by the user. If a parent has an LOD score exceeding the single parent threshold, it is considered as a true potential parent. If a parent pair has an LOD score greater than the threshold and comprises two true potential parents, it is considered as a true potential parent pair (Meagher & Thompson 1986). FAMOZ indicates the number of inferred parents according to the test.

Simulation of paternity/parentage test: decisions and gene flow

The test built in the preceding step can be applied to simulated data to measure their quality. To do so, the user chooses:

1 the number of mothers, selected either at random or in a list for paternity analysis;
2 the number of simulated offspring;
3 the total number N of individuals participating in the next generation. A [1,N] number is randomly generated. If N is smaller than G, the number of genotyped parents, the father is selected among the individuals 'inside the stand', otherwise the male gamete is generated according to allele frequencies; and
4 the thresholds of the LOD scores.

Parentage assignment is made according to the test and compared with the true situation and the number of correct assignments is recorded.

The following data about gene flow events are then computed.

1 The expected gene flow: the proportion of gametes not issued from the genotyped individuals, assuming equal success rate among genotyped and nongenotyped individuals: $(N - G)/N \times$ total number of gametes.
2 The true gene flow: the actual number of times a gamete generated according to allele frequencies (from 'outside the study stand') produced one of the offspring.
3 The apparent gene flow: the number of times no parent from 'inside the study stand' was detected for the simulated offspring according to the statistical tests.
4 The cryptic gene flow: the number of times a parent was detected among the genotyped parents according to the statistical tests whereas the true parent was 'outside the study stand'.

Half- and full-sibship

The half- and full-sibship likelihood ratio for any pair of individuals can be computed with FAMOZ codominant markers (Brenner 1997; http://dna-view.com/sibfmla.htm). Simulations can be computed to build statistical tests for measuring their ability to make correct decisions.

FAMOZ reconstructs parent/offspring relationships based on genotypic arrays obtained with different markers (dominant or codominant, biparental and uniparental inherited) or with a combination of markers. Other available softwares are usually limited to paternity analysis and, to our knowledge, there is none available using dominant and uniparentally inherited markers. The few assumptions and parameters required by the program make the results easy to obtain and to understand. Other programs often require more parameters that are not always easy to obtain for the species studied. FAMOZ can be used on any type of platform and is freely available. The template of the program can easily be used to create other specific programs.

Acknowledgements

We are very indebted and grateful to Georges Koepfler for his help in C programming. The help of Frédéric Austerlitz, Sylvie Oddou-Muratorio, Mathieu Lourmas, François Lefèvre and Clare Lord is also acknowledged. This project has been supported by a European project (OAKFLOW, QLK5-CT-2000-00960) and a national project (2000–2002) supported by BRG (Genetic Resources Board)/DERF (Rural Space and Forest Direction, French Ministry of Agriculture).

References

Brenner CH (1997) Symbolic kinship program. *Genetics*, **145**, 535–542.

Gerber S, Streiff R, Bodénès C, Mariette S, Kremer A (2000) Comparison of microsatellites and AFLP markers for parentage analysis. *Molecular Ecology*, **9**, 1037–1048.

Jamieson A, Taylor SS (1997) Comparisons of three probability formulae for parentage exclusion. *Animal Genetics*, **28**, 397–400.

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

Meagher TR, Thompson E (1986) The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. *Theoretical Population Biology*, **29**, 87–106.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Random numbers. In: *Numerical Recipes in C: the Art of Scientific Computing*, pp. 274–286. Cambridge University Press, Cambridge.

Waits LP, Luikart G, Taberlet P (2001) Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Molecular Ecology*, **10**, 249–256.

Welch B (1999) *Practical Programming in Tcl and Tk*, 3rd edn. Prentice Hall, Upper Saddle River, NJ.