Molecular Ecology Notes (2004) **4**, 779–782 doi: 10.1111/j.1471-8286.2004.00796.x

PROGRAM NOTE

KINGROUP: a program for pedigree relationship reconstruction and kin group assignments using genetic markers

DMITRY A. KONOVALOV,* CLINT MANNING* and MICHAEL T. HENSHAW†
*School of Information Technology, James Cook University, Townsville, QLD 4811, Australia, †School of Tropical Biology, James Cook University, Townsville, QLD 4811, Australia

Abstract

KINGROUP is an open source java program implementing a maximum likelihood approach to pedigree relationships reconstruction and kin group assignment. KINGROUP implements a new method (currently being performance tested) for reconstructing groups of kin that share a common relationship by estimating an overall likelihood for alternative partitions. A number of features found in KINSHIP (Goodnight & Queller 1999) have also been implemented to make them available outside the Classic Macintosh OS platform for the first time.

Keywords: kinship, likelihood, partition, pedigree, simulation, software

Received 24 March 2004; revision received 27 July 2004; accepted 9 August 2004

Increasingly, ecologists use molecular data to reconstruct pedigrees, estimate relatedness, and infer genealogical relationships. Suitable molecular markers have become more readily available, and methods for using such information to greatest effect need to be developed and improved (Blouin 2003; Jones & Ardren 2003). While simple exclusions based on Mendellian principles of inheritance can be very useful for determining some close pedigree relationships, they often fail for more distant relationships, or when close kin need to be sorted into subgroups.

The Macintosh computer program KINSHIP (Goodnight & Queller 1999) has provided one of the most powerful and flexible platforms for assessing pedigree relationships. In addition to estimating pairwise relatedness between individuals (Queller & Goodnight 1989), KINSHIP also calculates the likelihood that they share a hypothesized pedigree relationship (the primary hypothesis), and through simulation, tests whether that relationship is significantly more likely than a specified alternative relationship (the null hypothesis). Though there are often multiple possible alternatives, or null, relationships, conservative tests can usually be constructed by picking the next closest possible relationship to the hypothesized (primary) relationship as the null. This approach has proven to be both powerful and flexible, allowing for the assignment of individuals to pedigree relationships with relatively few loci (Henshaw *et al*. 2000; Ortega *et al*. 2003), and enabling tests of any conceivable pedigree relationship between both haploid and diploid individuals. However, despite its utility, the program is not available for current versions of the Macintosh operating system, or for any other operating systems such as WINDOWS or UNIX.

This program note introduces an open source java program using the likelihood formulas of Goodnight & Queller (1999), and with additional algorithms specifically designed for the reconstruction of groups of kin that share a common relationship. The Java* computer language is currently freely available for a variety of computer platforms, including WINDOWS† based PCs and MACINTOSH‡ OS X, making these calculations available outside of the Classic Macintosh OS platform for the first time. The program is available from www.kingroup.org website.

*Java is a trademark or registered trademark of Sun Microsystems, Inc. in the United States and other countries.
†Microsoft, MS EXCEL and WINDOWS are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries.
‡Macintosh is a trademark of Apple Computer, Inc., registered in the U.S. and other countries.

Correspondence: Dmitry A. Konovalov.
E-mail: Dmitry.Konovalov@jcu.edu.au

The User Interface

Because many users will already be familiar with KINSHIP, we have preserved many of KINSHIP's features in KINGROUP. It uses the same format for input files, which may or may not include the population allele frequencies. If the frequencies are not included in the input file, they are calculated by KINGROUP from the data (Queller & Goodnight 1989). Prior to loading the input file, users specify the file format in much the same way as in KINSHIP, however, KINGROUP differs from KINSHIP by displaying the dataset after it is loaded, allowing the user to easily refer to it and verify that it is correct before proceeding with further analysis.

Because the likelihood calculations are the same as in KINSHIP, the primary and null pedigree hypotheses are defined in the same way, by specifying the probabilities that the two individuals share an allele identical by descent via the maternal (R_m) and paternal (R_p) lines. R_m and R_p for many common relationships may be selected automatically in KINGROUP, or R_m and R_p may be entered manually for other relationships. This allows the user to specify any possible relationship between haploid and diploid individuals (for example, great-grandfather, haplodiploid aunt …). As in KINSHIP, users can also specify complex hypotheses which include a range of values for either R_m or R_p, e.g. full sib families vs. half sibs and unrelated, or half/full sibs vs. unrelated and parent-offspring. Once the primary and null hypotheses have been specified, the resulting likelihood calculations and their ratio can be displayed in separate windows with a variety of viewing options. These include significance flags, pairwise *P*-values, sorted matrices, and half matrices. Results can also be exported in formats suitable for other programs such as MS EXCEL.

The likelihood calculations implemented in KINGROUP are based on those of Goodnight & Queller (1999). We have extensively tested our calculations against those of KINSHIP, and they are in quantitative and qualitative agreement. Confidence levels for the pairwise likelihood values are estimated empirically through simulations for a given hypothesis [R_m, R_p] following Goodnight & Queller (1999). While we have not characterized the minimum number of simulations that are required, typically, 1000 simulated pairs are sufficient and up to 10 000 pairs are computationally feasible.

Kin Group Reconstructions

Typically, the pairwise likelihoods are used not just to infer the relationships between each pair of individuals, but to reconstruct groups of kin. Often this involves partitioning the dataset into subgroups which share a common relationship. For example, a user may want to sort the offspring of a multiply mated female into subgroups who are full siblings with each other and half siblings with the offspring in other subgroups. In these cases, an approach is required that recognizes the network of relationships that are implied by each partition.

KINGROUP implements such a method of sorting individuals into subgroups by evaluating alternative partitions of the data set according to an overall likelihood. The overall likelihood for a given partition is calculated from the pairwise likelihoods, where the primary hypothesis is that a pair belongs in the same subgroup, and the null hypothesis is that a pair should be split into different subgroups. For example to sort the offspring of a multiply mated female into full sib groups, the primary hypothesis would be that a pair of individuals were full sibs and the null hypothesis would be that they were half sibs. The user may also need to specify a complex hypothesis in some cases. For example, when sorting haplodiploid individuals into full sib groups, and in species where females mate once, individuals in different groups may be full cousins, non-relatives, or some intermediate relationship. In this case, the user could specify a complex null hypothesis with a range of R_m's from 0.375 (cousins) to 0 (unrelated), see examples in Figs 1 and 2. KINGROUP would then use the highest likelihood observed over the range of null relationships, creating a conservative test.

To create each partition, the first individual is arbitrarily assigned to the first subgroup and subsequent individuals are added, one at a time, to existing subgroups, or to newly created subgroups. The placement of each new individual is evaluated by placing it into each available subgroup in turn, as well as into a new subgroup of its own, and taking the product of the likelihoods for the pedigree relationships which result from each alternative. That is, the likelihood that the newly added individual shares the primary pedigree relationship with each member of the group it is in and the likelihood that it shares the null pedigree relationship with each individual in other subgroups. The individual is then placed in the subgroup that yields the highest overall likelihood so far.

The addition order has been implemented via the 'Descending Ratio' (Fig. 1) search algorithm that takes individuals from pairs with the clearest relationships (highest likelihood ratios), and adds them in a random order. Once the first pair is added, the next highest pair is added and so on until the entire dataset is partitioned. The 'Descending Ratio' algorithm has the benefit of grouping the most obvious individuals first which provides more information for subsequent assignments, and should make assigning less obvious individuals easier.

On the other hand the 'Exhaustive Descent' search algorithm (Fig. 2) builds every possible partition of subgroups by taking individuals in the order determined by the 'Descending Ratio' search algorithm. Then all resulting partitions are kept for the next step and not just the partition with the best overall likelihood as in the 'Descending Ratio'. The 'Exhaustive Descent' can be and is used to verify

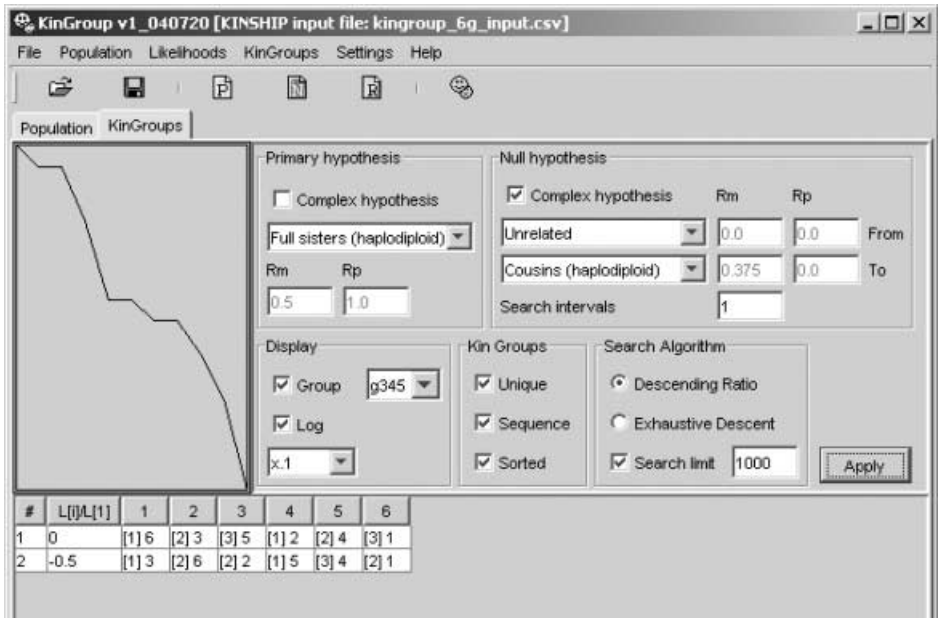

Fig. 1 Descending Ratio search (option in the 'Search Algorithm' panel) for the subgroups of haplodiploid sisters (see Primary and Null hypothesis panels) is performed on the group of six individuals. All unique partitions (the 'Unique' option in the 'Kin Groups' panel) are sorted in descending order based on the overall likelihood associated with each partition (the 'Sorted' option). The first (top most or #1) partition is used as a base for generated partitions where each individual is randomly placed into a different sub group (the 'Search Limit' option). The resulted overall likelihoods of such 'local search' are plotted as a graph for comparison. Newly assigned group id is shown as [*n*]. When results are saved, the individuals under consideration are sorted according to their new group ids in the first partition. The order in which individuals are added at each step is shown after the new group id (the 'Sequence' option in the 'Kin Groups' panel).

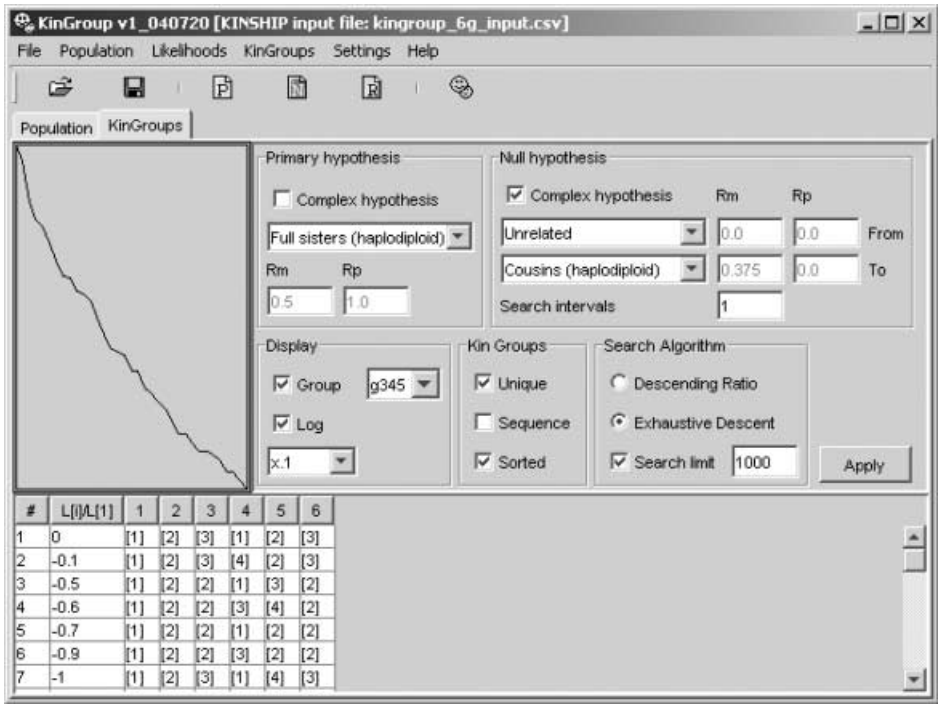

Fig. 2 Exhaustive Descent search (the same group individuals as in Figure 1) builds all possible (non-prohibited and limited by 'Search Limit') partitions resulting in 35 possible combinations (not shown, the last row in the column denoted by #). The overall likelihoods, divided by the likelihood of the first partition, are displayed in the column denoted 'L[i]/L[1]' and also plotted as a graph for comparison.

if the 'Descending Ratio' in fact arrives at the best possible partition. Such verification is important as a researcher may use it on a truncated population sample to see if the 'Descending Ratio' is viable, and then proceed with a full sample. However, such a test is limited to small data sets (order of tens) as the number of possible combinations in the 'Exhaustive Descent' explodes to an incomputable level very quickly, e.g. 35 partitions for only six individuals in Fig. 2.

The 'Descending Ratio' generally finds the correct partition provided that the partition is in fact the one with the highest overall likelihood obtained (at least in theory) by 'Exhaustive Descent'. Detailed research into the statistical properties of the 'Descending Ratio' is underway, including testing of its performance against the other partitioning methods, e.g. by Smith *et al*. (2001).

Acknowledgements

We would like to thank Bruce Litow and Ross Crozier for many useful discussions. We are grateful to Keith Goodnight and Jennifer Beyer for some clarifications regarding the simulation algorithm.

References

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

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

Henshaw MT, Strassmann JE, Queller DC (2000) The independent origin of a queen number bottleneck which promotes cooperation in the African swarm-founding wasp, *Polybioides tabidus*. *Behavioral Ecology and Sociobiology*, **48**, 478–483.

Jones AG, Ardren WR (2003) Methods of parentage analysis in natural populations. *Molecular Ecology*, **12**, 2511–2523.

Ortega J, Maldonado JE, Wilkinson GS, Arita HT, Fleischer RC (2003) Male dominance, paternity, and relatedness in the Jamaican fruit-eating bat (*Artibeus jamaicensis*). *Molecular Ecology*, **12**, 2409–2415.

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

Smith BR, Herbinger CM, Merry HR (2001) Accurate partition of individuals into full-sib families from genetic data without parental information. *Genetics*, **158**, 1329–1338.