Kinship 1.2

Manual by Keith F. Goodnight

Credits:

program by ..... **Keith F. Goodnight**
calculations by ..... **David C. Queller**
sort algorithm by ..... **Tal Poznansky**

Rice University
Dept. of Ecology & Evolutionary Biology MS-170
6100 Main Street
Houston, TX 77005-1892

Contents:

Developed with support from the Keck Center for Computational Biology, Rice University. Funded by National Library of Medicine grant 1T15LM07093 and National Science Foundation grant BIR-9419451.

programs available for downloading at http://www.bioc.rice.edu/~kfg/GSoft.html

Introduction

Kinship 1.2 is a program for testing hypotheses of pedigree relationships between pairs of individuals using data from codominant, single-locus genetic markers (such as DNA microsatellites). The program runs on Apple Macintosh computers and is released

as a "fat binary" application, i.e. one which runs native code on both Power-PC and 68K-based machines. It reads data in tab-formatted TEXT files, created by most spreadsheet programs such as Microsoft Excel using the "save as text" option. It outputs its results in the same format (in fact designating them as Excel files).

In *Kinship,* the user specifies a hypothesis about pedigree relationship using two variable, r_p and r_m. These variables define the probabilities that individuals in the pair share an allele by direct descent from their father or mother, respectively. For example, if the hypothesized relationship is diploid full siblings, both r values would be 0.5. If the hypothesis is half siblings sharing the mother but not the father, r_p is 0 and r_m is 0.5. And in haplodiploid fill siblings, r_m is still 0.5 but r_p is 1.0. (A full table of r values for common pedigree relationships is provided in the "Likelihood calculations" section.)

Given the hypothesis, *Kinship* uses the r values, the population allele frequencies, and the genotypes of the two individuals under consideration to calculate the likelihood that this genotype combination could have been produced by the relationship as specified. The calculation makes the simplifying assumptions of no linkage disequilibrium, no inbreeding, and no mutation (future versions of *Kinship* may allow the user to relax these assumptions and adjust calculations accordingly).

Kinship calculates a likelihood for two such hypotheses, the primary hypothesis and a null hypothesis, and reports the ratio between them (primary/null). A high value of the ratio favors the primary hypothesis and a low value rejects it in favor of the null hypothesis. The reason for reporting a ratio instead of a single likelihood value is that, as information increases (greater number of loci or alleles) the likelihood of a particular genotype drops. It is more useful to know whether it is more or less likely to produce the genotype one way than another. *Kinship* does include an option to output the likelihood of the primary hypothesis alone, should there be a need for it.

The program obtains a likelihood ratio for each pair of individuals in the population, or in each group, depending on the settings chosen by the user. It outputs these results in a symmetrical matrix, saved to a tab-formatted TEXT file readable my Microsoft Excel or other spreadsheet programs. There are some options to control the details of this output, which will be discussed below. The most important is a sort option which, if selected, will arrange the output matrix to bring together sets of individuals most likely to be related according to the primary hypothesis.

Kinship can also perform significance calculations to associate a p value for acceptance of the primary hypothesis to each pair's likelihood ratio. Intuitively, it might be expected that a ratio of 19 or higher would correspond to a $p < .05$ significance level (since it corresponds to the ratio of the probabilities .95/.05). However, investigation of the statistical properties of the likelihood ratio shows that the ratio needed for that (or any) significance value declines as information increases. With very good information, even a ratio less than one (which intuitively indicates the null hypothesis is the more probable) can still represent a significant rejection of the null hypothesis, as "null pairs" will have ratios very much lower still.

It is presently unclear how to analytically calculate the significance level of a given ratio. *Kinship* therefore does so empirically, by simulation. The simulation routine generates pairs of individuals using the hypothesis settings and the allele frequencies of the data set in memory and determines the ratio needed to reject the null hypothesis with $p = .05$, .01 and .001. For each ratio, it also finds the Type II error rate (rate of false

rejection of the primary hypothesis) that would result from using that level of significance.

Besides these options for likelihood calculation, *Kinship* can also output a matrix of pairwise relatedness values for all possible pairs in the population or group. Relatedness for this function is calculated in the same way as in the *Relatedness 4.2* program also available from Goodnight Software, documented in Queller & Goodnight 1989.

Reading Files

Kinship files are similar in format to TEXT files read by *Relatedness 4.2,* but with some new features. The basic organization is that (in spreadsheet format) each individual occupies one row and each variable one column. The file can contain comment lines which begin with the character * (asterisk); the program will ignore such lines completely. In addition, missing information can be simply left blank in the TEXT file or can be represented by the character • (option-8 on the Macintosh keyboard), which the program uses internally to represent missing data.

New features of the *Kinship* data file over the *Relatedness* file are the ability to specify allele frequencies in the file, and the ability to use up to 5-character names for alleles, instead of the alphabetic coding which *Relatedness* requires.

A data file consists of two main sections. Both of them are optional (although at least one of the two must be present or else the file is simply empty):

Allele frequency block: When an allele frequency block is present in the data file, *Kinship* will use its information exclusively for performing its calculations. When no allele frequency block is present, the program will calculate allele frequencies based on the individual data in the file.

Since the program does not require it, you will want to include a frequency block primarily when you have additional information on allele frequencies beyond that available in the current data set (for example, if the file is only one of several data sets which have been genotyped in the population of interest). You may also want to include a frequency block if you want to include a group ID variable but override the bias-correction to allele frequencies (see "Allele frequencies and group ID" below).

NOTE: When both and allele frequency block and an individual genotype block are present in a file, the allele frequency block must come first.

The data must be arranged with each locus occupying two columns: the first for the names of the alleles and the second for the frequency of that allele. The first locus must occupy columns 1 and 2, the second 3 and 4, and so on. The loci should read across, right to left, in the same order as they will appear in the individual genotype data.

The first non-comment line of the block must be a list of locus names, in the allele-names column for each locus. *Kinship 1.2* does not use this information, but it is included in the file specification to maintain compatibility with future versions which will do so. Starting with the second line, the information for allele frequencies begins.

The allele frequency block continues for as many lines as there are alleles in the most poly-allelic locus. Because the program has no *a priori* way of knowing how many alleles it will find, you must tell it when the allele frequency block is ending by including the word "end" on a line by itself at the end of the block.

The number of loci, and the number of alleles at each locus, is limited only by the available memory of the computer (see "Memory considerations" below).

Individual genotypes block: If this section of a data file is absent (in a file which does contain an allele frequency block) then *Kinship* can only perform simulations. No other function will be available (see the "Simulations" section below for information about the simulation function).

NOTE: When both an allele frequency block and an individual genotype block are present in the file, the individual genotype block must come second.

The first non-comment line of the individual genotype block must give the variable names of each column. *Kinship 1.2* does not use this information, but it is part of the file specification to maintain compatibility with future versions which will, and with *Relatedness 4.2* files which do use the information.

Kinship 1.2 reads the following variables: group ID, individual ID, Mother's ID, Father's ID, and genotype. Additional variables may be present in the file, although *Kinship* will not read them. These variables can occupy any column in any order, with the only restrictions being: 1) That loci occupy a set of consecutive columns, and that they read left to right in the same order as they appear in the allele frequency block (if one is present). 2) Parental IDs can only be used in a data set if individual IDs are also present (because it is to the individual ID that the parental ID variables refer).

Missing information for any variable can be indicated with the • character (option-8) on the Macintosh keyboard, or may simply be left blank. The • character *cannot* be used as an actual value for any variable: the program will always treat it as indicating missing information. All other characters both alphabetic and numeric, can be used. For alphabetic characters, case is significant (i.e. upper and lower case are treated as different).

Each locus should occupy one column. For diploid genotypes, both allele names are placed in the same column, separated by a delimiter character (which the user can specify).

> *EXAMPLE:*
> If the delimiter character is "/" and two alleles are named "100" and "110", then valid diploid genotype entries could be:
> `100/110  100/100  110/100  etc.`
> A haploid genotype could be written in any of the following ways:
> `100/•  100/  100  •/100  etc.`
> Likewise, a missing genotype could be any of the following:
> `•/•  /  <blank>`

Kinship can read TEXT files prepared for *Relatedness 4.2*, which uses single-character, lowercase letters for alleles. Specify no character as the delimiter, and *Kinship* will assume it is dealing with such a file.

The end of the individual genotype block is marked by the end of the file; no special designator is required for it.

Figure 1 shows a portion of a *Kinship* data file as displayed in Microsoft Excel. The *Kinship* distribution package also contains a sample data file which you can examine to see the file format.

	A	B	C	D
1	*Sample Data set			
2	*Lines beginning with "*" are comment lines			
3	*First, optional allele frequencies			
4	loc1		loc2	
5	100	0.5	155	0.5
6	110	0.25	163	0.5
7	117	0.25		
8				
9	end			
10	group	ID	loc1	loc2
11	1	1	100/110	•/•
12	1	2	110/117	155/163
13	1	3	110/117	155/155
14	2	1	100/100	163/163
15	2	2	110/110	155/163

Fig. 1: Part of a *Kinship 1.2* data set, showing both allele frequency and individual genotype blocks

Telling Kinship what to load:

Two commands in *Kinship's* File menu are used to specify the details of a data file to be loaded.

The Configure text... command brings up a dialog box similar to the file configuration box used by *Relatedness 4.2* and shown in figure 2. This dialog box allows you to specify which variables are present in your file and what columns they occupy, whether or not an allele frequencies block is present, how many loci are present, and what the delimiter character is in genotypes.

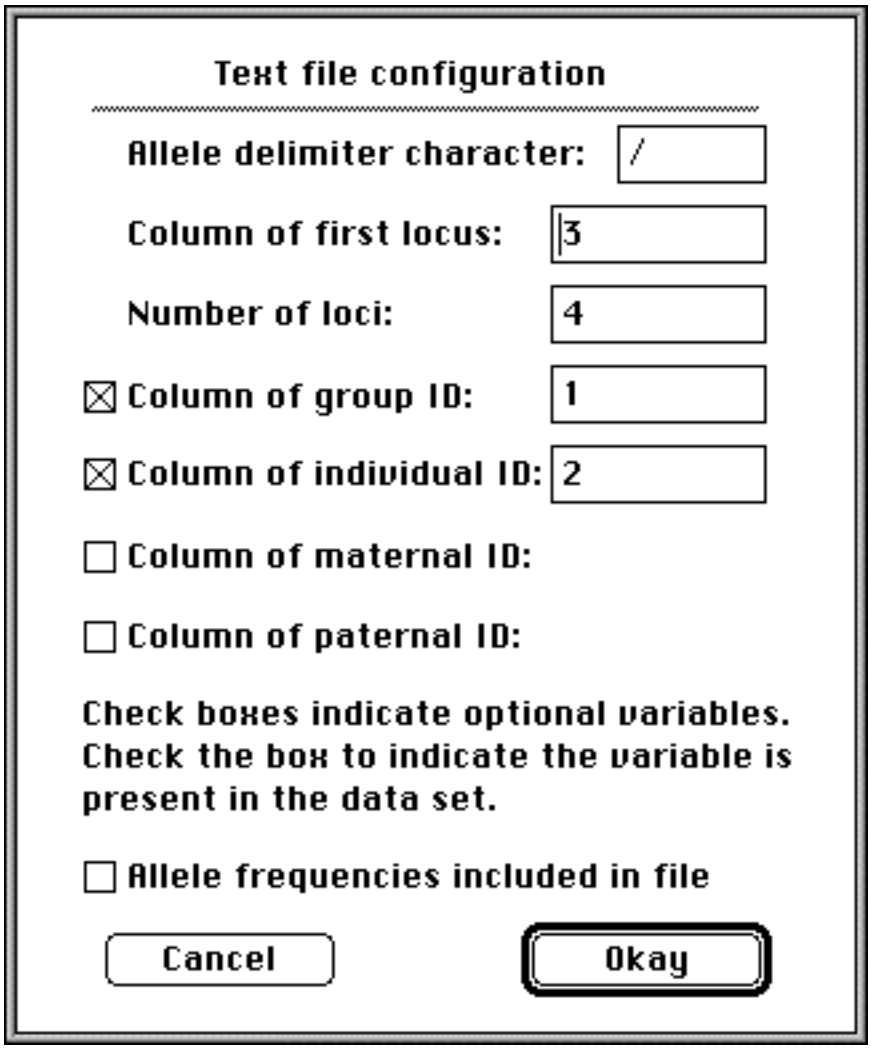

Figure 2: The Configure text dialog box.

From top to bottom, the items in this dialog are:

Allele delimiter character: The character in this box is the one the program will use to separate the alleles in individual genotypes. The default is "/". If you leave the box blank, *Kinship* will assume you have a file designed for *Relatedness 4.2* and will expect to find genotypes specified by 2 single-character, lowercase allele names.

Column of first locus: This is the first in the set of consecutive columns that give the genotypes at each locus. Remember that all loci must occupy a consecutive set of columns and must read from left-to-right in the same order as they appear in the allele frequencies block, if present, although the set may begin with any column.

Number of loci: This box indicates how many columns of locus information *Kinship* will read. It need not be the same as the full number present in the file: you can instruct the program, for example, to read only the first 3 out of 6 loci listed in the file.

Optional variables:

For each of these variables, the checkbox on the left specifies whether the variable is present in the data file or not. If you check that it is present, a box on the right becomes available for you to specify which column contains it.

Group ID: The primary function of the group ID, if present, is to specify sets of putative relatives for use in applying a bias correction to the allele frequency calculations (see below). When present, this variable also allows you to split the data file into groups and optionally to perform the likelihood comparisons only among pairs in the same group. If there is no group variable, the program will always perform comparisons among all pairs in the file. (This remains an option even when Group ID is present.)

Individual ID: When present, this variable allows *Kinship* to label the rows and columns of the output matrix with the ID of the individual each row and column represents. Without this variable, the program can still output the matrix but it will be unlabeled. While this might be acceptable if the only question is (for example) whether all individuals in a group are full siblings of one another, normally you will want to use this variable. Individual ID is also necessary to allow identification of individuals' parents in the data set.

Maternal ID, Paternal ID: Available only if Individual ID is checked, these variables give the ID label of each individual's mother and father (if known). This information is used to refine the likelihood calculations, improving the results. Using this variable does not require that every individual have its parents identified. *Kinship* will use the information when it is present and proceed without it if it is missing.

> *NOTE: It is important that, if present, parental IDs be correct. If examination of the genotypes shows that the indicated mother or father cannot actually be the parent of an individual, then the program will report "Excluded" instead of a ratio value for all that individual's comparisons (see "Output files" below for more information).*

Allele frequencies included in file: Check this box if your data file contains an allele frequencies block; uncheck it if it does not.

Once you have specified the format of your file using the Configure dialog, click "Okay" to accept your choices or "Cancel" to dismiss the dialog with no choices made.

The Open... command in the File menu is used to open the actual data file, once your configuration settings have been made. This command brings up a standard Macintosh Open File dialog, from which you can choose your data set.

Memory considerations:

Kinship limits data set size, including number of individuals, loci, alleles and groups, only by the available computer memory. The best way to save space if memory becomes a problem in running *Kinship* is to divide the data set into groups and perform comparisons by group only, not for the entire population. The reason is that the matrix of output results, a 2-dimensional matrix indexed by the number of individuals compared, varies in size as the square of that number. *Kinship* also uses high-precision variables for the results stored in the matrix, occupying considerable memory (this is necessary because with large numbers of loci, values for the likelihoods may be extremely small, as low as 10^{-2400} for data sets with 20 loci which we have tested).

With a large data set, the output matrix increasing as the square of the data set size can easily outstrip the memory required to hold the data itself. To perform calculations on all pairs in the population requires memory to hold a matrix corresponding to the full size of the data set, while to perform calculations by group requires a maximum a matrix corresponding to the size of the largest group. As an example, a data set of 100

individuals divided into 10 equal groups will require a 100x100 matrix to compare all pairs in the population and only a 10x10 matrix (re-used 10 times) to calculate by group— the by-group memory requirement is only 1/100th that of the full-population.

When it encounters a memory limitation, *Kinship* will attempt to determine what calculations it can and cannot perform, and notify the user. You will have the option to proceed with the stated limitation, or else to cancel file loading. You can then quit the program and increase its memory partition in the Macintosh Finder (see your Macintosh's documentation for details).

If the data set itself is too large to be held in memory, the only option will be to cancel file loading. More commonly, because of the scaling of the output matrix, *Kinship* will be able to calculate by group but not for the whole population (if a group variable is present— otherwise there will again be no option but to cancel loading). Another possibility is that *Kinship* will be able to perform all calculations but will not be able to run the results sorting routine, which requires some additional overhead.

If you are unable to increase *Kinship's* memory partition large enough to handle a large data set, the best way to conserve memory will be to find some way to subdivide the data set into groups (or further subdivide it if a group variable is already present) so that even if full-population calculation remains impossible comparisons by group will be available. Issues of bias correction in allele frequencies must be considered before subdividing a population in this way (see below), but if it can be done it will reduce memory requirements considerably.

If the data set cannot be divided, memory may still be saved (though at less than an N^2 rate of savings) by eliminating uninformative loci or alleles from the data set. If you are including an allele frequencies block in you file, and you know that some alleles are present in the overall population but not in the current data file, omit them from the alleles list in the data file. As long as the remaining frequencies are correct, *Kinship* does not need to allocate space to store information on alleles that will never be used. (The program will warn you that allele frequencies don't add up to 1.0, but you can dismiss the warning and proceed.) Also, dropping any single individuals who are uninformative (missing data for most loci, for example) will save memory at the same N^2 rate as subdividing the data.

In general, we have not found memory to be problematical until we reach data sets of around 1000 individuals or larger. For such a data set, a 1000x1000 matrix would be difficult to read and interpret in any case, and subdivision into by-group calculations would be desirable even without memory considerations.

Allele frequencies and group ID

As discussed in Queller & Goodnight (1989), when performing relatedness calculations using population allele frequencies obtained from the same data set as the individuals being measured, a bias correction must be applied to those frequencies. The same consideration applies to the likelihood calculations performed by *Kinship.*

Basically, any individuals which by hypothesis are relatives of the individual(s) under consideration will be expected to have allele frequencies closer to those individuals than the true population mean. Their inclusion in the limited sample of a data set thus biases its measure of population frequencies in that direction.

The solution is to exclude from background frequency calculations all individuals who might be relatives of the current individual.

Like our earlier program *Relatedness, Kinship* uses a "Group ID" variable to identify sets of such individuals. As it performs calculations on a given pair of individuals, *Kinship* will use allele frequencies obtained from the data set *excluding* the group(s) to which the current pair belong.

Unlike *Relatedness 4.2, Kinship's* group ID variable is optional. If no group ID is loaded (or if it is loaded but has only one value, an equivalent situation), then *Kinship* can only bias-correct population frequencies by excluding the 2 individuals in the current pair. Also, if an allele frequencies block is included in the data file then *Kinship* will assume that the frequencies provided are correct and will not attempt any corrections— it will use exactly the frequencies read from the file for all calculations.

Because of this function of the group ID, it is important for the accuracy of results to use this variable correctly. If your population can be clearly divided into sets of related individuals, you should designate this division with a group variable (even if you intend to do comparisons across the whole population). Likewise, if your population is more mixed, with no clear sets of relatives, you should not use a group ID.

If you want to use a variable to subdivide your comparisons but it is not appropriate to use it for this bias correction, then you can include an allele frequencies block in your data set along with loading the variable as a group ID. Since having a frequencies block overrides the bias correction, the group ID will only function to subdivide the pairwise comparisons.

Likelihood Calculations

Once a file is in memory, the commands under the Calculate menu become available. *Kinship's* primary function is accessed through the Likelihoods... command.

This command brings up the dialog box shown in figure 3. This large dialog presents all the options available for the pairwise likelihoods calculation. Settings controlling the hypotheses and calculations are on the left of the dialog, and output options on the right.

Taking the controls from top to bottom, hypothesis settings first, they are:

Treatment of haploids: Assigning which of an individual's two alleles is maternal or paternal in origin is necessary in performing the likelihood calculations (it controls whether the r_p or r_m value applies in comparison between the two individuals). For diploid individuals, *Kinship* uses the maternal and paternal genotypes, if parental IDs are in the data set, to decide which allele is which. If the parental genotypes leave the choice ambiguous, or if parental IDs are not a part of the data set, it performs calculations for all possible permutations and combines the result.

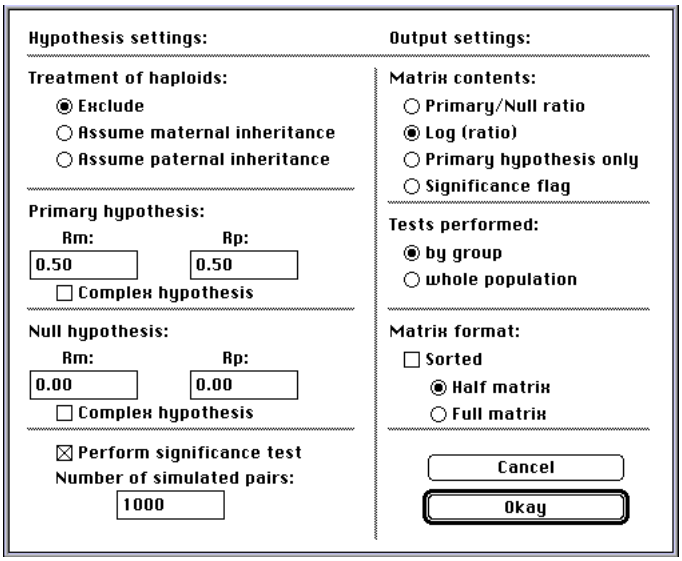

Figure 3: The Likelihoods dialog box (shown at half size)

For haploid individuals, however, you will generally know which parent the single allele comes from (if your data set includes haplodiploid males, the haploid genotype comes from the mother; if a sperm genotype, it comes from the "father", etc.). So, instead of checking permutations, *Kinship* lets you specify in advance whether haploid genotypes have maternal or paternal inheritance. The third option in this set of radio buttons is to exclude haploids altogether.

Hypothesis settings: The controls for specifying the primary and null hypothesis are the same. The primary hypothesis will be in the numerator when the ratio is taken, the null hypothesis in the denominator. Other than that, both are defined and calculated in the same way. The hypothesis settings consist of boxes for the r_p and r_m values, and a checkbox to specify if the hypothesis is "complex" (see below).

The two r values specify the probability that individuals related in the hypothesized way share an allele by descent through maternal (r_m) or paternal (r_p) inheritance. The values specify the probability over and above the baseline probability of identity-by-state due to the population frequency of the alleles. Different pedigree relationships may be described by different settings of the r values. Note that *Kinship* cannot (in its present version) distinguish between types of relationship that would have the same r values (such as mother-offspring vs. maternal half-siblings).

Table 1 gives a list of common pedigree relationships and the associated r values:

Table 1: r values for different relationships.

Relationship	r_m	r_p
Diploid full sibling	0.5	0.5
Diploid half sibling (mat)	0.5	0.0
Diploid half sibling (pat)	0.0	0.5
Haplodiploid full sister	0.5	1.0
Haplodiploid full brother	0.5	0.0
Haplodiploid sister-brother	0.5	0.0
Mother-offspring	1.0	0.0
Father-offspring	0	1.0

The tests performed by *Kinship* are always symmetrical, i.e. the ratio returned for individuals X to Y will be the same as for Y to X.

Complex hypotheses: A complex hypothesis is one in which the user specifies a range of r values instead of a single set. When a complex hypothesis is entered, *Kinship* will find the likelihoods for all r values within the range (the user can set how finely it divides the search space) and use the highest one that it finds in calculating the final ratio.

A complex hypothesis represents a hypothesis that a given pair is related in any one of several possibilities. Its most common use will be in the null hypothesis, for example to test whether pairs are full siblings as opposed to half-siblings, cousins, or unrelated. (A simple null hypothesis could only test for full-siblings vs. a single one of the possibilities at a time.)

Caution should be used in choosing the complex hypothesis option. *Kinship* does not confine its tests to particular significant values of r_p or r_m but checks the full range of possibilities. If your range is too broad, then the meaning of the results is less clear. In addition, *Kinship* is unable to perform significance tests when either hypothesis is complex.

If you do choose a complex hypothesis, then after you dismiss the main dialog box *Kinship* will present a second box for you to choose the second set of r values, defining the range to be searched (figure 4).

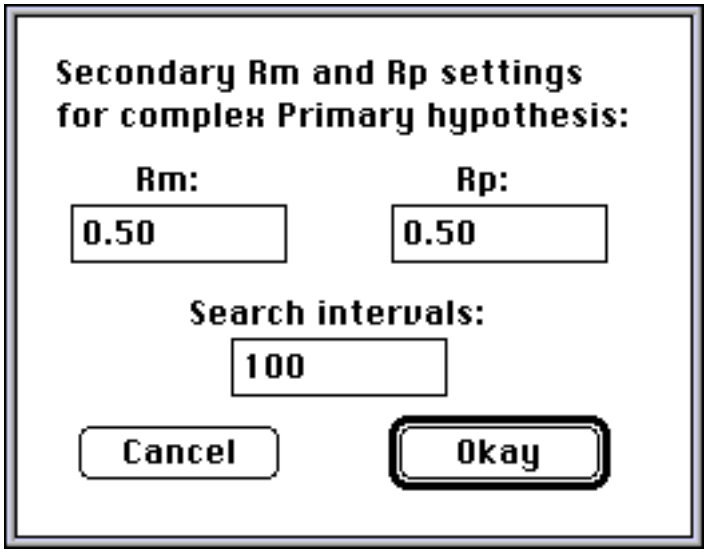

Figure 4: The Complex Hypothesis dialog

The choices in this dialog allow you to set the secondary r values, and the number of intervals to divide the range when *Kinship* searches for the highest likelihood value. The default r values in this dialog will be the ones chosen in the main dialog. If you only want one of the two to vary, leave the other at its default value.

Returning to the settings of the main likelihood dialog, the next is:

Perform significance test: When this checkbox is clicked, *Kinship* will run a simulation series to find what values of the likelihood ratio correspond to certain significance levels.

As noted in the introduction, the likelihood ratio does not match the significance level in an intuitive way, nor is it easy to calculate a significance level analytically. *Kinship* determines significance empirically, by generating a series of pairs at random (using the allele frequencies and *r* settings) and determining what values of the likelihood ratio result.

The box below the checkbox allows you to choose how many simulated pairs *Kinship* will generate in order to find the significance levels. The program will being by generating a series of pairs which match the null hypothesis. High values of the likelihood ratio for these pairs will constitute false positives or Type I errors. *Kinship* will find the ratios needed to exclude such errors at significance levels of $p = .05$, .01 and .001 (i.e. the ratio that represents the $p=.05$ significance level will be that which excludes 95% of the simulated pairs, etc.).

Once it has found these ratios, *Kinship* will generate a second series of pairs which match the primary hypothesis. Low values of these ratios represent false negatives or Type II errors. For each significance level, *Kinship* will find the rate of Type II errors which results.

Both the significant ratios and their associated Type II error rates will be reported in the header to the output file.

The default value for the number of pairs to simulate is 1000. This value represents a compromise between speed of calculation and precision of results. If your data allows you to obtain results significant at the $p = .001$ level, you will probably want to increase this number to improve the precision of the appropriate ratio.

The ratio reported for each level will be that which excludes exactly the appropriate number of pairs; normally, you will want to accept ratio values higher than that (not equal to it) and report your significance values as $p <$ __ .

The automatic significance testing that these controls provide is limited. This routine cannot calculate significance levels if either hypothesis is complex, because the range of *r* values requires additional assumptions about the distribution of values in the simulated pairs. Without precise information about the actual composition of the data set, the significance value will not be accurate. Therefore *Kinship* will not attempt such calculations.

Another limitation is that the simulated pairs are always both diploid (again, because the other alternative would require additional assumptions about the number of haploids vs. diploids in the data set). If you want to perform significance tests on haploid individuals, there is a manual simulation command (discussed below).

Down the right side of the likelihoods dialog are the settings for the output of *Kinship's* results.

Matrix Contents: This set of radio buttons allows you to choose what will actually be displayed in the output file. The default choice is the log (base 10) of the likelihood ratio. Choosing the log of the ratio has the advantage of being symmetrical about 0 for favoring the primary or the null hypothesis. You can also choose to report the actual (linear) value of the ratio.

A third option reports the likelihood value for the primary hypothesis alone, without taking a ratio.

The final option, available only if the "perform significance test" check box has been clicked, is to report a significance flag instead of the actual ratio. If this option is checked, *Kinship* will output "NS" for not significant, "*" for $p < .05$, "**" for $p < .01$ and "***" for $p < .001$. An output file with these flags replacing the ratio values can be used as an easy-to-read "index" complementing a file of the actual numerical values.

Tests performed: These two buttons give you the option of comparing all individuals in the data file vs. comparing only pairs within groups. As previously discussed, in cases of memory limitation sometimes only the "by group" option will be available.

Matrix format: The first of these choices gives you the option of sorting the output matrix or not. If selected, the sort routine attempts to group together sets of individuals which meet the primary hypothesis. For example, if the primary hypothesis is that of being full siblings, and the data file or group contains several sets of full siblings, the sort routine will tend to draw the sets of siblings together into a block. The sort will not infallibly group all the right individuals together, and should only be taken as an initial aid to identifying such sets. If the sort option is not checked, individuals will appear in the matrix in the order they occurred in the original data file.

The final two radio buttons allow you to output either a half matrix or whole matrix. Because *Kinship* calculations are symmetrical, the matrix entries for (Ind. X, Ind. Y) and (Ind. Y, Ind. X) will be the same. So it is only necessary to output one half of the symmetric matrix to display all the results. However, it is occasionally useful to have the full symmetric matrix printed out as well. *Kinship* offers both options.

If you choose "half matrix", the output will be a lower-triangular matrix while "full matrix" will output the complete symmetrical table.

Once you have chosen all your settings, click "Okay" to proceed with calculations. *Kinship* will present the complex hypothesis dialog box if necessary, and will follow it with a standard Macintosh Save dialog for you to name and locate the output file. Then calculations will begin.

The "Cancel" button will dismiss the dialog with no action taken.

Relatedness Calculations

Choosing Relatedness... from the Calculate menu brings up the dialog box shown in figure 5, which allows you to perform pairwise relatedness calculations on the individuals in your data set. The relatedness statistic obtained is the same as that discussed in Queller & Goodnight (1989) and used in our program *Relatedness 4.2*.

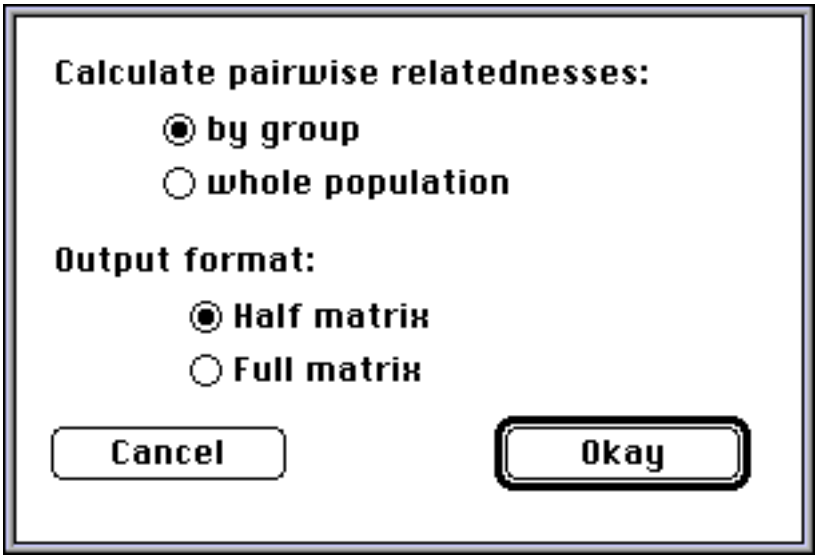

Figure 5: Relatedness dialog

The settings for relatedness calculations are much simpler than those for likelihoods. The only options are to perform the calculations by group or for the whole population, and to output either the lower-triangular half matrix or the full symmetric matrix. (See the discussion of these options in the *Likelihood calculations* section above for details.)

When you have chosen your settings, click "Okay" to proceed. *Kinship* will present a standard Macintosh save dialog for you to name and locate the output file, and then proceed with calculations.

Simulations

The automatic simulation used for significance testing in likelihoods calculation creates only diploid pairs, and cannot find significance for complex hypotheses. If you want to perform more complex simulations, you can use the Simulate command from the Calculate menu. Unlike other calculation commands, simulation is available even with a data set consisting only of allele frequencies, with no individual genotype data. The Simulate command brings up the simulation dialog box shown in figure 6.

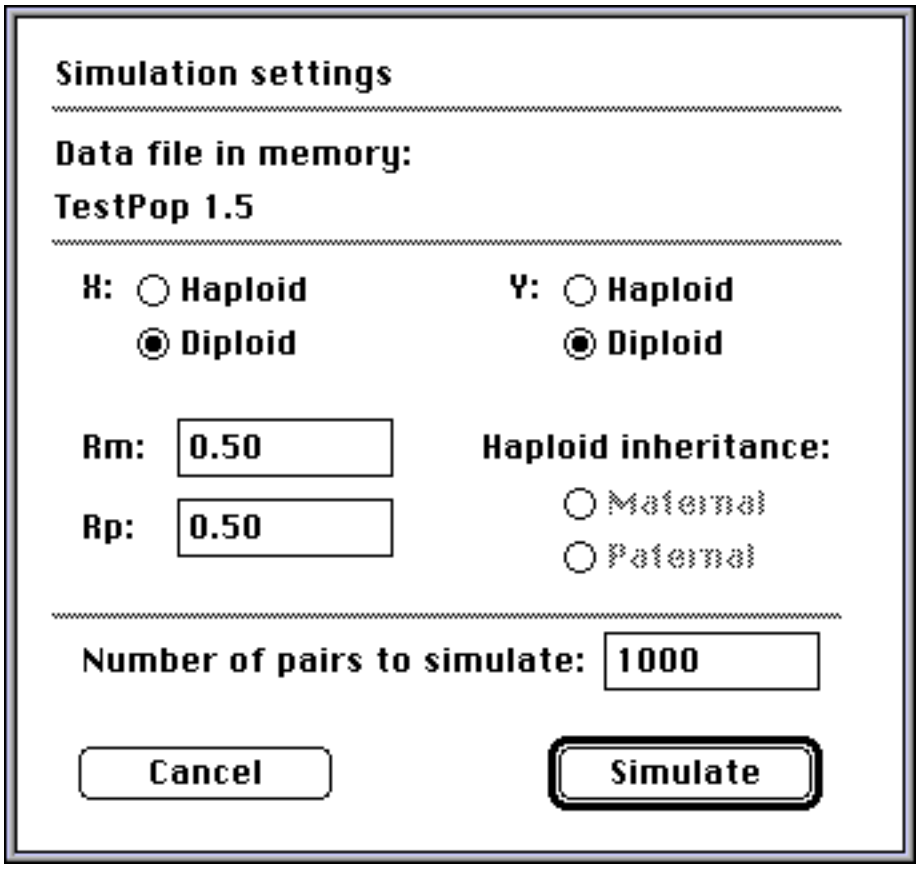

Figure 6: The Simulate Dialog

The simulation routine will create a series of simulated pairs of individuals, using the allele frequencies of the file in memory and the r_m and r_p values you enter in the boxes shown in the simulate dialog. The pairs will be created so that they match the r values you choose; i.e. they will "really" be related according to a hypothesis with the same r values as those you choose in this simulate dialog.

You can also choose to have either or both of the individuals in each simulated pair be haploid. If one individual is to be haploid and the other diploid, it will make no difference which one (X or Y as labeled in the dialog) you designate as which. If either or both individuals in a pair are haploid, the radio buttons labeled "Haploid Inheritance" will become available: you can specify whether the single allele in a haploid individual comes from the mother or father (and so, whether r_m or r_p is the value that applies to it).

The last setting in the simulate dialog lets you specify how many simulated pairs *Kinship* will create.

If you click "Okay" to proceed with the simulation, *Kinship* will next present the Likelihoods dialog (figure 3) where you can specify the test to be run on the simulated pairs. Use the Likelihoods dialog just as you would if you were specifying calculations on actual data, except that only some of the control will be applicable. Only the "Treatment of Haploids" radio buttons and the r settings for the primary and null hypotheses will apply to a simulation. Under "Matrix Contents", you can choose to display the log of the ratio or the ratio itself, but the other two choices will be ignored (*Kinship* will treat both as a choice for reporting the linear value of the ratio).

After you dismiss the Likelihood dialog, *Kinship* will present a standard Macintosh save dialog where you can specify the name and location of the simulation output file.

Kinship will then proceed with calculations. For each simulated pair, it will write into the output file the likelihood values for the primary and null hypotheses, the likelihood ratio, and the pairwise relatedness. You can load this output file into a statistical program or other application to investigate the distributions of these variables or any other property of interest. Because they were drawn from the same allele frequencies as the data set currently in memory, the distributions should estimate the results for actual pairs related in the same way as the simulated pairs (barring factors such as inbreeding or linkage disequilibrium).

The most common use for manual simulation is to provide significance tests in situations too complicated for the automatic routine, such as tests where one or both of the pair is haploid, or where one of the hypotheses is complex.

To perform a significance test, generate a set of simulated pairs which match your null hypothesis. Test them using the same primary/null combination you plan to use on your actual data. High values of the ratio will constitute false positives. With your statistical or spreadsheet program, examine the simulation output file and find the value of the ratio which is met or exceeded only by the fraction of the total pairs which matches the desired significance level.

To calculate Type II error rates, generate simulated pairs which match the primary hypothesis. Again, test them as you plan to test your actual data. Low ratio values constitute a false negative. You can find the number rejected at any given ratio level (such as one you obtained from a significance test).

Simulations and complex hypotheses: A complex hypothesis represents a mixture of different *r* values, indicating a mixture of different pedigree relationships in the data set. The simulation routine only generates one kind of relationship at a time. However, you can do a series of simulations, with different *r* values, and combine the output files using your spreadsheet or statistical program, putting the different relationships into the final file in the right proportions.

You should be very cautious in attempting to calculate significance for complex hypotheses, however. Your results will only be accurate if you have the correct proportions of the different relationships present. You would have to know what percentage of possible pairs in your data set are first cousins, half-siblings, full-siblings, etc., and correctly represent these proportions in your combined simulation file. If you have this information, and need *Kinship* only to identify to which class particular pairs belong, then a complex hypothesis with a significance test may be feasible.

Output Files

Kinship's output files are standard Macintosh TEXT files, tab-formatted to be read by most spreadsheet programs. *Kinship* identifies its output files as Microsoft Excel files, but any spreadsheet application or word processor equipped to read plain-text files can read them. If you open a *Kinship* output file with a word processor, you will need to set the tab stops to properly separate the columns of data; consult your word processor's documentation for details.

Kinship has three types of output file: a likelihoods file, a relatedness file, and a simulation file. The first two are similar in format.

The output file begins with a header giving information about the data file and the exact calculation settings used to produce the output. If a significance test was performed

for likelihood calculations, a table of the ratios required for significance at the .05, .01 and .001 levels will be included in the header.

The matrix or matrices of pairwise values comes after the header. If you performed calculations on the whole population, there will be a single large matrix. If you performed calculations by group, a series of smaller matrices will be shown. A part of a *Kinship* output file as displayed in Microsoft Excel is shown in figure 7.

	A	B	C	D	E	F	G
11	Group: 1						
12		h1	u5	**f1**	**f3**	**f2**	u3
13	h1	*					
14	u5	0.957	*				
15	**f1**	-1.014	1.885	*			
16	**f3**	-1.477	0.342	**7.518**	*		
17	**f2**	-2.431	0.242	**3.046**	**4.536**	*	
18	u3	-1.083	-3.871	-0.265	6.434	-0.231	*

Figure 7: Part of a *Kinship* output file.

The settings used for this file were half-matrix, and sorted. The bold-faced entries show a set of 3 full siblings which were pulled together by the sort routine, after being initially scattered through the group (the hypothesis tested was full-sibling/unrelated). (The bold-face highlighting was not in the original output file but was added to emphasize the grouping.)

This sample also displays the log of the likelihood ratio, rather than the actual ratio. When the actual ratio is displayed, the result appears in exponential notation.

If "Significance Flag" had been selected as the output choice, then in place of numerical values the matrix would contain the markers "NS" for not significant, "*" for .05 significance, "**" for .01, and "***" for .001. The asterisks along the diagonal would be replaced with "-" to avoid confusion. To display a given significance level, *Kinship* requires that the ratio be *higher* than the threshold ratio shown in the header to the results file, not simply equal to it.

Besides numerical or flag values, *Kinship* will also sometimes present one of the following entries as a matrix value:

MISSING: This label indicates that genetic information was missing from one or both individuals to the extent that no calculations could be performed.

UNDEFINED: The null hypothesis was excluded (i.e. its likelihood value was 0) so the ratio is undefined.

EXCL: The primary hypothesis was excluded (i.e. its likelihood value was 0) so the overall ratio value is 0. If you are reporting the log of the ratio, a value of 0 has an undefined log, so this text label is reported instead.

ZERO: Both the primary and null hypotheses were excluded, so the value of the ratio was 0/0. The most common reason for this value to appear is when parental IDs are in use and the stated parents for one or both individuals are incorrectly identified. Since the parental ID is considered a part of the test hypothesis, both primary and null are excluded.

Some of these text values may appear in the table of significant ratios as well. The simulation will never have MISSING ratios, but UNDEFINED, EXCL and ZERO can

occur. For purposes of comparison in applying significance flags to the ourput, the program treats an UNDEFINED result (null hypothesis excluded) as higher than any numerical ratio, EXCL as lower than any ratio, and ZERO as a ratio of 1.0 (since, though both excluded, both hypotheses did have the same "likelihood").

The format of a simulation output file is different. There is a header containing file information and the settings used for the simulation, and then 4 columns of values, representing the primary likelihood, the null likelihood, the ratio, and the relatedness value. Each line represents a single simulated pair.

The likelihood values will be shown in exponential notation, and the ratio will be in exponential notation if reported directly, and standard notation if you chose to report the log.

Reference

Queller, D.C. and Goodnight, K.F. 1989. Estimating relatedness using genetic markers. *Evolution* 43(2): 258-275.

KINSHIP version history:

1.2: Fixed a bug in the significance simulation which caused "missing" values to appear in the simulation output when allele frequencies included in the data set did not add up to 1.0.

1.1.2: Miscellaneous bug fixes: the significance-testing simulations were reporting 1.0 as the threshold for all significance levels under some circumstances, and the "Assume Maternal" option for haploid treatment was incorrectly being disabled when a complex hypothesis was in use. Also clarified the output of the manual simulation, which was listing r_p and r_m values in different order in different parts of the output, potentially causing confusion.

1.1.1: Fixed a bug that incorrectly reported out-of-memory errors for some large data sets; changed the file window to no longer report "There are 1 groups" when the group variable is not in use.

1.1: Corrected a bug in the allele frequency calculations. This bug caused incorrect results to be reported for data sets which did not include an allele frequencies block and which lacked a group ID variable (or equivalently had group ID but only one value, i.e. one single group).

The version 1.1 release package also included some revisions to the manual in response to questions from some users about *Kinship's* calculation of allele frequencies and bias correction.

1.0: Initial release