The kinship2 R Package for Pedigree Data: Supplementary Material

Jason P. Sinnwell, Terry M. Therneau and Daniel J. Schaid

Division of Biomedical Statistics and Informatics, Department of Health Sciences Research,

Mayo Clinic, Rochester, MN 55905

This supplementary document provides additional details and examples for the methods described in the application note entitled ``kinship2: An R package for pedigree data". The examples below make use of Figure 1 of the main document, a 17-subject pedigree containing four generations, identical twins, and a consanguineous marriage.

PEDIGREE PLOTS

As a brief example, let us assume we have a data frame in R named *fam1* with columns: subjid as a unique id; dad and mom for subject id of each subject's parents (0 if a founder, i.e., no parents in the pedigree); sexcode as 1 or 2 for male or female; and disease and vitalstatus indicator variables for disease and vital status. Below shows an example of how to construct the pedigree object and then plot it:

```
R> ped1 <- with(fam1, pedigree(id=subjid, dadid=dad,
            momid=mom,
            sex=sexcode, affected=disease,
            status=vitalstatus))
R> plot(ped1)
```

The status argument expects the vital status indicator and the affected argument expects an indicator of disease status. The affected argument can also be a matrix, which provides a way to store multiple indicator variables for each subject that can be used in plot symbols or simply used later in analyses.

Figure S1 shows a detailed pedigree plot created from subjects 1-10 of the full pedigree shown in Figure 1 of the main document. The plot shapes are split into as many portions as there are indicator variables in the affected matrix of the pedigree object. This example has three indicators, only one of which indicates disease status. The pedigree.legend function created the legend in the top right; indicating which portion of the shape shading indicates a positive status for the three variables in the affected matrix: disease, smoker, and availstatus. Additionally, the plot method allows the subject id to be alternate text. Supplementary figure S1 has the subject identifier concatenated with a character string created from the indicator available in the object's affected matrix, separated by a line-break character. Finally, though not shown here, the plot allows subject-specific colors, which applies to the plot symbol and identifier.

PEDIGREE TRIMMING

The pedigree.shrink function iteratively removes subjects given their availability status (e.g., genetic data available) and an affected status, from a pedigree in the following order,

until the pedigree is of a desired bit size.

1. Uninformative subjects, i.e., unavailable (no genetic data available) with no available descendants
2. Available terminal subjects with unknown affected status if both parents are available
3. Unkown affected status
4. Unaffected
5. Affected

The first two steps are always run to shrink the pedigree, regardless of the desired bit size. If after these steps the bit size is not met, the algorithm iteratively shrinks the pedigree by preferentially removing subjects, in the order of steps 3 to 5. If there are multiple subjects that meet the condition under consideration, one of them is randomly chosen to be removed. The method’s default bit size of 16 was chosen because some linkage software has memory requirements that increase exponentially by pedigree size, but a different bit size may be chosen for determining informative subjects in genome-wide association and sequencing studies where cost may be more of a limiting factor.

Given the affected information shown in Figure 1, and an availability status that is TRUE for all affected individuals, the pedigree in Figure 1 shrinks to the pedigree in supplementary Figure S1 after just the first two steps of pedigree.shrink, with a bit size of (2*6)-4=8. The command to shrink the pedigree in Figure 1 to the pedigree in Figure S1 is shown below, followed by a call to the pedigree.shrink print method.

```
       R> shrink1 <- pedigree.shrink(ped1, avail=availstatus)

       R> print(shrink1)

Pedigree Size:
                 N.subj Bits
Original             17   19
Only Informative     10    8
Trimmed              10    8

Unavailable subjects trimmed:
13 15 16 17

Non-informative subjects trimmed:
11 12 14
```

Trimming a pedigree can also be carried out by the built-in R sub-setting utilities, using the indices of the subjects that are to be either kept or removed. For example, the pedigree in Figure S1 can be created from the pedigree in Figure 1 with either of the following commands.

```
   R> pedS1 <- ped1[-c(11,12,13,14,15,16,17)]
   R> pedS1 <- ped1[1:10]
```

KINSHIP MATRIX EXAMPLES

Supplementary Table S1 shows the kinship matrix for the pedigree shown in supplementary Figure S1. Ignoring for a moment the fourth generation, subject 10, the kinship coefficients for the first three generations follow a simple pattern, where self kinship coefficients are 0.50, parent-offspring coefficients are 0.25, and grandparent-grandchild coefficients are 0.125. Avuncular coefficients for subject pairs 4-7, 3-8, and 3-9 are also 0.125, and cousins, pairs 7-8 and 7-9, are half that at 0.0625. Also, because subjects 8 and 9 are identical twins, the kinship coefficients with each other are the same as their self kinship coefficient. Subject 10 in the fourth generation introduces a few scenarios handled seamlessly by the kinship function. First, subject 10 is equally-related to her uncle as to her father because subjects 8 and 9 are identical twins. Furthermore, subject 10's self kinship coefficient is greater than 0.5 because of inbreeding; that is, there is a 1/32=0.03125 chance that the two alleles at a given locus are identical by descent from either of her great-grandparents. The inbreeding coefficient for any subject *i* can be obtained by $h_i=2K_{i,i}-1$; therefore, $h_{10}=0.0625$. The X chromosome kinship matrix for the same 10-subject pedigree is in Table S2, where we have included the sex (M=Male, F=Female) for each subject with the identifiers in rows and columns so it is clear to follow the differences from Table S1.

Table S1 Kinship matrix on autosomes for 10-member pedigree in Figure S1

ID(sex)	1(M)	2(F)	3(M)	4(F)	5(F)	6(M)	7(F)	8(M)	9(M)	10(F)
1(M)	0.50	0.00	0.25	0.25	0.00	0.00	0.13	0.13	0.13	0.13
2(F)	0.00	0.50	0.25	0.25	0.00	0.00	0.13	0.13	0.13	0.13
3(M)	0.25	0.25	0.50	0.25	0.00	0.00	0.25	0.13	0.13	0.19
4(F)	0.25	0.25	0.25	0.50	0.00	0.00	0.13	0.25	0.25	0.19
5(F)	0.00	0.00	0.00	0.00	0.50	0.00	0.25	0.00	0.00	0.13
6(M)	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.25	0.25	0.13
7(F)	0.13	0.13	0.25	0.13	0.25	0.00	0.50	0.06	0.06	0.28
8(M)	0.13	0.13	0.13	0.25	0.00	0.25	0.06	0.50	0.50	0.28
9(M)	0.13	0.13	0.13	0.25	0.00	0.25	0.06	0.50	0.50	0.28
10(F)	0.13	0.13	0.19	0.19	0.13	0.13	0.28	0.28	0.28	0.53

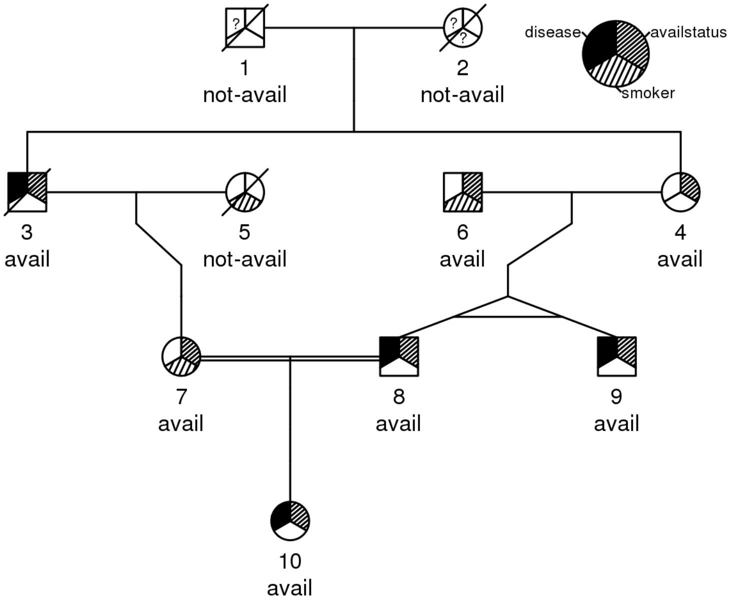
Table S2 Kinship matrix on X chromosome for 10-member pedigree in Figure S1

ID(sex)	1(M)	2(F)	3(M)	4(F)	5(F)	6(M)	7(F)	8(M)	9(M)	10(F)
1(M)	1.00	0.00	0.00	0.50	0.00	0.00	0.00	0.50	0.50	0.25
2(F)	0.00	0.50	0.50	0.25	0.00	0.00	0.25	0.25	0.25	0.25
3(M)	0.00	0.50	1.00	0.25	0.00	0.00	0.50	0.25	0.25	0.38
4(F)	0.50	0.25	0.25	0.50	0.00	0.00	0.13	0.50	0.50	0.31
5(F)	0.00	0.00	0.00	0.00	0.50	0.00	0.25	0.00	0.00	0.13

6(M)	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
7(F)	0.00	0.25	0.50	0.13	0.25	0.00	0.50	0.13	0.13	0.31
8(M)	0.50	0.25	0.25	0.50	0.00	0.00	0.13	1.00	1.00	0.56
9(M)	0.50	0.25	0.25	0.50	0.00	0.00	0.13	1.00	1.00	0.56
10(F)	0.25	0.25	0.38	0.31	0.13	0.00	0.31	0.56	0.56	0.56

FIGURE LEGENDS:

Figure S1: A detailed pedigree plot with text added to the subject identifier. Subject symbols show the status for three affected indicators, with a legend in the top right giving the portions attributed to each variable.


disease
availstatus
smoker
?
?
?
1
not-avail
2
not-avail
3
avail
5
not-avail
6
avail
4
avail
7
avail
8
avail
9
avail
10
avail