Original Paper

Human Heredity

Hum Hered 2014;78:91–93
DOI: 10.1159/000363105

Received: February 11, 2014
Accepted after revision: April 23, 2014
Published online: July 29, 2014

The kinship2 R Package for Pedigree Data

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

Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, Minn., USA

Key Words
Pedigrees · Genetic linkage analysis · Kinship · Graphics

Abstract
Background: The kinship2 package is restructured from the previous kinship package. Existing features are now enhanced and new features added for handling pedigree objects. ***Methods:*** Pedigree plotting features have been updated to display features on complex pedigrees while adhering to pedigree plotting standards. Kinship matrices can now be calculated for the X chromosome. Other methods have been added to subset and trim pedigrees while maintaining the pedigree structure. ***Conclusion:*** We make the kinship2 package available for R on the Contributed R Archives Network (CRAN), where data management is built-in and other packages can use the pedigree object.

Introduction

Pedigrees have long been used in genetic linkage and association studies, and pedigree kinship matrices are widely employed in mixed-effects models for survival and regression analyses. The original kinship package, developed by Terry Therneau and ported to R by Jing Hua Zhao [1], was developed to accompany the coxme package [2] which extends the Cox model to include kinship matrices for related subjects. The kinship2 package has major updates to the core functions to calculate kinship matrices, plot pedigrees, and trim pedigree objects. Family-based studies for linkage, association, and mixed effects need the capability to keep track of pedigree relationships, phenotypes, covariates, and subjects that are informative for the analyses. We utilize the object-oriented framework within R to manage pedigree objects, where the objects contain data for pedigree members.

Methods

Pedigree Plots

Pedigree plots are widely used to check the accuracy of pedigree data. Our pedigree plotting routines within R have compared favorably to other stand-alone and R plotting routines [1, 3–5]. Now in kinship2, we have made the plots more robust and adhere to more of the pedigree plotting standards discussed in Bennett et al. [6]. We provide an example of some of the key additions in figure 1; additional features are demonstrated in the online supplementary material (www.karger.com/doi/10.1159/000363105).

Pedigree Trimming

Pedigree analyses are often based on a subset of pedigree members that are informative for a specific analysis. For example, subjects who are not genotyped and do not have offspring are not

KARGER

E-Mail karger@karger.com
www.karger.com/hhe

0001–5652/14/0782–0091$39.50/0

Jason P. Sinnwell
Division of Biomedical Statistics and Informatics
Department of Health Sciences Research, Mayo Clinic
200 1st Ave SW, Rochester, MN 55905-0001 (USA)
E-Mail sinnwell.jason@mayo.edu

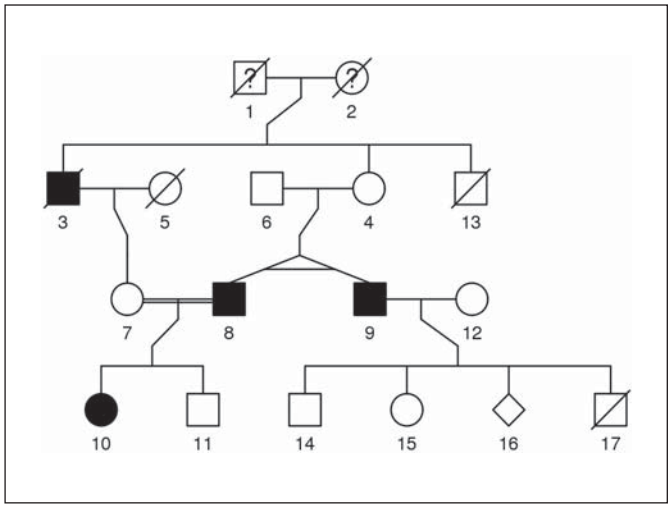

Fig. 1. Four-generation pedigree plot with shapes for male (squares), female (circles), and unknown gender (diamond). The shapes are black, white, or filled with a question mark (?) for a disease status of yes, no, and unknown, respectively. A diagonal line is drawn through the shape of those who have deceased. Twins are represented as diagonal lines split from a single mating event with a horizontal line drawn to both diagonal lines indicating the twins are monozygotic. A consanguineous marriage is indicated by two horizontal lines connecting the parents rather than by one line.

informative for linkage or association analyses and can be trimmed from the pedigree. However, some uninformative subjects are needed to maintain links among relatives to in turn maintain a valid pedigree. The pedigree.shrink function trims uninformative pedigree members while maintaining a valid pedigree structure down to a specific bit size, a metric used for allocating memory in some pedigree linkage analysis packages [7], defined as $bit\_size = (2 \times M) - N$, where M and N are the number of nonfounders and founders within the pedigree, respectively. For a more detailed description of the algorithm with a working example, see the online supplementary material.

Kinship Matrices

The kinship coefficient for any 2 subjects is the probability that an allele chosen at random for both subjects at a given locus is identical-by-descent, that is, inherited from a common ancestor. The computations in the kinship function are based on a recursive algorithm described in Lange [8], which assumes the founders are not inbred. Kinship matrix K can be transformed to a genetic correlation matrix by $2 \times K$, which can be used in mixed models to estimate the additive genetic effect of alleles on phenotypes.

The kinship function is now able to calculate the kinship matrix for the X chromosome. The recursive algorithm accounts for the asymmetry of the X chromosome transmission between males and females. Males will have $K_{i,i} = 1$ for the X chromosome, whereas for females $K_{i,i}$ is the same as for autosomes. The genetic correlation matrix for the X chromosome is $2 \times K_{i,j}$ for a pair of females, $K_{i,j}$ for a pair of males, and $(2^{1/2}) \times K_{i,j}$ for a male-female pair. How the genetic correlations influence mixed models depends on assumptions of dosage compensation, that is, the amount of inactivation of one of the two X chromosomes in females [9]. Examples for both autosomes and the X chromosome are provided in the online supplementary material.

We provide a framework to keep kinship matrices for multiple pedigrees in one object. A pseudo-code example for creating a kinship matrix from a pedigreeList object of multiple pedigrees is as follows. Given a data frame called 'families' with multiple families indexed by the 'pedid' variable:

```
R> pedlist <- with(families,
+     pedigree(id=sid, dadid=fid, momid=mid, sex=sex, famid=pedid))
R> kinmat <- kinship(pedlist)
```

The kinship matrix is symmetric, and the kinship matrix for multiple families is block-diagonal with each family a block along the diagonal. The Matrix R package (Matrix.R-forge.R-project.org/) provides methods for the efficient storage and manipulation of these sparse matrices. This is useful because the genetic correlation for related subjects in mixed models, as noted above, can be created using kinship matrices, which can be used with other covariance structures in modeling related subjects [10, 11]. Some simple tests on K replicates of the 17-member pedigree in figure 1 give a run time and storage of the kinship matrix of 1.2 s and 49 Kb, respectively, for 200 replicates, and of 1.85 s and 92 Kb, respectively, for 400 replicates (which is a linear increase on both measures), which are sufficient for most analyses.

Discussion

We have highlighted the pedigree object and the core functions plot.pedigree, pedigree.shrink, and kinship. While other methods for pedigree analysis and plotting exist [3–5], these functions have few competitors within the R framework, where data management is built-in and methods have been written to use pedigree objects. These advantages make the routines in kinship2 suitable for studies where screening steps and analyses are performed across many families. We have made the kinship2 package available on the Contributed R Archives Network cran.r-project.org/web/packages/kinship2 with a vignette. We invite feedback and contributions.

Acknowledgements

We would like to thank Elizabeth Atkinson, Martha Matsumoto, Shannon McDonnell, and Jing Hua Zhao for contributions and feedback. This research was supported by the U.S. Public Health Service, National Institutes of Health, contract grant No. GM065450.

References

1 Zhao JH: Pedigree-drawing with R and graphviz. Bioinformatics 2006;22:1013–1014.
2 Therneau TM, Grambsh PM: Modeling Survival Data: Extending the Cox Model. New York, Springer, 2000.
3 Fuchsberger C, Falchi M, Forer L, Pramstaller PP: PedVizApi: a Java API for the interactive, visual analysis of extended pedigrees. Bioinformatics 2008;24:279–281.
4 Trager EH, Khanna R, Marrs A, Siden L, Branham KE, Swaroop A, Richards JE: Madeline 2.0 PDE: a new program for local and web-based pedigree drawing. Bioinformatics 2007;23:1854–1856.
5 Mäkinen VP, Parkkonen M, Wessman M, Groop PH, Kanninen T, Kashi K: High-throughput pedigree drawing. Eur J Hum Genet 2005;13:987–989.
6 Bennett RL, French KS, Resta RG, Doyle DL: Standardized human pedigree nomenclature: update and assessment of the recommendations of the National Society of Genetic Counselors. J Genet Couns 2008;17:424–433.
7 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 2002;30:97–101.
8 Lange K: Mathematical and Statistical Methods for Genetics Analysis. New York, Springer, 1997.
9 Kent JW Jr, Dyer TD, Blangero J: Estimating the additive genetic effect of the X chromosome. Genet Epidemiol 2005;29:377–388.
10 Schifano ED, et al: SNP set association analysis for familial data. Genet Epidemiol 2012, Epub ahead of print.
11 Schaid DJ, McDonnell SK, Sinnwell JP, Thibodeau SN: Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genet Epidemiol 2013;37:409–418.