An Integrated Genetic Analysis Package Using R

Jing Hua Zhao

MRC Epidemiology Unit, Cambridge, UK
http://www.mrc-epid.cam.ac.uk

Contents

1 Introduction

This package was initiated to integrate some C/Fortran/SAS programs I have written or used over the years. As such, it would rather be a long-term project, but an immediate benefit would be something complementary to other packages currently available from CRAN, e.g. **genetics**, **hwde**, etc. I hope eventually this will be part of a bigger effort to fulfill most of the requirements foreseen by many, e.g. Guo and Lange (2000), within the portable environment of R for data management, analysis, graphics and object-oriented programming. My view has been outlined more formally in Zhao and Tan (2006b) and Zhao and Tan (2006a) in relation to other package systems. Also reported are Zhao (2005) and Zhao (2006) on package **kinship**.

The number of functions are quite limited and experimental, but I already feel the enormous advantage by shifting to R and would like sooner rather than later to share my work with others. I will not claim this workas exclusively done by me, but would like to invite others to join me and enlarge the collections and improve them.

2 Implementation

The following list shows the data and functions currently available.

```
BFDP                    Bayesian false-discovery probability
FPRP                    False-positive report probability
```

aldh2	ALDH2 markers and Alcoholism
apoeapoc	APOE/APOC1 markers and Schizophrenia
bt	Bradley-Terry model for contingency table
ccsize	Power and sample size for case-cohort design
chow.test	Chow's test for heterogeneity in two regressions
cf	Cystic Fibrosis data
comp.score	score statistics for testing genetic linkage of quantitative trait
crohn	Crohn disease data
fa	Friedreich Ataxia data
fbsize	Sample size for family-based linkage and association design
fsnps	A case-control data involving four SNPs with missing genotype
gc.em	Gene counting for haplotype analysis
gcontrol	genomic control
gcp	Permutation tests using GENECOUNTING
genecounting	Gene counting for haplotype analysis
gif	Kinship coefficient and genetic index of familiality
hap	Haplotype reconstruction
hap.em	Gene counting for haplotype analysis
hap.score	Score Statistics for Association of Traits with Haplotypes
hla	HLA markers and Schizophrenia
htr	Haplotype trend regression
hwe	Hardy-Weinberg equlibrium test for multiallelic marker
hwe.hardy	Hardy-Weinberg equlibrium test using MCMC
kbyl	LD statistics for two multiallelic loci
kin.morgan	kinship matrix for simple pedigree
makeped	A function to prepare pedigrees in post-MAKEPED format
mao	A study of Parkinson's disease and MAO gene
mia	multiple imputation analysis for hap
mtdt	Transmission/disequilibrium test of a multiallelic marker
muvar	Means and variances under 1- and 2- locus (biallelic) QTL model
nep499	A study of Alzheimer's disease with eight SNPs and APOE
pbsize	Power for population-based association design
pedtodot	Converting pedigree(s) to dot file(s)
pfc	Probability of familial clustering of disease
pfc.sim	Probability of familial clustering of disease
pgc	Preparing weight for GENECOUNTING
plot.hap.score	Plot Haplotype Frequencies versus Haplotype

```
                        Score Statistics
print.hap.score         Print a hap.score object
qqunif                  Q-Q plot for uniformly distributed random
                        variable
s2k                     Statistics for 2 by K table
snca                    A study of Parkinson's disease and SNCA makers
tbyt                    LD statistics for two SNPs
tscc                    Power calculation for two-stage case-control design
twinan90                Classic twin models
whscore                 Whittemore-Halpern scores for allele-sharing
```

Assuming proper installation, you will be able to obtain the list by typing *library(help=gap)* or view the list within a web browser via *help.start()*. A PDF version of this file can be viewed with command *vignette("gap",package="gap")*.

You can cut and paste examples at end of each function's documentation.

Both *genecounting* and *hap* are able to handle SNPs and multiallelic markers, with the former be flexible enough to include features such as X-linked data and the later being able to handle large number of SNPs. But they are unable to recode allele labels automatically, so functions *gc.em* and *hap.em* are in *haplo.em* format and used by a modified function *hap.score* in association testing.

It is notable that multilocus data are handled differently from that in **hwde** and elegant definitions of basic genetic data can be found in **genetics** package.

Incidentally, I found my C mixed-radixed sorting routine as in Zhao and Sham (2003) is much faster than R's internal function.

With exceptions such as function *pfc* which is very computer-intensive, most functions in the package can easily be adapted for analysis of large datasets involving either SNPs or multiallelic markers. Some are utility functions, e.g. *muvar* and *whscore*, which will be part of the other analysis routines in the future.

For users, all functions have unified format. For developers, it is able to incorporate their C/C++ programs more easily and avoid repetitive work such as preparing own routines for matrix algebra and linear models. Further advantage can be taken from packages in **Bioconductor**, which are designed and written to deal with large number of genes.

3 Examples

Examples can be found from most function documentations. You can also try several simple examples via *demo*:

```
library(gap)
demo(gap)
```

4 Known bugs

Unaware of any bug. However, better memory management is expected.

5 Bibliographic note

The main references are Chow (1960), Guo and Thompson (1992), Williams et al. (1992), Gholamic and Thomas (1994), Risch and Merikangas (1996), Spielman and Ewens (1996), Risch and Merikangas (1997), Miller (1997), Sham (1997), Sham (1998), Devlin and Roeder (1999), Zhao et al. (1999), Guo and Lange (2000), Hirotsu et al. (2001), Zhao et al. (2002), Zaykin et al. (2002), Zhao (2004), Wacholder et al. (2004), Wang (2005), Skol et al. (2006), Wakefield (2007).

References

G. C. Chow. Tests of equality between sets of coefficients in two linear regression. *Econometrica*, 28:591–605, 1960.

B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999.

K. Gholamic and A. Thomas. A linear time algorithm for calculation of multiple pairwise kinship coefficients and genetic index of familiality. *Comp Biomed Res*, 27:342–350, 1994.

S. W. Guo and K. Lange. Genetic mapping of complex traits: promises, problems, and prospects. *Theor Popul Biol*, 57:1–11, 2000.

S. W. Guo and E. A. Thompson. Performing the exact test of hardy-weinberg proportion for multiple alleles. *Biometrics*, 48:361–372, 1992.

C. Hirotsu, S. Aoki, T. Inada, and Y. Kitao. An exact test for the association between the disease and alleles at highly polymorphic loci with particular interest in the haplotype analysis. *Biometrics*, 57:769–778, 2001.

M. B. Miller. Genomic scanning and the transmission/disequilibrium test: analysis of error rates. *Genet Epidemiol*, 14:851–856, 1997.

N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(September):1516–1517, 1996.

N. Risch and K. Merikangas. Reply to scott el al. *Science*, 275:1329–1330., 1997.

P. C. Sham. Transmission/disequilibrium tests for multiallelic loci. *Am J Hum Genet*, 61: 774–778, 1997.

P. C. Sham. *Statistics in Human Genetics*. Arnold Applications of Statistics Series. Edward Arnold, London, 1998. 11-1-1999.

A. D. Skol, L. J. Scott, G. R. Abecasis, and M. Boehnke. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet*, 38(2): 209–13, 2006.

R. S. Spielman and W. J. Ewens. The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet*, 59(5):983–9, 1996.

S. Wacholder, S. Chanock, M. Garcia-Closas, L. El Ghormli, and N. Rothman. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst*, 96(6):434–42, 2004.

J. Wakefield. A bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet*, 81:208–226, 2007.

K. Wang. A likelihood approach for quantitative-trait-locus mapping with selected pedigrees. *Biometrics*, 61:465–473, 2005.

C. J. Williams, J. C. Christian, and J.A. Jr. Norton. Twinan90: A fortran programfor conducting anova-based and likelihood-based analyses of twin data. *Comp Meth Prog Biomed*, 38(2-3):167–76, 1992.

D. V. Zaykin, P. H. Westfall, S. S. Young, M. A. Karnoub, M. J. Wagner, and M. G. Ehm. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered*, 53(2):79–91, 2002.

J. H. Zhao. 2LD. GENECOUNTING and HAP: computer programs for linkage disequilibrium analysis. *Bioinformatics*, 20:1325–6, 2004.

J. H. Zhao. Mixed-effects Cox models of alcohol dependence in extended families. *BMC Genet*, 6(Suppl):S127, 2005.

J. H. Zhao. Pedigree-drawing with R and graphviz. *Bioinformatics*, 22(8):1013–4, 2006.

J. H. Zhao, S. Lissarrague, L. Essioux, and P. C. Sham. GENECOUNTING: haplotype analysis with missing genotypes. *Bioinformatics*, 18(12):1694–5, 2002.

J. H. Zhao and P. C. Sham. Generic number systems and haplotype analysis. *Comp Meth Prog Biomed*, 70:1–9, 2003.

J. H. Zhao, P. C. Sham, and D. Curtis. A program for the Monte Carlo evaluation of significance of the extended transmission/disequilibrium test. *Am J Hum Genet*, 64:1484–1485, 1999.

J. H. Zhao and Q. Tan. Genetic dissection of complex traits in silico: approaches, problems and soluations. *Current Bioinformatics*, 1(3):359–369, 2006a.

J. H. Zhao and Q. Tan. Integrated analysis of genetic data with R. *Hum Genomics*, 2(4): 258–65, 2006b.