An Integrated Genetic Analysis Package Using R

Jing hua Zhao

Department of Epidemiology and Public Health, Unversity College London
http://www.ucl.ac.uk/∼rmjdjhz, http://www.hgmp.mrc.ac.uk/∼jzhao

Contents

1 Introduction

This package was designed 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 in R, e.g. **genetics**, **hwde**, **haplo.score**, 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.

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

2 Implementation

The following, extracted from the package INDEX, shows the data and functions currently available.

```
aldh2                   ALDH2 markers and Alcoholism
apoeapoc                APOE/APOC1 markers and Schizophrenia
bt                      Bradley-Terry model for contingency table
chow.test               Chow's test for heterogeneity in two
```

	regressions
cf	Cystic Fibrosis data
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
s2k	Statistics for 2 by K table
snca	A study of Parkinson's disease and SNCA makers
tbyt	LD statistics for two SNPs
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()*.

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 (not incorporated yet) and the later being able to handle large number of SNPs, an advantage over algorithms in **haplo.score**. But the latter is able to recode allele labels automatically, so functions *gc.em* and *hap.em* are in **haplo.score**'s *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 mixed-radixed sorting routine in C (Zhao & 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 after hwe.hardy was fixed. However, better memory management is expected.

5 References

Chow GC (1960). Tests of equality between sets of coefficients in two linear regression. Econometrica 28:591-605

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997-1004

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

Guo S-W, Thompson EA (1992) Performing the Exact Test of Hardy-Weinberg Proportion for Multiple Alleles. Biometrics. 48:361–372.

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

Hirotsu C, Aoki S, Inada T, Kitao Y (2001) 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

Miller MB (1997) Genomic scanning and the transmission/disequilibrium test: analysis of error rates. Genet Epidemiol 14:851-856

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

Risch N, Merikangas K (1997). Reply to Scott el al. Science 275(February): 1329-1330.

Sham PC (1997) Transmission/disequilibrium tests for multiallelic loci. Am J Hum Genet 61:774-778

Sham PC (1998). Statistics in Human Genetics. Arnold

Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59:983-989

Zapata C, Carollo C, Rodriquez S (2001) Sampleing variance and distribution of the D' measure of overall gametic disequilibrium between multiallelic loci. Ann Hum Genet 65: 395-406

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53:79-91

Zhao JH, Lissarrague S, Essioux L, Sham PC (2002). GENECOUNTING: haplotype analysis with missing genotypes. Bioinformatics 18(12):1694-1695

Zhao JH, Sham PC, Curtis D (1999) A program for the Monte Carlo evaluation of significance of the extended transmission/disequilibrium test. Am J Hum Genet 64:1484-1485

Zhao JH, Sham PC (2003). Generic number systems and haplotype analysis. Comp Meth Prog Biomed 70: 1-9

Zhao JH (2004). 2LD, GENECOUNTING and HAP: Computer programs for linkage disequilibrium analysis. Bioinformatics, 20, 1325-1326