

# An Integrated Genetic Analysis Package Using R

Jing hua Zhao

Department of Epidemiology and Public Health, University College London  
<http://www.ucl.ac.uk/~rmjdjh/>, <http://www.hgmp.mrc.ac.uk/~jzhao>

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Implementation</b>	<b>1</b>
<b>3</b>	<b>Examples</b>	<b>2</b>
<b>4</b>	<b>Known bugs</b>	<b>3</b>
<b>5</b>	<b>References</b>	<b>3</b>

## 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.

<code>aldh2</code>	ALDH2 markers and Alcoholism
<code>apoeapoc</code>	APOE/APOC1 markers and Schizophrenia
<code>bt</code>	Bradley-Terry model for contingency table
<code>chow.test</code>	Chow's test for heterogeneity
<code>fbsize</code>	Sample size for family-based linkage and association design
<code>fsnps</code>	A case-control data involving four SNPs for missing genotype
<code>gc.em</code>	Gene counting for haplotype analysis
<code>gcontrol</code>	Genomic control
<code>genecounting</code>	Gene counting for haplotype analysis
<code>gif</code>	Kinship coefficient and genetic index of familiarity

<code>hap</code>	Haplotype reconstruction
<code>hap.em</code>	Gene counting for haplotype analysis
<code>hap.score</code>	Score Statistics for Association of Traits with haplotypes
<code>hla</code>	HLA markers and Schizophrenia
<code>hwe.hardy</code>	Hardy-Weinberg equilibrium test
<code>htr</code>	Haplotype trend regression (with permutation)
<code>kbyl</code>	LD statistics for two multiallelic loci
<code>kin.morgan</code>	kinship matrix for simple pedigree
<code>makeped</code>	A function to prepare post-MAKEPED format file
<code>mia</code>	Multiple imputation analysis for hap
<code>mtdt</code>	Transmission/disequilibrium test of a multiallelic marker
<code>muvar</code>	Means and variances under 1- and 2- locus QTL model
<code>pbsize</code>	Sample size for population-based association design
<code>pfc</code>	Probability of familial clustering of disease
<code>pgc</code>	Preparing frequency weight for GENECOUNTING
<code>s2k</code>	Statistics for 2 by K table
<code>tbyt</code>	LD statistics for two SNPs
<code>whscore</code>	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 being flexible enough to include features such as X-linked data (not incorporated yet) and the latter 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.demo)
```

## 4 Known bugs

There appears to be some problem with the memory management in *hwe.hardy* which needs to be fixed.

## 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
- Gholami K, Thomas A (1994) A linear time algorithm for calculation of multiple pairwise kinship coefficients and genetic index of familiarity. *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 et 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, Rodriguez S (2001) Sampling 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). Gene-counting for 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