Journal of Statistical Software

December 2007, Volume 23, Issue 8. doi: 10.18637/jss.v000.i00

# gap: Genetic Analysis Package

**Jing Hua Zhao**
MRC Epidemiology Unit

**Abstract**

A preliminary attempt at collecting tools and utilities for genetic data as an R package called **gap** is described. Genomewide association is then described as a specific example, linking the work of Risch and Merikangas (1996), Long *et al.* (1997) for family-based and population-based studies, and the counterpart for case-cohort design established by Cai and Zeng (2004). Analysis of staged design as outlined by Skol *et al.* (2006) and associate methods are discussed. The package is flexible, customizable, and should prove useful to researchers especially in its application to genomewide association studies.

*Keywords*: genetic data analysis, genomewide association, R.

## 1. Introduction

Approaches to understanding the genetic basis of human diseases have been widely discussed, e.g., Morton *et al.* (1983), Khoury *et al.* (1993), Thomas (2004). Methods include the assessment of familial aggregation for heritability, identification of major gene effect, study of cosegregation of genetic marker with putative disease-predisposing loci in the so-called linkage studies and association studies in search of frequency differences between cases and controls and/or correlation between genotype and phenotype as a quantitative trait. Recently, owing to the availability of large number of genetic variants and particularly single nucleotide polymorphisms (SNPs), attention has focused on association designs including both families and unrelated individuals from general populations. Three initiatives of interest are:

1. The hapmap project (`http://www.hapmap.org/`), a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals.

2. The Wellcome Trust Case Control Consortium (WTCCC, `http://www.wtccc.org.uk/`), a collaboration of human geneticists who analyse thousands of DNA samples

from patients suffering with different diseases to identify common genetic variations for each condition. It is hoped that by identifying these genetic signposts, researchers will be able to understand which people are most at risk, and also produce more effective treatments. The WTCCC searches for the genetic signposts for tuberculosis, coronary heart disease, type 1 diabetes, type 2 diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder and hypertension. The research is conducted at a number of institutes throughout the UK, including the Wellcome Trust Sanger Institute, Cambridge University and Oxford University.

3. The genetic association information network (GAIN, `http://www.fnih.org/GAIN2/home_new.shtml`), a public-private partnership of the Foundation for the National Institutes of Health, Inc., which include corporations, private foundations, advocacy groups, concerned individuals, and the National Institutes of Health. This initiative will take the next step in the search to understand the genetic factors influencing risk for complex human diseases.

Following successes in localization of Mendelian diseases such as Huntington's disease (Gusella *et al.* 1983), cystic fibrosis (Tsui *et al.* 1985), some recent successes in non-Mendelian diseases come from breast cancer (Hall *et al.* 1990), macular degenration (Klein *et al.* 2005), and non-insulin-dependent (Type-2) diabetes (Grant *et al.* 2006). Here, we will focus specifically on methodological issues concerning the design and analysis of genomewide association studies as largely foreseen by the seminal paper of Risch and Merikangas (1996). As it dealt with the power of association studies using family-based designs, an immediate argument for case-control designs as an alternative to family-based designs was made by Long *et al.* (1997) on the basis of (1). its ease of implementation; (2). the increased prospects for extension from established epidemiological cohorts; (3). for late onset diseases such as Type-2 diabetes and hypertension, the difficulty of typing parents for family-based studies; (4). the issue that, for equivalent power, the number of individuals genotyped needs to be doubled for linkage, tripled for singleton, and quadrupled for sib-pair designs, assuming both parents are genotyped in an affected sibling study. Concerns over cost-efficiency have led to the adoption of staged design in some studies. In the most popular two-staged design, the first stage uses only a proportion of individuals and are genotyped at all SNPs, among which a percentage of SNPs showing statistical significance is carried over to the second stage. For analysis of two-staged design, Skol *et al.* (2006) recognized that joint analysis could be more powerful. The high cost of genotyping has motivated researchers to seek generic controls for a range of traits as in a case-cohort design, in which a small random sample of the whole cohort and all the diseased subjects are used. It is possible the subcohort also contains some cases but it is fully representative of the population and can be used in conjunction with a range of case definitions. The merits of such a design in genetic association studies have only recently been recognized (Langholz *et al.* 1999; Manolio *et al.* 2006). Some of the terms used in this paper can be found in the paper by Burkett *et al.* (2006), and more generally in a review paper by Elston and Anne Spence (2006) on recent developments in statistical genetics, and in a tutorial specifically for genetic association analysis by Balding (2006).

Analysis for population association studies generally involves preliminary analyses such as quality control, Hardy-Weinberg equilibrium tests, examination of linkage disequilibrium and recombination, to be followed by tests of association for single and/or multiple SNPs, both may involve case-control or binary phenotype or continuous outcomes. Further issues involve

handling of large data, multiple testing, among others. The implementation of analytical approaches to these problems has so far scattered and not integrated with established software systems. Preliminary reviews have shown that R is a strong alternative which offers many desirable facilities and can be used in conjunction with other software systems (Zhao and Tan 2006). It is possible that in the near future, many tools for genetic data analysis will be available on these software systems.

We have recently carried out a study of obesity using Affymetrix (http://www.affymetrix.com/) 500K and Illumina (http://www.illumina.com/) 317K systems following an earlier pilot using Perlegen (http://www.perlegen.com/) platform with about 250K SNPs. The samples were based on European Prospective Investigation of Cancer (EPIC) Norfolk study (http://www.srl.cam.ac.uk/epic/). Our particular concern is cost-efficiency, not only because the experiment is expensive but we wish to take full advantage of it. We adopted a two-stage case-cohort design which is advantageous in cost-efficiency over the standard case-control design. While seeking for appropriate methods for joint analyses, we have recast the problem within the missing data framework. We believe our approach will be invaluable to other colleagues.

Here, I describe a genetic analysis package, **gap**, created and maintained by the author and is available from the Comprehensive R Archive Network (CRAN, http://CRAN.R-project.org/), taking advantages of the portable environment of R for data management, analysis, graphics and object-oriented programming. The functions implemented for genetic association analysis include Hardy-Weinberg equilibrium tests, measures of linkage disequilibrium between SNPs or multiallelic markers and haplotype analysis. I will give a brief overview of the package, and a specific example concerning design and analysis of genetic association, which was used in the design of the case-cohort study aforementioned. I will also use a dataset from our case-control association study of Type-2 diabetes with chromosome 20, to illustrate joint analysis within R. It is clear that many statistical issues can be handled with ease and used in conjunction with other packages available from CRAN. The calculation as in some original work is also given for reference, with results not seen in the genetic analysis literature. At the end I will provide a brief summary.

## 2. Overview of implemented methods

This package was designed in 2003 to integrate some programs in C/Fortran and SAS (SAS Institute Inc. 2004) I have written or used over the years, by taking advantage of the ability of R to use foreign language routines in C/Fortran or its facilities similar to SAS. The recent implementations are more native to R. The collection therefore somewhat reflects the evolution of (or more exactly my appreciation of) statistical genetics (or genetic epidemiology) over years. The description of the main functions is given in Table 1, while a list of data sets is given in Table 2. They can be conveniently obtained with R command `library(help = "gap")` once **gap** is installed.

Some of the functions are highlighted as follows:

- `BFDP`: Bayesian false-discovery probability according to Wakefield (2007).
- `FPRP`: False-positive report probability according to Wacholder *et al.* (2004).

| Function name | Description |
|---|---|
| `BFDP` | Bayesian false-discovery probability |
| `FPRP` | False-positive report probability |
| `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 |
| `fbsize` | Sample size for family-based linkage and association design |
| `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 |
| `htr` | Haplotype trend regression |
| `hwe` | Hardy-Weinberg equilibrium test for a multiallelic marker |
| `hwe.hardy` | Hardy-Weinberg equilibrium test using MCMC |
| `kbyl` | Linkage disequilibrium statistics for two multiallelic loci |
| `kin.morgan` | Kinship matrix for simple pedigree |
| `makeped` | A function to prepare pedigrees in post-**MAKEPED** format |
| `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 |
| `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 |
| `tbyt` | Linkage disequilibrium statistics for two SNPs |
| `tscc` | Power calculation for two-stage case-control design |
| `twinan90` | Classic twin models |
| `whscore` | Whittemore-Halpern scores for allele-sharing |

Table 1: A list of functions in **gap**.

| Data sets | Description |
|---|---|
| `aldh2` | ALDH2 markers and alcoholism |
| `apoeapoc` | APOE/APOC1 markers and schizophrenia |
| `cf` | Cystic fibrosis data |
| `crohn` | Crohn's disease data |
| `fa` | Friedreich ataxia data |
| `fsnps` | A case-control data involving four SNPs with missing genotype |
| `hla` | HLA markers and schizophrenia |
| `mao` | A study of Parkinson's disease and MAO gene |
| `nep499` | A study of Alzheimer's disease with eight SNPs and APOE |
| `snca` | A study of Parkinson's disease and SNCA markers |

Table 2: A list of data sets in **gap**.

- `gif`: kinship coefficient and genetic index of familiality according to Gholami and Thomas (1994).
- `genecounting`, `gcp`, `gc.em`, `pgc`: expectation-maximization (EM) method to infer haplotype and case-control haplotype association similar to SAS/**Genetics** (SAS Institute Inc. 2005) as reported in Zhao *et al.* (2002), Zhao (2004).
- `hap`, `hap.em`, `mia`: a multiallelic version of progressive EM algorithm for haplotype inference as reported in Zhao (2004).
- `tbyt`, `kbyl`: linkage disequilibrium statistics between two biallelic or multiallelic markers, including standard error for $D'$ as reported in Zhao (2004).
- `pfc`, `pfc.sim`: probability of familial clustering of diseases according to Yu and Zelterman (2002).
- `htr`: haplotype-trend regression as accorded to Zaykin *et al.* (2002).
- `hap.score`, `print.hap.score`, `plot.hap.score`: score statistics for haplotype-trait association with revised algorithms according to Schaid *et al.* (2002).
- `pbsize`, `fbsize`, `ccsize`: sample size/power for population-based, family-based case-control designs, or population-based case-cohort design, according to Long *et al.* (1997), Risch and Merikangas (1996), Cai and Zeng (2004).
- `s2k`: statistics from $2 \times K$ table according to Hirotsu *et al.* (2001).
- `hwe`, `hwe.hardy`: Hardy-Weinberg equilibrium test for multiallelic markers, the latter implements a Markov chain Monte Carlo (MCMC) algorithm according to Guo and Thompson (1992).
- `makeped`: a utility to prepare for post-**MAKEPED** format file as required by **LINKAGE** (Lathrop *et al.* 1984) or other computer programs, a version distributed with **LINKAGE** by Wentian Li.

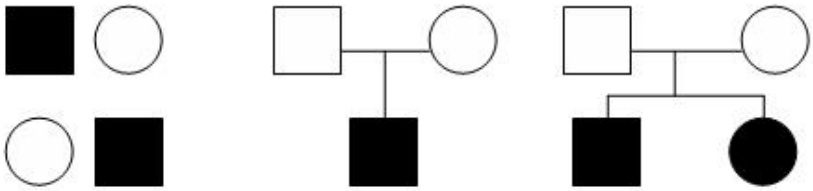

Figure 1: Three common genetic association designs involving unrelated individuals (left), nuclear families with affected singletons (middle) and affected sib-pairs (right). Males and females are denoted by squares and circles with affected individuals filled with black and unaffect individuals being empty.

- `pedtodot`: a utility to produce **Graphviz** (AT&T Research Lab 2008) command file based on a GAWK script by David Duffy.
- `tscc`: power of two-stage case-control design under several models, similar to Skol *et al.* (2006).
- `muvar`, `whscore`: utility functions to obtain mean and variance under two-locus model, and score (Whittemore and Halpern 1994).

A brief summary of a given function, including appropriate reference(s), can be obtained with `help(function-name, package = "gap")` within R. There are a number of example datasets. For instance, the cystic fibrosis data (Kerem *et al.* 1989) has been used in many papers on fine-mapping.

## 3. Example: Study design and analysis of genetic association

Since the range of problems to be tackled by **gap** is quite broad, I now give specific examples of how some functions have been developed to facilitate our design and analysis of genetic association. The three popular designs outlined earlier are shown in Figure 1.

Because of the importance of work by Risch and Merikangas (1996) and Long *et al.* (1997), functions `pbsize` and `fbsize` have been written to implement their methods and correct a programming error in Risch and Merikangas (1996). The table to be obtained could be used as a quick reference and the code be tailored to particular designs. Power for case-cohort design established by Cai and Zeng (2004) is given in function `ccsize`. The method by Skol *et al.* (2006) for two-stage design has been implemented in function `tscc`.

### 3.1. Some preliminaries

A widely used general model consists of a disease susceptible locus with two alleles (A and a) and population frequencies $p$, $q = 1 - p$, and penetrances as $f_0$, $f_1$ and $f_2$ (i.e., the

conditional probability of getting the disease, the subscripts 0, 1 and 2 are copies of the disease-predisposing allele A). Hardy-Weinberg equilibrium states that the distribution of the three genotypes, aa, Aa, and AA are according to $(p+q)^2$, so that the population prevalence of the disease $K = p^2f^2 + 2pqf_1 + q^2f_0$. Other statistics, which are often used, are additive $V_A = 2pq[p(f_2 - f_1) + q(f1 - f_0)]^2$ and dominance $V_D = p^2q^2(f_2 - 2f_1 + f_0)^2$, variances, respectively. It is also common to specify further the relationship between the three genotypes. As for multiplicative models to be considered here, the genotypic relative risk (GRR) is $\gamma$ and $\gamma^2$ for genotypes Aa and AA compared to aa as baseline, the population disease prevalence, the additive and dominance variances are therefore $K = p^2\gamma^2 + 2pq\gamma + q^2 = (p\gamma + q)^2$, $V_A = 2pq(\gamma - 1)^2(p\gamma + q)^2$ and $V_D = p^2q^2(\gamma - 1)^4$, respectively. These can be used to define offspring/sibling relative risks, which are the probability ratios of a offspring/sibling of an affected parent/child being affected relative to the general population and they are shown to be $\lambda_O = 1+1/2V_A/K^2$ and $\lambda_S = 1+(V_A/2 + V_D/4)/K^2$ so $\lambda_O = 1+w$ and $\lambda_S = (1+1/2w)^2$, where $w = (pq(\gamma - 1)^2)/(p\gamma + q)^2$.

Linkage and association designs using family data were the state-of-the-art designs for localising disease-predisposing loci. A key concept relates to genes shared by relatives from common ancestor and is called identity by descent (IBD), which is associated with a model-free approach of linkage whereby allele-sharing between affected members in a pedigree is compared to test for linkage between a marker locus and a disease locus. The simplest allele-sharing method uses affected sib-pairs. Assuming the recombination fraction between the disease locus and a marker to be $\theta$, and defining $\Psi = \theta^2+(1-\theta)^2$, Suarez *et al.* (1978) derived the IBD probabilities for affected sib-pair, $P(IBD = 0) = 1/4-[(\Psi-1/2)V_A+(2\Psi-\Psi^2-3/4)V_D]/[4(K^2+V_A/2+V_D/4)]$, $P(IBD = 1) = 1/2-[2(\Psi^2-\Psi+1/4)V_D]/[4(K^2+V_A/2+V_D/4)]$, $P(IBD = 2) = 1/4 + [(\Psi - 1/2)V_A + (\Psi^2 - 1/4)V_D]/[4(K^2 + V_A/2 + V_D/4)]$, respectively. Under no linkage, the probabilities of affected sib pair sharing 0, 1, and 2 alleles IBD are 1/4, 1/2, and 1/4. This result can be used to form a nonparametric test for linkage. The corresponding expressions under multiplicative model can also be obtained, where probabilities of siblings sharing none or one allele by descent is $z_0 = 1/(4\lambda_S)$ and $z_1 = \lambda_O/(2\lambda_S)$, the nonshared probability is $Y = 1 - z_1/2 - z_0 = 1 - (\lambda_O + 1)/(4\lambda_S) = (1 + w)/(2 + w)$. When a candidate or dense collection of markers is available, one can use nuclear families in which affected offsprings and their parents are genotyped, the transmitted versus nontransmitted alleles from parents to offsprings are compared in the so-called transmission disequilibrium test (TDT), which can be based on nuclear families with affected singletons only.

In the following standard results related to normal distribution will be used. For a set of $N$ independent identically distributed random variables $B_i$ with mean and variance being 0 and 1 under the null hypothesis, $\mu$ and $\sigma^2$ under the alternative hypothesis, the statistic $\sum_{i=1}^{N} B_i/N$ has mean 0 and variance 1 under the null but mean $\sqrt{N}\mu$ and variance $\sigma^2$ under the alternative. The sample size $(N)$ for a given significance level $\alpha$ and power $1 - \beta$ can be estimated by $(Z_\alpha - \sigma Z_{1-\beta})^2/\mu^2$. In the actual calculation below, the type I error rate $(\alpha)$ and type II error rate $(\beta)$ are $5 \times 10^{-8}$ and 0.2, respectively.

### 3.2. Power/sample size for family and case-control/case-cohort designs

We recall some results on affected sib-pair linkage analysis, and define the alleles shared and nonshared from the $i$th parent as a random variable $B_i, i = 1, \ldots, N$, scoring 1 and $-1$. The mean $(\mu)$ and variance $(\sigma^2)$ of $B_i$ are 0 and 1 under the null hypothesis as the shared

and nonshared each has probability 0.5, and $2Y-1$ and $4Y(1-Y)$ under the alternative. Assuming sharing of alleles from both parent to be independent, the required sample size ($N$) for affected sib-pair under $\theta = 0$ and no linkage disequilibrium is $(Z_\alpha - \sigma Z_{1-\beta})^2/2\mu^2$, where $Y$ and $w$ are defined as before.

As with TDT, we assume that the disease locus and a nearby locus are in complete disequilibrium, the number of transmissions of allele A are scored from heterozygous parents, where the probability ($h$) of a parent of an affected child being heterozygous is given by $pq(\gamma+1)/(p\gamma+q)$ and $pq(\gamma+1)^2/[2(\gamma p+q)^2+pq(\gamma-1)^2]$ for singletons and sib-pairs, respectively. For singletons, a random variable $B_i$ is defined taking values $1/\sqrt{h}$ if parent is heterozygous and transmits A, 0 if parent is homozygous, $-1/\sqrt{h}$ if parent is heterozygous and transmits a. The mean and variance of $B_i$ are 0 and 1 under the null hypothesis, $\sqrt{h}(\gamma-1)/(\gamma+1)$ and $1-h[(\gamma-1)/(\gamma+1)]^2$ under the alternative. When sib-pairs instead of singletons are used in TDT analysis, the same formula for sample size calculation can be applied and the required number of families is half the expected number since there are two independent affected sibs.

The results of `example("fbsize", package = "gap")` are shown Table 3. Column $N_L$ contains the correct calculation corresponding to the original paper. The Alzheimer's disease model is based on Scott *et al.* (1997).

It turns out that with $\gamma \leq 2$, the expected marker-sharing only marginally exceeds 50% for any allele frequency ($p$). The use of linkage would need large sample size, but direct tests of association with a disease locus itself can still be quite strong although it may involve large amount of statistical testing of associated alleles. Clearly, it is most favorable for diseases that are relatively common, which has important implications for complex traits.

Now we consider the case-control design with a statistic directly testing association between marker and disease. Following Long *et al.* (1997) and supposing we have a randomly ascertained population sample, the frequencies of the three disease genotypes AA, Aa and aa in cases are $\pi\gamma^2p^2$, $2\pi\gamma pq$, and $\pi q^2$, respectively, where $\pi$ is the "baseline" probability that an individual with *aa* genotype being affected. Similarly, the three frequencies in controls are $(1-\pi\gamma^2)p^2$, $2(1-\pi\gamma)pq$ and $(1-\pi)q^2$. A unit $\chi^2$ statistic can be constructed using Table 4 as $X^2 = \sum(O-E)^2/E$, where $E$s indicate expected unit frequencies ($\pi p(\gamma p+q)^2$, $\pi q(\gamma p+q)^2$, $p-\pi p(\gamma p+q)^2$ and $q-\pi q(\gamma p+q)^2$), and the discrepancies between observed and expected frequencies all have factor $\pi pq(\gamma p+q)(\gamma-1)$ but with negative sign before the second and the third items. The statistic $X^2$ has $\chi^2_1$ distribution under the null hypothesis ($\gamma = 1$), and noncentral $\chi^2_{1,\delta}$ distribution with noncentrality parameter $\delta = [\pi pq(\gamma-1)^2]/[1-\pi(\gamma p+q)^2]$ or $[\gamma^2 p+q-(\gamma p+q)^2]/[1-\pi(\gamma p+q)^2]$ under the alternative ($\gamma > 1$). It is equivalent to derive the power by $Y = \sqrt{X^2} \sim N(\sqrt{\delta}, 1)$. $1-\beta = \Phi(-Z < Y < Z)$, where $Z$ is a preassigned standard normal deviate, $\Phi(.)$ is the cumulative normal distribution function.

The power for random case-control ascertainment with three different disease prevalences as from `example("pbsize", package = "gap")` is shown in Table 5.

So the most favorable scenario is when both the genotypic relative risk ($\gamma$) and allele frequency ($p$) are high for a given locus, and the disease is common.

Lastly we turn to case-cohort design. Cai and Zeng (2004) showed that power of a case-cohort design can be obtained by $\Phi(Z_\alpha + m^{0.5}\theta\sqrt{p_1p_2p_D/q+(1-q)p_D})$, where $\alpha$ is the significance level, $\theta$ is the log-hazard ratio for two groups, $p_j, j = 1,2$, are the proportion of the two groups in the population, $m$ is the total number of subjects in the subcohort, $p_D$ is the

| | | Linkage | | | Association | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\gamma$ | $p$ | $Y$ | $N_L$ | $P_A$ | $H_1$ | $N_{tdt}$ | $H_2$ | $N_{asp/tdt}$ | $\lambda_o$ | $\lambda_s$ |
| 4.00 | 0.01 | 0.520 | 6400 | 0.800 | 0.048 | 1098 | 0.112 | 235 | 1.08 | 1.09 |
| | 0.10 | 0.597 | 277 | 0.800 | 0.346 | 151 | 0.537 | 48 | 1.48 | 1.54 |
| | 0.50 | 0.576 | 445 | 0.800 | 0.500 | 104 | 0.424 | 62 | 1.36 | 1.39 |
| | 0.80 | 0.529 | 3023 | 0.800 | 0.235 | 223 | 0.163 | 162 | 1.12 | 1.13 |
| 2.00 | 0.01 | 0.502 | 445839 | 0.667 | 0.029 | 5824 | 0.043 | 1970 | 1.01 | 1.01 |
| | 0.10 | 0.518 | 8085 | 0.667 | 0.245 | 696 | 0.323 | 265 | 1.07 | 1.08 |
| | 0.50 | 0.526 | 3752 | 0.667 | 0.500 | 340 | 0.474 | 180 | 1.11 | 1.11 |
| | 0.80 | 0.512 | 17904 | 0.667 | 0.267 | 640 | 0.217 | 394 | 1.05 | 1.05 |
| 1.50 | 0.01 | 0.501 | 6942837 | 0.600 | 0.025 | 19321 | 0.031 | 7777 | 1.00 | 1.00 |
| | 0.10 | 0.505 | 101898 | 0.600 | 0.214 | 2219 | 0.253 | 941 | 1.02 | 1.02 |
| | 0.50 | 0.510 | 27041 | 0.600 | 0.500 | 950 | 0.490 | 485 | 1.04 | 1.04 |
| | 0.80 | 0.505 | 101898 | 0.600 | 0.286 | 1663 | 0.253 | 941 | 1.02 | 1.02 |
| Alzheimer's: | | | | | | | | | | |
| 4.50 | 0.15 | 0.626 | 163 | 0.818 | 0.460 | 100 | 0.621 | 37 | 1.67 | 1.78 |

Table 3: Comparison of linkage and association in nuclear families required for identification of disease gene: $\gamma$=genotypic risk ratio; $p$=frequency of disease allele A; $Y$=probability of allele sharing; $N_L$=number of ASP families required for linkage; $P_A$=probability of transmitting disease allele A; $H_1$, $H_2$=proportions of heterozygous parents; $N_{tdt}$=number of family trios; $N_{asp/tdt}$=number of ASP. families

proportion of the failures in the full cohort, and $q$ is the sampling fraction of the subcohort. Alternatively, the sample size required for the subcohort is $m = nBp_D/(n - B(1 - p_D))$, where $B = (Z_{1-\alpha} + Z_\beta)^2/(\theta^2 p_1 p_2 p_D))$ and $n$ is the size of cohort. The method has been implemented in function `ccsize`.

### 3.3. Joint analysis in staged design and related issues

Skol *et al.* (2006) examined tests of allele frequency differences between cases and controls in a two-stage design and is described here. The usual test of proportions can be written as $z(p_1, p_2, n_1, n_2, \pi_{samples})=(p_1-p_2)/\sqrt{p_1(1-p_1)/(2n_1\pi_{sample}) + p_2(1-p_2)/(2n_2\pi_{sample})}$, where $p_1$ and $p_2$ are the allele frequencies, $n_1$ and $n_2$ are the sample sizes, $\pi_{samples}$ is the proportion of samples to be genotyped at stage 1. The test statistics for stage 1, for stage 2 as replication and for stages 1 and 2 in a joint analysis are then $z_1 = z(\hat{p}_1, \hat{p}_2, n_1, n_2, \pi_{samples})$, $z_2 = z(\hat{p}_1, \hat{p}_2, n_1, n_2, 1 - \pi_{samples})$, $z_j = \sqrt{\pi_{samples}}z_1 + \sqrt{1 - \pi_{samples}}z_2$, respectively. Let $C_1$, $C_2$, and $C_j$ be the thresholds for these statistics, Skol *et al.* (2006) derived the false positive rates according to $P(|z_1| > C_1)P(|z_2| > C_2, sign(z_1) = sign(z_2))$ and $P(|z_1| > C_1)P(|z_j| >$

| | Affected genotype | Nonaffected genotype | |
|---|---|---|---|
| A | $\pi\gamma^2p^2 + \pi\gamma pq$ | $(1-\pi\gamma^2)p^2 + (1-\pi\gamma)pq$ | $p$ |
| a | $\pi\gamma pq + \pi q^2$ | $(1-\pi\gamma)pq + (1-\pi)q^2$ | $q$ |
| | $\pi(\gamma p + q)^2$ | $1 - \pi(\gamma p + q)^2$ | 1 |

Table 4: Expected frequencies for allele by case/control genotypes.

| | | $K$ | | |
|---|---|---|---|---|
| $\gamma$ | $p$ | 1% | 5% | 10% |
| 4.0 | 0.01 | 46638 | 8951 | 4240 |
| | 0.10 | 8173 | 1569 | 743 |
| | 0.50 | 10881 | 2089 | 990 |
| | 0.80 | 31444 | 6035 | 2859 |
| 2.0 | 0.01 | 403594 | 77458 | 36691 |
| | 0.10 | 52660 | 10107 | 4788 |
| | 0.50 | 35252 | 6766 | 3205 |
| | 0.80 | 79317 | 15223 | 7211 |
| 1.5 | 0.01 | 1598430 | 306770 | 145312 |
| | 0.10 | 191926 | 36835 | 17448 |
| | 0.50 | 97922 | 18793 | 8902 |
| | 0.80 | 191926 | 36835 | 17448 |

Table 5: Estimated sample sizes required for association detection using population data.

$C_j||z_1| > C_1$) for replication-based and joint analyses, respectively. As our primary interest is the power for the two types of analyses, we implement it in function `tscc` whose format is

```
tscc(model, GRR, p1, n1, n2, M, alpha.genome, pi.samples, pi.markers, K)
```

which requires specification of disease model (multiplicative, additive, dominant, recessive), genotypic relative risk (GRR), the estimated risk allele frequency in cases ($p_1$), total number of cases ($n_1$) total number of controls ($n_2$), total number of markers ($M$), the false positive rate at genome level ($\alpha_{genome}$), the proportion of markers to be selected ($\pi_{markers}$, also used as the false positive rate at stage 1) and the population prevalence ($K$). Note the disease risks involving the three genotypes for additive, dominant and recessive models are calculated as $(1, GRR, 2 \cdot GRR - 1)$, $(1, GRR, GRR)$, $(1, 1, GRR)$, respectively. Power for a number of scenarios for two-stage designs can be obtained with slight modification of example code in the documentation of `tscc`. Interestingly, the R implementation is much shorter than the C program by Skol *et al.* (2006).

Lin (2006) recently proposed a method of analysis for two-stage designs. Like other contributions, the focus was on the markers genotyped at both stages. Given that in most studies such as the case-cohort design outlined above, there is a rich collection of covariate information and

perhaps an equally important aspect is to use all markers and covariates at both stages in a unified analysis. This amounts to a standard analysis with missing independent variables with which statistical methods have been well developed, e.g., Ibrahim (1990), Schafer (1997a), van Buuren *et al.* (1999). The packages for multiple imputation available from CRAN, such as **cat**, **mix**, **norm**, **pan** (Schafer 1997b,c,d,e), **mice** (van Buuren and Oudshoorn 2007), **mitools** (Lumley 2006) can be used.

A useful point relates to prospective versus retrospective methods. The retrospective counterpart requires more attention but far from clear (Elston and Anne Spence 2006), although the equivalence of retrospective versus prospective methods is known (Prentice and Pyke 1979; Seaman and Richardson 2004). It has been shown (Kraft and Thomas 2000; Tan *et al.* 2005) that prospective likelihood can give biased estimate due to over representation of cases or the extremes of the traits compared with the general population, so that a retrospective model of allele/genotype/haplotype frequencies conditional on the disease phenotype via logistic model is preferred. The model is applicable to a wide range of traits and provides unbiased estimates. This is in contrast to prospective methods (Schaid *et al.* 2002; Seaman and Muller-Myhsok 2005).

These two aspects are well illustrated with a concrete example. We have recently conducted a study of Type-2 diabetes involving 5,013 Ashkenazi and four UK populations using two-staged design and 4,570 SNPs across a 10Mb region of chromosome 20q. A subset of 2,502 individuals were genotyped on these SNPs at stage one, from which SNPs with significance level 0.1 together and those within regions of interest (e.g., HNF4A) were further genotyped on remaining sample at stage two. In the analysis, a meta-analysis was applied to take into account the heterogeneity between populations. The two-stage method by Lin (2006) would be rather complicated. A useful prospective formulation for a particular marker not genotyped at stage two is to treat it as missing and considered in a weighted regression on trait, where the weight is associated with the three genotypes, similar to Schaid *et al.* (2002). For the retrospective method, one can augment the missing genotype similarly, in much the same way as Burkett *et al.* (2006). However, recall that this is exactly the kind of problem multiple imputation will deal with, such that we simply apply the appropriate packages aforementioned which accounts for the uncertainty due to missing genotypes. This has not been used in staged design although it was recognized recently in haplotype analysis (Sorensen *et al.* 2006; Cordell 2006). Now with our data contained in `chr20`, and stage one marker rs1419383 is to be analyzed, we can use **mice** as follows,

```
R> data("chr20")
R> imp <- mice(chr20)
R> fit <- lm.mids(cc ~ rs1419383 + ethnicity, data = imp)
R> pool(fit)
```

The function `mice` by default performs five imputations and therefore five analyses whose results are combined with `pool` command. We can also load Bioconductor (`http://www.bioconductor.org/`) package **RSNPper** (Carey 2006) and report metadata (e.g., genome annotation) and population information based on hapmap, as follows.

```
R> library("RSNPper")
R> mysnp <- SNPinfo("rs1419383")
R> popDetails(mysnp)
```

```
R> geneInfo("HNF4A")
```

Markers involved in a multistage design can largely be seen as monotonic missing by design and can be dealt with similarly. For large data from a typical genomewide association involving much more SNPs, we can use database connection mechanisms such as open database connectivity (ODBC).

### 3.4. The EPIC-Norfolk study of obesity

Our EPIC-Norfolk genomewide association study of obesity serves as a good example which many principles just outlined apply. The study was initially designed as a case-control study with the following definition of cases and controls: cases are those with body mass index (BMI) $> 30kg/m^2$ and controls with $20kg/m^2 \leq$ BMI $< 25kg/m^2$. This leads to all obese individuals of the EPIC-Norfolk cohort (3425), and 3400 controls. Half of the samples are to be genotyped at stage one, to be followed by the remaining half at stage two. When adapting this into a case-cohort design, we wished to know the power/sample size required and this turned to be straightforward to obtain. Part of calculation in Cai and Zeng (2004) and the required sample size or power in our case can be obtained with `example("ccsize", package = "gap")`. It turned out a comparable case-cohort design has approximately 2500 controls. Interestingly, this also corresponds to about 10% of the size of our cohort ($n$), which would expect to give stable estimate for a range of covariates measured in the cohort.

## 4. Summary

This short report shows that the design and implementation of the **gap** package allows for a wide range of functions available in a unified fashion. The package is still under development and can be seen as a prototype for a future more fully established environment. As a reviewer pointed out that "A strength of the '**gap**' package is that it aims to be the seed point for a general compilation of such methods, which means it implements many different kinds of methods, not just the ones developed by the author", the author would like to take this opportunity to invite comments, suggestions and contributions to consolidate the package. Because of the broad scope of applications, we focus on genomewide association as a particular application. Part of Table 3, Table 5, and the use of multiple imputation in staged design, have not been seen in the literature. The latter requires more elaborate development with respect to probability weighting. Function `tscc` is a much more transparent implementation than the original authors' C/C++ software. For power calculation, it is desirable to implement other commonly used models such as additive, dominant and recessive models. Recently, Skol *et al.* (2007) further examined the design issues associated with two-stage design. I am aware of related efforts which may be useful, such as **pbatR** (Hoffmann and Lange 2006), **SNPassoc** (Gonzalez *et al.* 2007), **GenABEL** (Aulchenko *et al.* 2007) and **snpMatrix** (Clayton and Leung 2007). Among others, the **RSNPper** (Carey 2006) is useful for accessing SNP metadata while **biomaRt** (Durinck 2006) enables a large collection of biological data to be available in R. The focus here is largely single point analysis, but multilocus methods such as haplotype analysis have been implemented in **gap**. For instance, the function `hap.score` is an adaptation of `haplo.score` function based on Schaid *et al.* (2002) which can easily take haplotype frequency estimates and haplotype assignments from other source and use them within the generalized linear models framework.

Although the availability of SNP data is overwhelming, the **gap** package was conceived such that it will handle not only SNPs but multiallelic loci and X-chromosome data. Examples include `genecounting` and `hap`, with the former being flexible enough to include features such as X-linked data and the latter 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. Owing to limitation in timing, we have used SAS for many tasks in analyzing our obesity data (Zhao *et al.* 2007) which partly contributes to the delay in consolidating SNP specific analyses in the package. Nevertheless, it will be of value to adapt some of the utilities developed in SAS.

After the package was posted to CRAN, other packages have appeared with somewhat overlapping functions, notably **powerpkg** (Weeks 2005); a recent version of **genetics** package (Warnes 2003) also contains a function `power.casectrl`. It would be useful to make some comparisons. Another point relates to the sequence of development of functions for a package. In retrospect, it would have been easier if all codes were implemented in native R language and avoiding foreign language calls, but the latter may be advantageous in speed given some tuning is done. Notably, the following functions need further work: `bt`, `kin.morgan`, `twinan90` (Williams *et al.* 1992), `gcontrol` (Devlin and Roeder 1999), `mtdt` (Sham 1997), `s2k` (Hirotsu *et al.* 2001). The overall target would involve better memory management, maintaining numerical precision or extensions, and use more native `.Call` and S4 classes of R. It is likely that existing functions and data would continue to evolve.

The major driving force to integrate many tools for genetic data analysis is in line with the observation that statistical genetic methods or genetic epidemiology is increasingly becoming part of the general epidemiology with focus on both genetic and nongenetic factors for diseases, where statistical and computational methods are more established. On this point, I wish to emphasize the benefit of shifting to or working on R from my own experience, which has been a very rewarding one, e.g., Zhao (2005) and Zhao (2006) in relation to package **kinship** (Therneau 2003). I noted a similar but more formal effort has been launched (`http://rgenetics.org/`) but certainly an informal approach also has its place, as for example in the case of **haplo.stats** (Sinnwell and Schaid 2007). The best data structures have yet to be established and collaborative work would be beneficial. 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).

## Acknowledgments

The R development team, particularly Prof. Kurt Hornik, gave many inputs during the package development. Dr. Anthony Long kindly provided me their results similar to Table 5 based on an approximation algorithm. Dr. Jian'an Luan provided the chromosome 20 data. Dr. Mike Weale kindly read through the manuscript and provided useful comments. Comments from anonymous referees and the editor Achim Zeileis also help to improve the manuscript.

## References

AT&T Research Lab (2008). ***Graphviz*** - *Graph Visualization Software*. URL `http://www.graphviz.org/`.

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007). "**GenABEL**: An R Library for Genome-Wide Association Analysis." *Bioinformatics*, **23**(10), 1294–1296.

Balding DJ (2006). "A Tutorial on Statistical Methods for Population Association Studies." *Nature Reviews Genetics*, **7**(10), 781–791.

Burkett K, Graham J, McNeney B (2006). "**hapassoc**: Software for Likelihood Inference of Trait Associations with SNP Haplotype and Other Attributes." *Journal of Statistical Software*, **16**(2), 1–19.

Cai J, Zeng D (2004). "Sample Size/Power Calculation for Case-Cohort Studies." *Biometrics*, **60**(4), 1015–1024.

Carey V (2006). "SNP Metadata Access and Use with Bioconductor." *R News*, **6**(5), 36–39. URL http://CRAN.R-project.org/doc/Rnews/.

Clayton D, Leung HT (2007). "An R Package for Analysis of Whole-Genome Association Studies." *Human Heredity*, **64**(1), 45–51.

Cordell HJ (2006). "Estimation and Testing of Genotype and Haplotype Effects in Case-Control Studies: Comparison of Weighted Regression and Multiple Imputation Procedures." *Genetic Epidemiology*, **30**(3), 259–275.

Devlin B, Roeder K (1999). "Genomic Control for Association Studies." *Biometrics*, **55**(4), 997–1004.

Durinck S (2006). "Integrating Biological Data Resources into R with **biomaRt**." *R News*, **6**(5), 40–45. URL http://CRAN.R-project.org/doc/Rnews/.

Elston RC, Anne Spence M (2006). "Advances in Statistical Human Genetics over the Last 25 Years." *Statistics in Medicine*, **25**(18), 3049–3080.

Gholami K, Thomas A (1994). "A Linear Time Algorithm for Calculation of Multiple Pairwise Kinship Coefficients and Genetic Index of Familiality." *Computers and Biomedical Research*, **27**, 342–350.

Gonzalez JR, Armengol L, Sole X, Guino E, Mercader JM, Estivill X, Moreno V (2007). "**SNPassoc**: An R Package to Perform Whole Genome Association Studies." *Bioinformatics*, **23**(5), 644–645.

Grant SF, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, Sainz J, Helgason A, Stefansson H, Emilsson V, Helgadottir A, Styrkarsdottir U, Magnusson KP, Walters GB, Palsdottir E, Jonsdottir T, Gudmundsdottir T, Gylfason A, Saemundsdottir J, Wilensky RL, Reilly MP, Rader DJ, Bagger Y, Christiansen C, Gudnason V, Sigurdsson G, Thorsteinsdottir U, Gulcher JR, Kong A, Stefansson K (2006). "Variant of Transcription Factor 7-like 2 (TCF7L2) Gene Confers Risk of Type 2 Diabetes." *Nature Genetics*, **38**(3), 320–323.

Guo SW, Lange K (2000). "Genetic Mapping of Complex Traits: Promises, Problems, and Prospects." *Theoretical Population Biology*, **57**, 1–11.

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

Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY, et al (1983). "A Polymorphic DNA Marker Genetically Linked to Huntington's Disease." *Nature*, **306**(5940), 234–238.

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC (1990). "Linkage of Early-onset Familial Breast Cancer to Chromosome 17q21." *Science*, **250**(4988), 1684–1689.

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.

Hoffmann T, Lange C (2006). "**P2BAT**: a Massive Parallel Implementation of **PBAT** for Genome-Wide Association Studies in R." *Bioinformatics*, **22**(24), 3103–3105.

Ibrahim JC (1990). "Incomplete Data in Generalized Linear Models." *Journal of the American Statistical Association*, **85**, 765–769.

Kerem B, Rommens J, Buchanan J, Markiewicz D, Cox T, Chakravarti A, Buchwald M, Tsui L (1989). "Identification of the Cystic-Fibrosis Gene: Genetic Analysis." *Science*, **245**(4922), 1073–1080.

Khoury MJ, Beaty TH, Cohn BH (1993). *Fundamentals of Genetic Epidemiology*. Oxford University Press.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005). "Complement Factor H Polymorphism in Age-related Macular Degeneration." *Science*, **308**(5720), 385–389.

Kraft P, Thomas DC (2000). "Bias and Eficiency in Family-based Gene-Characterization Studies: Conditional, Prospective, Retrospective, and Joint Likelihoods." *American Journal of Human Genetics*, **66**, 1119–1131.

Langholz B, Rothman N, Wacholder S, Thomas DC (1999). "Cohort Studies for Characterizing Measured Genes." *Journal of National Cancer Institute Monograph*, **26**, 39–42.

Lathrop G, Lalouel J, Julier C, Ott J (1984). "Strategies for Multilocus Linkage Analysis in Humans." *Proceedings of the National Academy of Science USA*, **81**, 3443–3446.

Lin DY (2006). "Evaluating Statistical Significance in Two-stage Genomewide Association Studies." *American Journal of Human Genetics*, **78**(3), 505–509. Author reply 1095-1096.

Long AD, Grote MN, Langley CH (1997). "Genetic Analysis of Complex Traits." *Science*, **275**, 1328.

Lumley T (2006). ***mitools**: Tools for Multiple Imputation of Missing Data*. URL http://CRAN.R-project.org/.

Manolio TA, Bailey-Wilson JE, Collins FS (2006). "Genes, Environment and the Value of Prospective Cohort Studies." *Nature Reviews Genetics*, **7**(10), 812–820.

Morton NE, Rao DC, Lalouel JM (1983). *Methods in Genetic Epidemiolgy*. Karger.

Prentice RL, Pyke R (1979). "Logistic Disease Incidence Models and Case-Control Studies." *Biometrika*, **66**, 403–411.

Risch N, Merikangas K (1996). "The Future of Genetic Studies of Complex Human Diseases." *Science*, **273**, 1516–1517.

SAS Institute Inc (2004). ***SAS/STAT***® *9.1 User's Guide*. Cary, NC. URL http://www.sas.com/.

SAS Institute Inc (2005). ***SAS/Genetics***® *9.1.3 User's Guide*. Cary, NC. URL http://www.sas.com/.

Schafer JL (1997a). *Analysis of Incomplete Multivariate Data*. Monographs on Statistics and Applied Probability. Chapmand and Hall.

Schafer JL (1997b). ***cat***: *Analysis of Categorical-Variable Datasets with Missing Values*. URL http://www.stat.psu.edu/~jls/misoftwa.html.

Schafer JL (1997c). ***mix***: *Estimation/Multiple Imputation for Mixed Categorical and Continuous Data*. URL http://www.stat.psu.edu/~jls/misoftwa.html.

Schafer JL (1997d). ***norm***: *Analysis of Multivariate Normal Datasets with Missing Values*. URL http://www.stat.psu.edu/~jls/misoftwa.html.

Schafer JL (1997e). ***pan***: *Multiple Imputation for Multivariate Panel or Clustered Data*. URL http://www.stat.psu.edu/~jls/misoftwa.html.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002). "Score Tests for Association Between Traits and Haplotypes when Linkage Phase is Ambiguous." *American Journal of Human Genetics*, **70**(2), 425–434.

Scott WK, Pericak-Vance MA, Haines JL (1997). "Genetic Analysis of Complex Diseases." *Science*, **275**, 1327.

Seaman SR, Muller-Myhsok B (2005). "Rapid Simulation of $P$ Values for Product Methods and Multiple-Testing Adjustment in Association Studies." *American Journal of Human Genetics*, **76**(3), 399–408, author reply 514–515.

Seaman SR, Richardson S (2004). "Equivalence of Prospective and Retrospective Models in the Bayesian Analysis of Case-Control Studies." *Biometrika*, **91**(1), 15–25.

Sham PC (1997). "Transmission/Disequilibrium Tests for Multiallelic Loci." *American Journal of Human Genetics*, **61**, 774–778.

Sinnwell JP, Schaid DJ (2007). *haplo.stats: Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous*. URL http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm.

Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006). "Joint Analysis is More Efficient than Replication-based Analysis for Two-stage Genome-wide Association Studies." *Nature Genetics*, **38**(2), 209–213.

Skol AD, Scott LJ, Abecasis GR, Boehnke M (2007). "Optimal Designs for Two-Stage Genome-Wide Association Studies." *Genetic Epidemiology*, **31**, 776–788.

Sorensen TI, Boutin P, Taylor MA, Larsen LH, Verdich C, Petersen L, Holst C, Echwald SM, Dina C, Toubro S, Petersen M, Polak J, Clement K, Martinez JA, Langin D, Oppert JM, Stich V, Macdonald I, Arner P, Saris WH, Pedersen O, Astrup A, Froguel P (2006). "Genetic Polymorphisms and Weight Loss in Obesity: A Randomised Trial of Hypo-Energetic High- versus Low-Fat Diets." *PLoS Clinical Trials*, **1**(2), e12.

Suarez BK, Rice JP, Reich T (1978). "The Generalised Sib-Pair IBD Distribution: Its Use in the Detection of Linkage." *Annals of Human Genetics*, **42**, 87–94.

Tan Q, Christiansen L, Christensen K, Bathum L, Li S, Zhao JH, Kruse TA (2005). "Haplotype Association Analysis of Human Disease Traits Using Genotype Data of Unrelated Individuals." *Genetical Research*, **86**, 223–231.

Therneau T (2003). *On Mixed-Effects Cox Models, Sparse Matrices, and Modeling Data from Large Pedigrees*. URL http://www.mayo.edu/biostatistics.

Thomas DC (2004). *Statistical Methods in Genetic Epidemiology*. Oxford University Press, Oxford.

Tsui LC, Buchwald M, Barker D, Braman JC, Knowlton R, Schumm JW, Eiberg H, Mohr J, Kennedy D, Plavsic N, et al (1985). "Cystic Fibrosis Locus Defined by a Genetically Linked Polymorphic DNA Marker." *Science*, **230**(4729), 1054–1057.

van Buuren S, Boshuizen HC, Knook DL (1999). "Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis." *Statistics in Medicine*, **18**(6), 681–94.

van Buuren S, Oudshoorn CGM (2007). ***mice**: Multivariate Imputation by Chained Equations*. URL http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm.

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

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

Warnes GR (2003). "The **genetics** Package." *R News*, **3**(1), 9–13. URL http://CRAN.R-project.org/doc/Rnews/.

Weeks DE (2005). ***powerpkg**: Power Analyses for the Affected Sib Pair and the TDT Design*. URL http://CRAN.R-project.org/.

Whittemore AS, Halpern J (1994). "A Class of Tests for Linkage Using Affected Pedigree Members." *Biometrics*, **50**, 118–127.

Williams CJ, Christian JC, Norton JA (1992). "**TWINAN90**: A Fortran Program for Conducting ANOVA-based and Likelihood-based Analyses of Twin Data." *Computer Methods and Programs in Biomedicine*, **38**(2-3), 167–176.

Yu C, Zelterman D (2002). "Statistical Inference for Familial Disease Clusters." *Biometrics*, **58**(3), 481–491.

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." *Human Heredity*, **53**(2), 79–91.

Zhao JH (2004). "**2LD**. **GENECOUNTING** and **HAP**: Computer Programs for Linkage Disequilibrium Analysis." *Bioinformatics*, **20**, 1325–1326.

Zhao JH (2005). "Mixed-Effects Cox Models of Alcohol Dependence in Extended Families." *BMC Genetics*, **6**, S127.

Zhao JH (2006). "Drawing Pedigree Diagrams with R and **Graphviz**." *R News*, **6**(2), 38–41. URL http://CRAN.R-project.org/doc/Rnews/.

Zhao JH, Lissarrague S, Essioux L, Sham PC (2002). "**GENECOUNTING**: Haplotype Analysis with Missing Genotypes." *Bioinformatics*, **18**(12), 1694–1695.

Zhao JH, Luan JA, Tan Q, Loos R, Wareham N (2007). "Analysis of Large Genomic Data *in Silico*: The EPIC-Norfolk Study of Obesity." In DS Huang, L Heutte, M Loog (eds.), *Communications in Computer and Information Science (CCIS)*, volume 2, pp. 781–790. Springer-Verlag, Berlin Heidelberg.

Zhao JH, Tan Q (2006). "Genetic Dissection of Complex Traits *in Silico*: Approaches, Problems and Solutions." *Current Bioinformatics*, **1**(3), 359–369.

## Affiliation:

Jing Hua Zhao
MRC Epidemiology Unit
Institute of Metabolic Science
Addenbrooke's Hospital
Hills Road
Cambridge CB2 0QQ
United Kingdom
E-mail: jinghua.zhao@mrc-epid.cam.ac.uk

*Journal of Statistical Software* https://www.jstatsoft.org/
published by the Foundation for Open Access Statistics https://www.foastat.org/
December 2007, Volume 23, Issue 8 *Submitted:* 2007-01-03
doi:10.18637/jss.v000.i00 *Accepted:* 2007-06-08