A Genetic Analysis Package with R

Jing Hua Zhao

MRC Epidemiology Unit, University of Cambridge, Cambridge, UK
http://www.mrc-epid.cam.ac.uk/people/jing-hua-zhao

Contents

1 Introduction 1
2 Implementation 2
3 Independent programs 10
4 Demos 11
5 Examples 11
  5.1 Pedigree drawing .......................... 11
  5.2 Kinship calculation ......................... 11
  5.3 Study design ............................... 13
  5.4 Graphics examples ......................... 18
6 Polygenic modeling 23
7 Known bugs 23
8 Summary 23
9 Bibliographic note 24

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 (2006a) and Zhao and Tan (2006b) 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 work as 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.
* ANALYSIS *

AE3  AE model using nuclear family trios
bt   Bradley-Terry model for contingency table
ccsie  Power and sample size for case-cohort design
fbsize  Sample size for family-based linkage and association design
ge.cm  Gene counting for haplotype analysis
gcontrol  genomic control
gcontrol2  genomic control based on p values
gcp  Permutation tests using GENECOUNTING
genecounting  Gene counting for haplotype analysis
gif  Kinship coefficient and genetic index of familiality
grmMCMC  Mixed modeling with genetic relationship matrices
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.cc  A likelihood ratio test of population Hardy-Weinberg equilibrium
hwe.hardy  Hardy-Weinberg equilibrium test using MCMC
kin.morgan  kinship matrix for simple pedigree
LD22  LD statistics for two diallelic markers
LDkl  LD statistics for two multiallelic markers
masize  Sample size calculation for mediation analysis
mia  multiple imputation analysis for hap
mttdt  Transmission/disequilibrium test of a multiallelic marker
mttdt2  Transmission/disequilibrium test of a multiallelic marker
   by Bradley-Terry model
mvmeta  Multivariate meta-analysis based on generalized least squares
pbsize  Power for population-based association design
pbsize2  Power for case-control association design
pfc  Probability of familial clustering of disease
pfc.sim  Probability of familial clustering of disease
pgc  Preparing weight for GENECOUNTING
print.hap.score  Print a hap.score object
s2k  Statistics for 2 by K table
tscc  Power calculation for two-stage case-control design
* GRAPHICS *

- asplot: Regional association plot
- ESplot: Effect-size plot
- mhtplot: Manhattan plot
- mhtplot2: Manhattan plot with annotations
- pedtodot: Converting pedigree(s) to dot file(s)
- plot.hap.score: Plot haplotype frequencies versus haplotype score statistics
- qqfun: Quantile-comparison plots
- qqunif: Q-Q plot for uniformly distributed random variable

* DATASETS *

- PD: A study of Parkinson’s disease and APOE, LRRK2, SNCA makers
- 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
- l51: An example pedigree data
- lukas: An example pedigree
- mao: A study of Parkinson’s disease and MAO gene
- meyer: A pedigree data on 282 animals deriving from two generations
- nep499: A study of Alzheimer’s disease with eight SNPs and APOE
* UTILITIES *

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP</td>
<td>Functions for single nucleotide polymorphisms (SNPs)</td>
</tr>
<tr>
<td>BFDP</td>
<td>Bayesian false-discovery probability</td>
</tr>
<tr>
<td>FPRP</td>
<td>False-positive report probability</td>
</tr>
<tr>
<td>ab</td>
<td>Test/Power calculation for mediating effect</td>
</tr>
<tr>
<td>b2r</td>
<td>Obtain correlation coefficients and their variance-covariances</td>
</tr>
<tr>
<td>chow.test</td>
<td>Chow’s test for heterogeneity in two regressions</td>
</tr>
<tr>
<td>comp.score</td>
<td>Score statistics for testing genetic linkage of quantitative trait</td>
</tr>
<tr>
<td>h2</td>
<td>Heritability estimation according to twin correlations for case-control studies</td>
</tr>
<tr>
<td>klem</td>
<td>Haplotype frequency estimation based on a genotype table of two multiallelic markers</td>
</tr>
<tr>
<td>makeped</td>
<td>A function to prepare pedigrees in post-MAKEPED format</td>
</tr>
<tr>
<td>metap</td>
<td>Meta-analysis of p values</td>
</tr>
<tr>
<td>metareg</td>
<td>Fixed and random effects model for meta-analysis</td>
</tr>
<tr>
<td>muvar</td>
<td>Means and variances under 1- and 2- locus (di allelic) QTL model</td>
</tr>
<tr>
<td>read.ms.output</td>
<td>A utility function to read ms output</td>
</tr>
<tr>
<td>twinan90</td>
<td>Classic twin models</td>
</tr>
<tr>
<td>whscore</td>
<td>Whittemore-Halpern scores for allele-sharing</td>
</tr>
<tr>
<td>GRM functions</td>
<td>ReadGRM, ReadGRMBin, ReadGRMPLINK, ReadGRMPCA, WriteGRM, WebGRMPCA,</td>
</tr>
<tr>
<td></td>
<td>handle genomic relationship matrix involving other software</td>
</tr>
<tr>
<td>heritability functions</td>
<td>h2G, VR, h2GC, h2l give point estimates as with their variances</td>
</tr>
</tbody>
</table>

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()`. Assuming that you have already loaded the package via `library(gap)`, you can use `lsf.str("package:gap")` and `data(package="gap")` to generate a list of functions and a list of datasets, respectively. If this looks odd to you, you might try `search()` within R to examine what is available in your environment before issuing the `lsf.str` command.

```r
> library(gap)
> search()

[1] ".GlobalEnv"   "package:gap"   "package:stats"
[10] "package:base"

> lsf.str("package:gap")
```
fbsize : function (gamma, p, alpha = c(1e-04, 1e-08, 1e-08), beta = 0.2, debug = 0, 
    error = 0)
g2a : function (g)
g2a.c : function (g)
gc.control : function (xdata = FALSE, convll = 1, handle.miss = 0, eps = 1e-06, tol =
    maxit = 50, pl = 0.001, assignment = "assign.dat", verbose = T)
gc.em : function (data, locus.label = NA, converge.eps = 1e-06, maxiter = 500,
    handle.miss = 0, miss.val = 0, control = gc.control())
gcode : function (a1, a2)
gcontrol : function (data, zeta = 1000, kappa = 4, tau2 = 1, epsilon = 0.01, ngib =
    burn = 50, idum = 2348)
gcontrol2 : function (p, col = palette()[4], lcol = palette()[2], ...
    quietly = FALSE)
geneaccounting : function (data, weight = NULL, loci = NULL, control = gc.control())
genoc.recode : function (geno, miss.val = 0)
getPTE : function (b1, b2, rho, sdx1 = 1, sdx2 = 1)
getb1star : function (b1, b2, rho, sdx1 = 1, sdx2 = 1)
gif : function (data, gifset)
grec2g : function (h, n, t)
h2 : function (mzDat = NULL, dzDat = NULL, rmz = NULL, rdz = NULL, selV = NULL)
h2G : function (V, VCOV, verbose = TRUE)
h2GE : function (V, VCOV, verbose = TRUE)
h2l : function (K = 0.05, P = 0.5, h2, se, verbose = TRUE)
 hap : function (id, data, nloci, loci = rep(2, nloci), names = paste("loci",
    1:nloci, sep = ""), control = hap.control())
 hap.control : function (mb = 0, pr = 0, po = 0.001, to = 0.001, th = 1, maxit = 100,
    n = 0, ss = 0, rs = 0, rp = 0, ro = 0, rv = 0, sd = 0, mm = 0, mi = 0,
    mc = 50, ds = 0.1, de = 0, q = 0, hapfile = "hap.out", assignfile = "assign.out"
    hap.em : function (id, data, locus.label = NA, converge.eps = 1e-06, maxiter = 500,
    miss.val = 0)
 hap.score : function (y, geno, trait.type = "gaussian", offset = NA, x.adj = NA, skip.
    locus.label = NA, miss.val = 0, n.sim = 0, method = "gc", id = NA,
    handle.miss = 0, mlcni = NA, sexid = NA)
hmht.control : function (data = NULL, colors = NULL, yoffset = 0.25, cex = 1.5, boxe.
    htr : function (y, x, n.sim = 0)
hwe : function (data, data.type = "allele", yates.correct = FALSE, miss.val = 0)
hwe.cc : function (model, case, ctrl, k0, initial1, initial2)
hwe.hardy : function (a, alleles = 3, seed = 3000, sample = c(1000, 1000, 5000))
 invlogit : function (x = 0)
k : function (r, N, adjust = TRUE)
kin.morgan : function (ped, verbose = FALSE)
klem : function (obs, k = 2, l = 2)
logit : function (p = 0.5)
m2plem : function (a1, a2)
makeped : function (pifile = "pedfile.pre", pofile = "pedfile.ped", auto.select = 1,
    with.loop = 0, loop.file = NA, auto.proband = 1, proband.file = NA)
masize : function (model, opts, alpha = 0.025, gamma = 0.2)
metap : function (data, N, verbose = "Y", prefixp = "p", prefixn = "n")
metareg : function (data, N, verbose = "Y", prefixb = "b", prefixse = "se")
mht.control : function (type = "p", usepos = FALSE, logscale = TRUE, base = 10, cutoff = NULL,
    colors = NULL, labels = NULL, srt = 45, gap = NULL, cex = 0.4, yline = 3,
    xline = 3)
mhtplot : function (data, control = mht.control(), hcontrol = hmht.control(), ...)
mhtplot2 : function (data, control = mht.control(), hcontrol = hmht.control(), ...)
mia : function (hapfile = "hap.out", assfile = "assign.out", miafile = "mia.out",
    so = 0, ns = 0, mi = 0, allsnps = 0, sas = 0)
imicombine : function (est, std.err, confidence = 0.95)
mtdt : function (x, n.sim = 0)
mtdt2 : function (x, verbose = TRUE, n.sim = NULL, ...)
muvar : function (n.loci = 1, y1 = c(0, 1, 1), y12 = c(1, 1, 1, 1, 1, 0, 0, 0,
    0), p1 = 0.99, p2 = 0.9)
mvmeta : function (b, V)
pbsize : function (kp, gamma = 4.5, p = 0.15, alpha = 5e-08, beta = 0.2)
pbsize2 : function (N, fc = 0.5, alpha = 0.05, gamma = 4.5, p = 0.15, kp = 0.1, mode = "additive")
pedtodot : function (pedfile, makeped = FALSE, sink = TRUE, page = "B5", url = "http://www.mrc-epid.cam.ac.uk",
    height = 0.5, width = 0.75, rotate = 0, dir = "none")
pfc : function (famdata, enum = 0)
pfc.sim : function (famdata, n.sim = 1e+06, n.loop = 1)
pgc : function (data, handle.miss = 1, is.genotype = 0, with.id = 0)
plem2m : function (a)
plot.hap.score : function (x, ...)
print.hap.score : function (x, ...)
qqfun : function (x, distribution = "norm", ylab = deparse(substitute(x)), xlab = paste(distribution,
    "quantiles"), main = NULL, las = par("las"), envelope = 0.95, labels = FALSE,
    col = palette()[4], lcol = palette()[2], xlim = NULL, ylim = NULL,
    lwd = 1, pch = 1, bg = palette()[4], cex = 0.4, line = c("quartiles",
    "robust", "none"), ...)
qqunif : function (u, type = "unif", logscale = TRUE, base = 10, col = palette()[4],
    lcol = palette()[2], ci = FALSE, alpha = 0.05, ...)
read.ms.output : function (msout, is.file = TRUE, xpose = TRUE, verbose = TRUE, outfileonly = FALSE)
revhap : function (loci, hapid)
revhap.i : function (loci, hapid)
s2k : function (y1, y2)
se.exp : function (p, se.p)
se.invlogit : function (logit.p, se.logit)
snp.ES : function (beta, SE, N)
snp.HWE : function (g)
snp.PAR : function (RR, MAF, unit = 2)
solve_skol : function (rootfun, target, lo, hi, e)
toETDT : function (a)
tscc : function (model, GRR, p1, n1, n2, M, alpha.genome, pi.samples, pi.markers, K)
ungcode : function (g)
whscore : function (allele, type)
x2 : function (p1, p2, n1, n2)
z : function (p1, p2, n1, n2, r)

> data(package="gap")$results

<table>
<thead>
<tr>
<th>Package</th>
<th>LibPath</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,] &quot;gap&quot;</td>
<td>&quot;/tmp/RtmpKD4nRJ/Rinst2474a1f555c&quot;</td>
<td>&quot;PD&quot;</td>
</tr>
<tr>
<td>[2,] &quot;gap&quot;</td>
<td>&quot;/tmp/RtmpKD4nRJ/Rinst2474a1f555c&quot;</td>
<td>&quot;aldh2&quot;</td>
</tr>
<tr>
<td>[3,] &quot;gap&quot;</td>
<td>&quot;/tmp/RtmpKD4nRJ/Rinst2474a1f555c&quot;</td>
<td>&quot;apoeapoc&quot;</td>
</tr>
<tr>
<td>[4,] &quot;gap&quot;</td>
<td>&quot;/tmp/RtmpKD4nRJ/Rinst2474a1f555c&quot;</td>
<td>&quot;cf&quot;</td>
</tr>
<tr>
<td>[5,] &quot;gap&quot;</td>
<td>&quot;/tmp/RtmpKD4nRJ/Rinst2474a1f555c&quot;</td>
<td>&quot;crohn&quot;</td>
</tr>
<tr>
<td>[6,] &quot;gap&quot;</td>
<td>&quot;/tmp/RtmpKD4nRJ/Rinst2474a1f555c&quot;</td>
<td>&quot;fa&quot;</td>
</tr>
<tr>
<td>[7,] &quot;gap&quot;</td>
<td>&quot;/tmp/RtmpKD4nRJ/Rinst2474a1f555c&quot;</td>
<td>&quot;fsnps&quot;</td>
</tr>
<tr>
<td>[8,] &quot;gap&quot;</td>
<td>&quot;/tmp/RtmpKD4nRJ/Rinst2474a1f555c&quot;</td>
<td>&quot;hla&quot;</td>
</tr>
<tr>
<td>[9,] &quot;gap&quot;</td>
<td>&quot;/tmp/RtmpKD4nRJ/Rinst2474a1f555c&quot;</td>
<td>&quot;151&quot;</td>
</tr>
<tr>
<td>[10,] &quot;gap&quot;</td>
<td>&quot;/tmp/RtmpKD4nRJ/Rinst2474a1f555c&quot;</td>
<td>&quot;lukas&quot;</td>
</tr>
<tr>
<td>[11,] &quot;gap&quot;</td>
<td>&quot;/tmp/RtmpKD4nRJ/Rinst2474a1f555c&quot;</td>
<td>&quot;mao&quot;</td>
</tr>
<tr>
<td>[12,] &quot;gap&quot;</td>
<td>&quot;/tmp/RtmpKD4nRJ/Rinst2474a1f555c&quot;</td>
<td>&quot;meyer&quot;</td>
</tr>
<tr>
<td>[13,] &quot;gap&quot;</td>
<td>&quot;/tmp/RtmpKD4nRJ/Rinst2474a1f555c&quot;</td>
<td>&quot;mfblong&quot;</td>
</tr>
<tr>
<td>[14,] &quot;gap&quot;</td>
<td>&quot;/tmp/RtmpKD4nRJ/Rinst2474a1f555c&quot;</td>
<td>&quot;nep499&quot;</td>
</tr>
</tbody>
</table>

Title

| [1,] "A study of Parkinson's disease and APOE, LRRK2, SNCA makers" |
| [2,] "ALDH2 markers and Alcoholism" |
| [3,] "APOE/APOC1 markers and Alzheimer's" |
| [4,] "Cystic fibrosis data" |
| [5,] "Crohn's disease data" |
| [6,] "Friedreich Ataxia data" |
| [7,] "A case-control data involving four SNPs with missing genotype" |
| [8,] "The HLA data" |
| [9,] "An example pedigree data" |
| [10,] "An example pedigree" |
| [11,] "A study of Parkinson's disease and MAO gene" |
| [12,] "A pedigree data on 282 animals deriving from two generations" |
| [13,] "Example data for ACEnucfam" |
| [14,] "A study of Alzheimer's disease with eight SNPs and APOE" |
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 multialelic 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 the `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 multialelic markers. Some are utility functions, e.g. `muvar` and `whscore`, which will be part of the other analysis routines in the future.

The benefit with R compared to standalone programs is that 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 Independent programs

To facilitate comparisons and individual preferences, The source codes for 2LD, EHPLUS, GENECOUNTING, HAP, now hosted at GitHub, have enjoyed great popularity ahead of the genomewide association studies (GWAS) therefore are likely to be more familiar than their R counterparts in `gap`. However, you need to follow their instructions to compile for a particular computer system.

I have included ms code (which is required by `read.ms.output`) and .xls files to accompany `read.ms.output` and `FPRP` and `BFDP` functions as with
a classic twin example for ACE model in OpenMx. The package is now available from CRAN.
For these models it is actually simpler to use facilities as in package mets, which I now suggest.

A final category is twinan90, which is now dropped from the package function list due to difficulty to keep up with the requirements by the R environment but nevertheless you will still be able to compile and use otherwise.

4 Demos
You can also try several simple examples via demo:

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

5 Examples
I would like to highlight pedtodot pbsize, fbsize and ccsize functions used for pedigree drawing and power/sample calculations in a genome-wide association study as reported in Zhao (2007).

5.1 Pedigree drawing
I have included the original file for the R News as well as put examples in separate vignettes. They can be accessed via vignette("rnews",package="gap") and vignette("pedtodot", package="gap"), respectively.

5.2 Kinship calculation
Next, I will provide an example for kinship calculation using kin.morgan. It is recommended that individuals in a pedigree are ordered so that parents always precede their children. In this regard, package pedigree can be used, and package kinship2 can be used to produce pedigree diagram as with kinship matrix.

Pedigree diagram
```
> # pedigree diagram
> data(lukas,package="gap")
> library(kinship2)
> ped <- with(lukas,pedigree(id,father,mother,sex))
```
null device

1

The pedigree diagram is as follows,

Kinship calculation

We then turn to the kinship calculation.

> # unordered individuals
> library(gap)
> gk1 <- kin.morgan(lukas)
> write.table(gk1$kin.matrix,"results/gap_1.txt",quote=FALSE)
> library(kinship2)
> kk1 <- kinship(lukas[,1],lukas[,2],lukas[,3])
> write.table(kk1,"results/kinship_1.txt",quote=FALSE)
> d <- gk1$kin.matrix-kk1
> sum(abs(d))

[1] 2.443634

> # order individuals so that parents precede their children
> library(pedigree)
> op <- orderPed(lukas)
> olukas <- lukas[order(op),]
> gk2 <- kin.morgan(olukas)
> write.table(olukas,"olukas.csv",quote=FALSE)
> write.table(gk2$kin.matrix,"results/gap_2.txt",quote=FALSE)
> kk2 <- kinship(olukas[,1],olukas[,2],olukas[,3])
> write.table(kk2,"results/kinship_2.txt",quote=FALSE)
> z <- gk2$kin.matrix-kk2
> sum(abs(z))

[1] 0

We see that in the second case, the result agrees with kinship2.

5.3 Study design

Family-based design

The example involving family-based design is as follows,

> library(gap)
> models <- matrix(c(
+   4.0, 0.01,
+   4.0, 0.10,
+   4.0, 0.50,
+   4.0, 0.80,
+   2.0, 0.01,
+   2.0, 0.10,
+   2.0, 0.50,
+   2.0, 0.80,
+   1.5, 0.01,
+   1.5, 0.10,
+   1.5, 0.50,
+   1.5, 0.80), ncol=2, byrow=TRUE)
```r
> outfile <- "fbsize.txt"
> cat("gamma","p","Y","N_asp","P_A","H1","N_tdt","H2","N_asp/tdt","L_o","L_s\n", +    file=outfile,sep="\t")
> for(i in 1:12) {
+  g <- models[i,1]
+  p <- models[i,2]
+  z <- fbsize(g,p)
+  cat(z$gamma,z$p,z$y,z$n1,z$pA,z$h1,z$n2,z$h2,z$n3,z$lambdao,z$lambda, +      file=outfile,append=TRUE,sep="\t")
+  cat("\n",file=outfile,append=TRUE)
+ }
> table1 <- read.table(outfile,header=TRUE,sep="\t")
> nc <- c(4,7,9)
> table1[,nc] <- ceiling(table1[,nc])
> dc <- c(3,5,6,8,10,11)
> table1[,dc] <- round(table1[,dc],2)
> unlink(outfile)
> # APOE-4, Scott WK, Pericak-Vance, MA & Haines JL
> # Genetic analysis of complex diseases 1327
> g <- 4.5
> p <- 0.15
> cat("\nAlzheimer's:

")
> fbsize(g,p)

$gamma
[1] 4.5

$p
[1] 0.15

$y
[1] 0.6256916

$n1
[1] 162.6246

$pA
[1] 0.8181818

$h1
[1] 0.4598361
```

Alzheimer's:

```r
> fbsize(g,p)

$gamma
[1] 4.5

$p
[1] 0.15

$y
[1] 0.6256916

$n1
[1] 162.6246

$pA
[1] 0.8181818

$h1
[1] 0.4598361
```
Population-based design

The example involving population-based design is as follows,

```R
> library(gap)
> kp <- c(0.01,0.05,0.10,0.2)
> models <- matrix(c(
+ 4.0, 0.01,
+ 4.0, 0.10,
+ 4.0, 0.50,
+ 4.0, 0.80,
+ 2.0, 0.01,
+ 2.0, 0.10,
+ 1.5, 0.50,
+ 1.5, 0.80,
+ 1.5, 0.10,
+ 1.5, 0.01), nrow=12, ncol=6)
```

15
Case-cohort design

For case-cohort design, we obtain results for ARIC and EPIC studies.

```r
> library(gap)
> # ARIC study
> outfile <- "aric.txt"
> n <- 15792
> pD <- 0.03
> p1 <- 0.25
> alpha <- 0.05
```
> theta <- c(1.35,1.40,1.45)
> beta1 <- 0.8
> s_nb <- c(1463,722,468)
> cat("n","pD","p1","hr","q","power","ssize\n",file=outfile,sep="\t")
> for(i in 1:3)
+ {
+   q <- s_nb[i]/n
+   power <- ccsize(n,q,pD,p1,alpha,log(theta[i]))
+   ssize <- ccsize(n,q,pD,p1,alpha,log(theta[i]),beta1)
+   cat(n,"\t",pD,"\t",p1,"\t",hr,"\t",q,"\t",signif(power,3),"\t",ssize,"\n",file=outfile,append=TRUE)
+ }
> read.table(outfile,header=TRUE,sep="\t")

<table>
<thead>
<tr>
<th>n</th>
<th>pD</th>
<th>p1</th>
<th>hr</th>
<th>q</th>
<th>power</th>
<th>ssize</th>
</tr>
</thead>
<tbody>
<tr>
<td>15792</td>
<td>0.03</td>
<td>0.25</td>
<td>1.35</td>
<td>0.09264184</td>
<td>0.8</td>
<td>1463</td>
</tr>
<tr>
<td>15792</td>
<td>0.03</td>
<td>0.25</td>
<td>1.40</td>
<td>0.04571935</td>
<td>0.8</td>
<td>722</td>
</tr>
<tr>
<td>15792</td>
<td>0.03</td>
<td>0.25</td>
<td>1.45</td>
<td>0.02963526</td>
<td>0.8</td>
<td>468</td>
</tr>
</tbody>
</table>

> unlink(outfile)

> # EPIC study
> outfile <- "epic.txt"
> n <- 25000
> alpha <- 0.00000005
> power <- 0.8
> s_pD <- c(0.3,0.2,0.1,0.05)
> s_p1 <- seq(0.1,0.5,by=0.1)
> s_hr <- seq(1.1,1.4,by=0.1)
> cat("n","pD","p1","hr","alpha","ssize\n",file=outfile,sep="\t")
> # direct calculation
> for(pD in s_pD)
+ {
+   for(p1 in s_p1)
+   {
+     for(hr in s_hr)
+     {
+       ssize <- ccsize(n,q,pD,p1,alpha,log(hr),power)
+       if (ssize>0) cat(n,"\t",pD,"\t",p1,"\t",hr,"\t",alpha,"\t",ssize,"\n",file=outfile,append=TRUE)
+     }
+   }
+ }
> read.table(outfile,header=TRUE,sep="\t")
> unlink(outfile)

5.4 Graphics examples

I now include some figures from the documentation that may be of interest.

Genome-wide association

The following code is used to obtain a Q-Q plot via qqunif function,

> library(gap)
> pdf("figures/qqunif.pdf", height=10, width=10)
> u_obs <- runif(1000)
> r <- qqunif(u_obs,pch=21,bg="blue",bty="n")
> u_exp <- r$y
> hits <- u_exp >= 2.30103
> points(r$x[hits],u_exp[hits],pch=21,bg="green")
> dev.off()

null device

1

Based on a chicken genome scan data, the code below generates a Manhattan plot, demonstrating the use of gaps to separate chromosomes.

> library(gap)
> ord <- with(w4,order(chr,pos))
> w4 <- w4[ord,]
> pdf("figures/w4.pdf",height=9,width=10)
> oldpar <- par()
> par(cex=0.6)
> colors <- c(rep(c("blue","red"),15),"red")
> mhtplot(w4,control=mht.control(colors=colors,gap=1000),pch=19,srt=0)
> axis(2,cex.axis=2)
> suggestiveline <- -log10(3.60036E-05)
The code below obtains a Manhattan plot with gene annotation,

```r
> library(gap)
> png("figures/mhtplot.png", height=10, width=16, units="cm", res=300)
> data <- with(mhtdata, cbind(chr, pos, p))
> glist <- c("IRS1", "SPRY2", "FTO", "GRIK3", "SNED1", "HTR1A", "MARCH3", "WISP3", "PPP1R3B",
  "RP1L1", "FDFT1", "SLC39A14", "GFRA1", "MC4R")
> hdata <- subset(mhtdata, gene%in%glist)[c("chr", "pos", "p", "gene")]
> color <- rep(c("lightgray", "gray"), 11)
> glen <- length(glist)
> hcolor <- rep("red", glen)
> par(las=2, xpd=TRUE, cex.axis=1.8, cex=0.4)
> ops <- mht.control(colors=col, yline=1.5, xline=3)
> mhtplot(data, ops, hops, pch=19)
```
All these look familiar, so revised form of the function called `mhtplot2` was created for additional features such as centering the chromosome ticks, allowing for more sophisticated coloring schemes, using prespecified fonts, etc. Please refer to the function’s documentation example.

The code below obtains a regional association plot with the `asplot` function,

```r
> library(gap)
> pdf("figures/asplot.pdf",height=14,width=14)
> asplot(CDKNlocus, CDKNmap, CDKNgenes, best.pval=5.4e-8, sf=c(3,6))
> title("CDKN2A/CDKN2B Region")
> dev.off()
```
The function predates the currently popular locuszoom software but leaves the option open for generating such plots on the fly and locally.

**Effect size plot**

The code below obtains an effect size plot via the ESplot function.

```r
> library(gap)
> pdf("figures/ESplot.pdf",height=10,width=10)
> options(stringsAsFactors=FALSE)
> testdata <- data.frame(models=c("Basic model","Adjusted","Moderately adjusted", "Heavily adjusted","Other"),
+ OR = c(4.5,3.5,2.5,1.5,1),
+ SElogOR = c(0.2,0.1,0.5,0.5,0.2))
> ESplot(testdata,v=1)
```
6 Polygenic modeling

In line with the recent surge of interest in the polygenic models, a separate vignette is available through `vignette("h2", package="gap")` demonstrating aspect of the models on heritability.

7 Known bugs

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

8 Summary

I believe by now the package should have given you a flavour of initiatives I have made so far in relation to how the project was envisaged. More
importantly, it is clear that availability of the package will serve as a platform on which future work can be accumulated and collaboration can be built.

9 Bibliographic note

The main references are Chow (1960); Guo and Thompson (1992); Williams et al. (1992); Gholamic and Thomas (1994); Hartung et al. (2008); Risch and Merikangas (1996); Spielman and Ewens (1996); Risch and Merikangas (1997); Miller (1997); Sham (1997); Elston (1975); 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


