GENECOUNTING: Haplotype Analysis with Missing Genotypes

GENECOUNTING implements an EM algorithm for haplotype analysis of unrelated subjects (Zhao et al. 2002). It has been recently extended to handle data on X chromosome so both males and females can be used in a single analysis. The algorithm can handle individuals with missing genotype data. However, currently it is limited to missing with both alleles at these loci.

This distribution also contains a module GENECOUNTING/PREPARE for preparing input data to GENECOUNTING and a module GENECOUNTING/PERMUTE obtaining permutation tests for global association and significance of specific haplotypes using Freeman-Tukey and proportion tests.

1 Installation

GENECOUNTING program and associates can be obtained from the following URLs:

`http://www.mrc-epid.cam.ac.uk/~jinghua/software.htm`

They are distributed in a zip file, so you will need unzip them first. e.g. by unzip

unzip gc22.zip

to a directory under both Unix and Windows. You can also use WINZIP (available from `http://www.winzip.com`). Full description of all files contained is available in packing.lst.

2 Input

Assuming there are n markers with alleles a1, a2, ..., an, the program accepts raw genotype data in the following format,

```
line 1: a1 a2 ... an
line 2- ID w <marker 1 genotype> ... <marker n genotype>
```

where w is a column of weights associated with specific multilocus genotypes. Marker genotype at each marker consists of two integers indicating numbered alleles. Nonpositive integers indicate missing genotypes.

For data involves X chromosome, an extra column is required to indicate sex (0=male, 1=female) of each individual (record), so that the format above becomes,

```
line 1: a1 a2 ... an
line 2- ID w <sex> <marker 1 genotype> ... <marker n genotype>
```

Note that marker genotype could either be "allele 1 allele 2" or "allele 1/allele 2" if two alleles are involved at a locus.

3 Output

GENECOUNTING reports number of individuals with genotypes at each marker, allele frequencies, haplotype frequency estimates assuming equilibrium and disequilibrium and their associated log-likelihoods. It also gives haplotype assignment and posterior probabilities for all subjects based on their observed multilocus genotypes.

4 Running GENECOUNTING

The command syntax is as follows,

gc <input file> [output file] [threshold]

where <input file> contains the raw genotypes and [output file] contains result of the analysis. [threshold] is a cutoff value for trimming posterior probabilities, with default value 0.001. Both output file and threshold are optional so they are put in []. Without specifying output file the output will be sent to computer screen, and can be redirected to a text file. e.g.

gc my-input-file > my-screen-output

The format with X chromosome version for gcx is similar.

5 Examples

A simulated data of four SNPs is contained in file 4snps.4m, whose first few lines are shown below,

```
2  2  2  2
   1    29  1  1  1  1  1  1  1  1
   2     9  1  1  1  1  1  1  1  2
   3     2  1  1  1  1  1  1  2  2
   4     4  1  1  1  1  1  1  0  0
   5    67  1  1  1  1  1  2  1  1
   ...
```

where Line 1 contains 4 integers representing numbers of alleles at four SNPs Line 2- contain the following columns: Column 1, sequence of numbers associated with 78 observed multilocus genotypes; Column 2, number of subjects with these genotypes; Columns 3-10, the actual genotypes.

For example, line 2 shows there are 29 individuals homozygous at all four loci, line 3 shows there are 9 individuals homozygous at the first three SNPs.

The following command can be used to obtain haplotype frequency estimates:

gc 4snps.4m 4snps.out

The HLA data as described in Zhao et al. (2000), Zhao and Sham (2002) is provided as file hla.dat. The example is to illustrate haplotype reconstruction with both missing data and multiallelic markers.

An example file involving X chromosome data is mao.inp.

6 Utility programs

6.1 GENECOUNTING/PREPARE

This program is able to condense the raw genotype data into the form as required by GENECOUNTING and has the following command syntax,

pgc <parameter file> <data file> <output file>

where ¡parameter file¿ and ¡data file¿ are EHPLUS (Zhao et al. 2000, also available from the URL above) parameter and data files and collapsing individuals with similar information into single category and recording number of instances occurs. Always select option for marker-marker analysis to prepare input file for GENECOUNTING.

Briefly, the data file can be in the form of either alleles

[ID] [label] [1a] [1b] [2a] [2b] ...

or genotype identifiers

[ID] [label] [1] [2] ...

where [ID] and [label] are individual's ID and case-control status, and [1a], [1b], [2a], [2b] are pairs of numbered alleles at each marker separated by spaces. As a genotype identifier of a marker uniquely determines both alleles it is also accepted. For instance a SNP the genotype identifiers for marker genotypes 1/1, 1/2 and 2/2 are 1, 2, and 3, respectively. In general, let L and U ($L <= U$) be the two alleles at a marker, then the identifier is calculated as follows

$$L + U(U-1)/2$$

File 4snps.dat contains 207 cases and 225 controls genotyped at four SNPs and can be used to generate 4snps.4m. Some lines of the file are shown as follows.

```
  1 1      1 2     1 2     2 2     1 2
  2 1      2 2     2 2     2 2     2 2
  3 1      1 2     0 0     1 2     1 2
  4 1      1 1     1 1     1 1     1 1
  5 1      1 1     1 1     2 2     1 2
...

208 0      1 1     1 1     1 2     1 1
209 0      1 1     1 1     0 0     1 1
210 0      1 1     1 1     1 2     1 2
211 0      1 2     1 1     1 1     1 1
212 0      0 0     1 1     1 1     1 2
...
```

where

Column 1 is individual's ID

Column 2 is a label showing the individual to be case (=1) or control (=0)

Columns 3-10 are the actual genotypes

It is simply the raw genotype data plus columns 1 and 2 showing individual IDs and case/control labels.

The parameter file consists of six lines indicating basic information of the data file and the analysis to be performed. This makes it possible to generate data for analysis involving only subset of markers. A parameter file appropriate for file 4snps.dat is 4snps.par containing the following lines,

```
4 0 0 0     << number of loci, case/control label, label permutation, replicates
2 2 2 2     << number of alleles
0 0         << raw genotype data, suppress screen reports
1 1 1 1     << select all markers for analysis
0 0 0 0     << no permutation for all 4 SNPs
0.001 0.05 0.2 0.8 << a putative disease model
```

Use the following command

pgc 4snps.par 4snps.dat 4snps.4m

to obtain 4snps.4m

It is possible to make the program to use raw genotype directly, as shown by 4snp.inp, whose first few lines of are shown as follows,

```
2 2 2 2
  1   1    1 2      1 2      2 2      1 2
  2   1    2 2      2 2      2 2      2 2
  3   1    1 2      0 0      1 2      1 2
  4   1    1 1      1 1      1 1      1 1
  5   1    1 1      1 1      2 2      1 2
  ...
```

Note all individuals have weight 1 at column 2. Now the command to perform the analysis becomes,

gc 4snps.inp 4snps.out

While this is rather straightforward, it is more time-consuming for running the analysis. It is also important that individuals with no information are excluded from the data file.

6.2 GENECOUNTING/PERMUTE

This module has been available from version 1.3. It has the following command syntax,

gcp <**parameter file**> <**data file**> [<**output file**> [**random number** seed]]

where <parameter file> and <data file> are EHPLUS format files as before, while the optional <output file> specifies the name of the output file. If <output file> is omitted, then the output will be sent to the computer screen. The seed for pseudorandom number generator can be specified after <output file>; its default value is 3,000.

For case-control data, it generates a sequence of heterogeneity statistics by permuting case-control labels. For a marker-marker analysis, it tests for linkage disequilibrium of a set of markers or association between two marker blocks. More details can be found in Zhao et al. (2000), Zhao and Sham (2002, 2003).

For example, to obtain p value of marker-disease association based on 10,000 replicates of cases and controls in 4snps.dat, we can alter 4snps.par to be as follows,

```
4 1 0 10000  << ---- This has been changed
2 2 2 2
0 0
1 1 1 1
0 0 0 0
0.001 0.05 0.2 0.8
```

GENECOUNTING/PERMUTE can be run as follows,

gcp 4snps.par 4snps.dat 4snps.out

For both case-control and marker-marker analyses, the program also yields haplotype specific tests.

7 Other utilities

It would be useful to extract haplotype assignment to be used by other programs, which can be be done using the following **awk**. Suppose 4snps.awk has the following line

```
/\[1\]|\[2\]/{gsub(/\[|\]/,""); print;}
```

On a Linux/Unix system we can extract the haplotype assignment by

awk -f 4snps.awk 4snps.out > assign.dat

It can also be done similarly Under Windows.

8 How to cite

Please cite the following references if you use the program in a publication.

Zhao, J. H., Lissarrague, S., Essioux, L. and P. C. Sham (2002). GENECOUNTING: haplotype analysis with missing genotypes. Bioinformatics 18(12):1694-1695

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

9 Acknowledgement

Thanks to Dr Sebastien Lissarrague for providing the SNP data, and to Dr Andrew Pakstis for providing HAPLO result of 4snps.dat as comparison during the program development.

10 References

Zhao, J. H., Curtis, D. and Sham, P. C. (2000). Model-free analysis and permutation test for allelic associations. Hum Hered 50:133-139

Zhao, J. H. and P. C. Sham (2002). Faster allelic association analysis using unrelated subjects. Hum Hered 53(1):36-41

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

Zhao, J. H. (2007). gap: Genetic analysis package. J Stat Soft 23(8):1-18

11 Contact information

Jing Hua Zhao
MRC Epidemiology Unit
Institute of Metabolic Science
Box 285
Addenbrooke's Hospital, Hills Road
Cambridge CB2 0QQ
United Kingdom

Tel: +44 (0)1223 769165
email: jinghua.zhao@mrc-epid.cam.ac.uk
Web: `http://www.mrc-epid.cam.ac.uk/~jinghua.zhao/`

Date of last change 04-04-2008 by Jing Hua Zhao