Documentation for FASTEHPLUS

1 Program description

FASTEHPLUS performs model-free analysis and permutation test(s) of allelic association based on EH (Xie and Ott 1993) and EHPLUS (Zhao et al. 2000). It uses marker genotypes from a group of unrelated individuals or a group of cases and a group of controls and employs gene-counting algorithm to estimate haplotype frequencies and output asymptotic and permutation test statistics. As yet it does not handle missing genotypes. Please check for GENECOUNTING program on this site if you wish to use individuals with missing genotypes in your analysis.

Assume m loci are involved each with ai alleles, $i = 1, \ldots, m$, they can form $a_1 \times a_2 \times ... \times a_m$ haplotypes. Two hypotheses can be considered.

H_0: no association between markers;

H_1: there is marker-marker association.

Under H_0, haplotype frequencies can be obtained from product of their constituent allele frequencies, so that the number of model parameters are simply $N_0 = (a_1 - 1) + (a_2 - 1) + ...(a_m - 1)$. Under H_1, the haplotype frequencies are obtained from gene counting method, a particular form of EM algorithm. The number of parameters becomes $N_1 = a_1 \times a_2 \ldots, \times a_m - 1$. Denote the log-likelihoods under both assumptions as $\ln L_0$ and $\ln L_1$, the log-likelihood ratio test statistic $2(\ln L_1 - \ln L_0)$ asymptotically has χ^2 distribution with $N_1 - N_0$ degrees of freedom.

A block of markers in the data can be assumed to be associated but not the others. The marker block could either be within a chromosome segment or any particular combination of typed markers. Let log-likelihoods from block 1 be $\ln L_0'$, $\ln L_1'$ and those from block 2 be $\ln L_0''$, $\ln L_1''$, and their numbers of parameters as N_0', N_1', N_0'', N_1'', then log-likelihood ratio test of two block association can be specified as $2(\ln L_1 - \ln L_1' - \ln L_1'')$ with $(N_1 - N_1' - N_1'')$ degrees of freedom.

When a group of cases and a group of controls are involved, the program outputs heterogeneity statistic, defined as

$$T_5 = -2(\ln L_1[cases + controls] - \ln L_1[cases] - \ln L_1[controls])$$

with labels in the brackets indicating sources of data, and the statistic can be referred to a χ^2 distribution with N_1 degrees of freedom.

Since potentially there may be many haplotypes involved and asymptotic approximation is likely to be unreliable, the program also performs empirical inference via permutation tests. A large number of replicates is generated by randomly shuffling marker data or case-control labels. This should break up any hidden association between markers or between markers and the putative disease locus as considered by the original EH. Statistics from these replicates then constitute an empirical distribution. The location of the observed statistic in this distribution provides empirical evidence for allelic association. In practice this is achieved by calculating proportion of replicates that produce values of statistics at least as large as the observed.

When only marker data are involved, FASTEHPLUS considers three kinds of permutations: permuting every marker loci in the data, permuting block 1 while keeping the second block 2 intact, and reporting result of block 1 only after permutation. For case-control data, only the observed case-control labels need to be permuted.

Replicate statistics from permutation procedure can be used to measure linkage disequilibrium (Zhao et al. 1999). The permutation-based LD measure is denoted ξ, and its sample estimate is denoted as $\hat{\xi}$, and $\hat{\xi} = \sqrt{2f}((t-\mu)/\sigma)/N$, where t is the log-likelihood ratio test statistic from the observed data, f its degrees of freedom and N the number of individuals in the sample. The mean (μ) and variance (σ^2) of the likelihood ratio test statistic are based on its empirical distribution obtained by permutation. The sample variance of $\hat{\xi}$, $2(f + 2N\hat{\xi})/N^2$, can be used to construct confidence interval.

2 Input and output

FASTEHPLUS needs two input files, a data file containing individual's ID, affection status (group identity) with original genotyping, and a parameter file describing these data.

2.1 Data file

The data file is a list of records containing individual's ID, affection status (0=unaffected, 1=affected) and marker genotypes either in the format of

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

where [ID] and [label] are the individual's ID and case-control status respectively. For case-control analysis (specified in parameter file) [label] takes values of 1 for cases and 0 for controls.

For the first format, columns [1a], [1b], [2a], [2b] are pairs of numbered alleles at each marker separated by spaces. For the second format columns [1], [2], etc. are genotype identifiers calculated from $(L+U(U-1)/2$, where L, U, $L \leq U$ are the actual alleles of a specific marker. For example with biallelic marker genotypes 1/1, 1/2 and 2/2 then the genotype identifiers are 1, 2 and 3.

2.2 Parameter file

The parameter file contains control information such as number of marker loci, type of analysis, number of permutation, and marker block, etc. specified in six lines.

line 1: #1, #2, #3, #4
line 2: alleles1, ..., alleles#1
line 3: *1, *2
line 4: selected1, ..., selected#1
line 5: permuted1, ..., permuted#1
line 6: q, f_0, f_1, f_2

Line 1 specifies four numbers: #1, number of loci (i.e., m), #2, type of analysis (0=marker-marker analysis, 1=case-control analysis), #3, label-permutating indicator for case-control analysis and #4, number of permutations. If "type of analysis" is specified to be 0 for a case-control data a marker-marker analysis will be performed for cases and controls combined together.

Line 2 specifies number of alleles for all loci (i.e., $a_1, \ldots, a_m$) as indicated in line 1.

Line 3 specifies whether the original marker data is single genotype identifier (*1=1) or actual alleles (*1=0), and whether to output the identifier to screen (0=no, 1=yes).

Line 4 specifies marker selection status for each locus in the analysis, i.e., to be used in the analysis if its value is 1, not used if its value is 0.

Line 5 specifies marker permutation status: those taking values of 1 formed one block and to be permuted; those taking values of 0 formed the other. This option is only for marker-marker analysis.

Line 6 specifies the disease model, only kept to be compatible with EHPLUS, i.e., specified but not used. These are frequency of disease allele, q, and three penetrances f_0, f_1, f_2 indicating probabilities of being affected given there are 0, 1, 2 disease alleles at the putative disease locus. Note in the original EH penetrances for disease genotypes are prompted as +/+, D/+ and D/D, where + and D represent normal and disease alleles at a putative disease locus, respectively.

2.3 Output

General information about data file and program control is reiterated on the computer screen. When only one analysis is requested all-subset analyses will be conducted.

For marker-marker analysis, one block association is just as if an ordinary statistic obtained from an EH-type analysis. The other two statistics are as described above.

For case-control data, statistics are given for cases only, controls only, and heterogeneity statistics for all subsets of markers. The main interests are the heterogeneity statistic, especially when only one marker is involved in a subset.

When replicate analysis is indicated, empirical p values, $\hat{\xi}$ (xihat) and its standard error are also calculated. However this estimate for heterogeneity statistic of case-control data is not so straightforward compared to that of marker-marker analysis.

3 Running the program

MicroSoft Windows users will have to enter MS-DOS Prompt first. For example with WIN9x the operations will be

Click Start → Select Programs → Select MS-DOS Prompt.

Then change to the directory where FASTEHPLUS locates by **cd** command (use **cd /?** to obtain more information).

The parameter, data and output files are supplied as command-line arguments to FASTEHPLUS. Optionally the program also reads a random number seed in place of the default value 3000 for permutation test(s). In other words, the syntax of command is as follows.

fpmp <parameter file><data file><output file> [seed]

Those in angled brackets (<>) are compulsory, i.e., they need explicitly specified, whereas the random number seed in squared bracket is optional, i.e., it may or may not be specified.

4 Example: Association of alcoholism and ALDH2

File aldh2.dat contains data of 130 Japanese alcoholics and 136 controls as reported in Koch et al. (2000). Six microsatellite markers and two single nucleotide polymorphisms (SNP) in the ALDH2 region (D12S2070, D12S839, D12S821, D12S1344, EXON12, EXON1, D12S2263 and D12S1341) were genotyped. They have alleles 8, 8, 13, 14, 2, 2, 13 and 10 in the sample.

Apart from the subject ID at column 1, case-control indicators at column 2, it simply a list of subjects with their marker genotypes at columns 3-18.

We now use data on two markers on either side of the functional locus EXON12. They are conveniently numbered as 1, 2, 3 and 4.

File aldh2cc.par is created with the following lines.

```
8 1 1 10000 << nloci, case/control, label permutation (1=permuted), npermute
8 8 13 14 2 2 13 10 << alleles
0 0 << is genotype, output genotypes (0=no, 1=yes)
0 0 1 1 0 1 1 0 << marker selection status (0=unselected, 1=selected)
0 0 0 0 0 0 0 0 << marker permutation status (0=not permuted, 1=permuted)
0.001 0.05 0.2 0.8 << disease model for case-control design
```

Line 1 has four numbers specifying aldh2.dat has 8 markers. This is a case-control analysis and permutation is conducted on affection status. The number of permutations to be performed is 10,000.

Line 2 lists the actual alleles at each marker mentioned above.

Line 3 indicates that the marker genotypes in aldh2.dat are provided as pairs of alleles and that the genotype identifiers will not be shown on the screen.

Line 4 selects markers D12S821, D12S1344, EXON1 and D12S2263 for the analysis.

Line 5 specifies markers are not to be permuted.

Line 6 defines the disease model. We can leave this line intact since we currently only use heterogeneity statistic for case-control data.

Now we specify

fpmp aldh2cc.par aldh2.dat aldh2cc.out

to obtain output in file aldh2cc.out.

The heterogeneity χ^2 statistic is 214.72 with degrees of freedom 4055 based on asymptotic approximation. We have to rely on permutation tests.

The same dataset can be used to perform marker-marker analysis. File aldh2mm.dat is obtained by slight modification of aldh2cc.par.

```
8 0 0 10000 << This line has been changed
8 8 13 14 2 2 13 10
0 0
0 0 1 1 0 1 1 0
0 0 1 1 0 0 0 0 << This line has been changed
0.001 0.05 0.2 0.8
```

Changes are made at the second number of line 1, which tells FASTEHPLUS to do marker-marker analysis. The fifth line has also been changed. Since this is a marker-marker analysis we use the option to see if any association between block 1, containing D12S821, D12S1344 and block 2, containing EXON1, D12S2263. We will use random number seed 50,000.

Our command now becomes

fpmp aldh2mm.par aldh2.dat aldh2mm.out 50000

The output will be written to aldh2mm.out.

As shown by these two analyses, we can modify our parameter file for any desirable subset analysis and permutation test. Multiple runs can be initiated via DOS or Unix batch files. Unless the original problem is small, this would be computer-intensive.

5 Notes on changes, program constants and compiling

There is one noticeable difference in marker-marker analysis between EHPLUS and FASTEHPLUS when blockwise association is examined: the third statistic in FASTEHPLUS is only for the permuted block. For case-control analysis, it suppresses statistics base on disease model. It is possible to modify FASTEHPLUS source code slightly to allow for case-control type analysis using parametric model. A single EH-type analysis could be achieved by PREPFASTEHPLUS and FASTEHPLUS. PREPFASTEHPLUS program uses input files for FASTEHPLUS to generate an output file, to be used by FASTEHPLUS for ordinary EH analysis.

For instance, to run a marker-marker analysis of the ALDH2 dataset we use the following command:

pfehp aldh2mm.par aldh2.dat aldh2mm.dat

fehp

and answer the queries as follows
Do you wish to use the case-control sampling option? [N]
Enter name of data file [EHPLUS.DAT]
aldh2mm.dat
you entered: aldh2mm.dat
Enter name of output file. [EHPLUS.OUT]
aldh2mm.out

The result is then stored in file aldh2mm.out.
A full list of the distributed files is given as follows.

File name	Description
fehp.c	FASTEHPLUS single analysis source file
fehp.h	FASTEHPLUS single analysis header file
fehp.exe	FASTEHPLUS single analysis executable file
fpmp.c	FASTPMPLUS, permutation/model-free analysis source file
fpmp.h	FASTPMPLUS, permutation/model-free analysis header file
fpmp.exe	FASTPMPLUS, permutation/model-free analysis executable file
fpmp.doc	Documentation in ASCII format
fpmp.pdf	This file
pfehp.c	PREPFASTEHPLUS source file
pfehp.h	PREPFASTEHPLUS header file
pfehp.exe	PREPFASTEHPLUS executable
aldh2.dat	ALDH2 data file
aldh2cc.par	ALDH2 parameter file for case-control analysis
aldh2mm.par	ALDH2 parameter file for marker-marker analysis

Program fpmp.c integrates the functionality of EHPLUS programs ehplus.c and pmplus.c. Program fehp.c is a faster version of ehplus.c for a single analysis without permutation, and allows for disease model of a putative disease locus to be specified (Xie and Ott 1993; Ott 1998). Program pfehp.c prepares input file for fehp.c.

Two program constants are maximum number of loci (MAX_LOC) and maximum number of alleles (maxalleles) at a locus, set to MAX_LOC=30, maxalleles=50. To alter them simply locate them in fpmp.h and change to the desired values.

Examples for building WIN9x/NT or Unix executables are as follows.

Borland/Inprise C

bcc -Iinclude -Llib -mh fpmp.c

or

bcc32 -Iinclude -Llib fpmp.c
assuming current directory contains include and lib subdirectories for C header and library files.

Symantec C

sc -mn fpmp.c

assuming include and lib subdirectories are properly set.

MicroSoft Visual C

vcvars32
cl fpmp.c

Appropriate environments can be set via runing batch file VCVARS32.

Cygwin gcc

gcc fpmp.c -o fpmp

Unix gcc

gcc fpmp.c -lm -o fpmp

Under Unix, pfehp, fehp and fpmp can be created by a single command

make

6 Acknowledgement

Thanks to Dr Wentian Li for providing makefile and Dr Dimitri Zaykin for many suggestions. Thanks also to many colleagues for comments and program testing.

7 Contact information

If you have any questions, comments and suggestions, please contact me via **e-mail** j.zhao@public-health.ucl.ac.uk, or by post to

Jing Hua Zhao
Department of Epidemiology & Public Health
University College London
1-19 Torrington Place
London WC1E 6BT
The United Kingdom
Tel +44 (0)20 7679 5627

8 How to cite

Zhao JH and Sham PC (2002) Faster allelic association analysis using unrelated subjects. Hum Hered, 53(1):36-41

9 References

Koch HG, McClay J, Loh E-W, Higuchi S, Zhao J-H, Sham P, Ball D, et al (2000) Allele association studies with SSR and SNP markers at known physical distances within a 1 Mb region embracing the ALDH2 locus in the Japanese, demonstrates linkage disequilibrium extending up to 400 kb. Hum Mol Genet 9:2993-2999

Ott J. (1998) User's Guide to EH. http://linkage.rockefeller.edu

Xie X. and J. Ott (1993): Testing linkage disequilibrium between a disease gene and marker loci. Am J Hum Genet 53:1107

Zhao JH, Curtis D, Sham PC (2000) Model-free analysis and permutation tests for allelic associations. Hum Hered 50:133-139

Zhao H, Pakstis AJ, Kidd JR, Kidd KK (1999) Assessing linkage disequilibrium in a complex genetic system. I. Overall deviation from random association. Ann Hum Genet 63:167-179