Haplotype Frequency Estimation and Association Analysis with FAMHAP

Tim Becker, Christine Herold, Michael Knapp
Institute for Medical Biometry, Informatics and Epidemiology
Sigmund-Freud-Str. 25
D-53105 Bonn
E-mail: Tim.Becker@ukb.uni-bonn.de

Table of Contents

1. About FAMHAP

FAMHAP is a software for **single-marker analysis** and, in particular, **haplotype analysis**. It can be used both for the analysis of **case-control data** and the analysis of **nuclear family data** (parents with an arbitrary number of children). The program is optimized for the **haplotype frequency estimation of many markers**. A theoretically possible number of 2^{62} haplotypes is allowed, i.e., for instance, 62 SNPs can be handled. (On Dos32 machines the limit is 30). FAMHAP provides **LD measures** and selects **tagging markers**. The program also implements a method that **search**es **for potential genotyping errors** using haplotypes. FAMHAP provides several methods for **haplotype** and **diplotype association analysis**. Particular emphasis is on **Monte-Carlo simulations** and the issue of **multiple testing**. For case-control data, FAMHAP also provides testing strategies for the **simultaneous analysis of unlinked genomic regions**, **quantitative traits** and a method to test for an excess of seldom haplotypes in cases. For nuclear family data, a test for association and **imprinting** is available. For case-control data, **imputed genotypes** can be tested for association. Methods that allow **conditional analysis** accounting for LD are provided, as well.

The new release comes along with a **graphical user interface** (**GUI**) for Windows users. On top of this, there is a new feature that makes it possible to run FAMHAP commands repeatedly for varying marker sets. This feature greatly facilitates **association analysis of GWAS**.

The haplotype frequencies the program computes are maximum-likelihood estimates (MLEs) which are obtained with the expectation-maximization (EM) algorithm. It is the philosophy of the program to store haplotype explanations and to not recalculate them. In this way, it is possible to keep the EM-algorithm fast also for many markers. In order to allow quick testing procedures based on MC simulations, haplotype explanations are stored together with their conditional likelihood weights which can be derived from the haplotype frequency estimates.

FAMHAP is written in C. The GUI is written in C#.

2. How to Get Started

The current version **famhap18** (2008) can be downloaded from http://famhap.meb.uni-bonn.de/.

Windows: FAMHAP can now be called with the graphical user interface. The GUI will work with Windows XP® and Windows Vista®, provided that Microsoft .NET (version 2 upwards) is installed. This program should be installed on all computers with Windows XP Service Pack 2 upwards and all Windows Vista® computers.. If you do not have the programme, please download and install it from http://www.microsoft.com/net ->Download .Net Framework. For usage with other Windows versions please contact us.

To run the GUI version, unzip the package **famhap18.zip** and click on **Famhap.exe** in the folder famhapGUI to run the program! You will also find the DOS executable famhap18dos32.exe in this folder. It can also be run from the command prompt, but do not remove it from the folder! There is a dos32 version and a dos64 version for 64-Bit machines.

Linux: A unix executable **famhap18Linux** is provided which can be run from the command line (cf section 4). Mind to change the user rights with **chmod +x famhap18LINUX**! If the executable does not work on your machine, you obtain a running version of the program by compiling the C source code **famhap18.c** with **gcc**. You will find the source code file in the folder **SOURCE_CODE**. If you are in that directory, you can compile the program with the command

```
gcc famhap18.c  -o famhap18Linux -lm -O3
```

to get an executable called **famhap18**. The “-O3” statement is optional, but it improves the running time of the program. I strongly recommend to use it. The program has been developed and tested with this option.

It also should be possible to run the GUI on a unix machine. The mono program has to be installed on your machine and you can start the GUI with the command

```
mono Famhap.exe
```

Depending on your system, it might be necessary to compile the program first as described before. In this case copy the new compile famhap18Linux into the folder were Famhap.exe is located.

Mac: Should work as unix. Use the mono program to run the GUI!

3. Input Format and Example Files

FAMHAP uses a **simplified LINKAGE format (pedfile)**. Have a look at the example files **casecontrol** and **nuclearfamilies**.

The rows of the pedfile are:

FID PID FA MO SEX AFF M1_a M1_b M2_a M2_b ……

The columns must be separated by blanks or tabs.

The first line of the infile must not contain leading blanks/tabs or additional blanks/tabs at the end! Each line, in particular the last line, must end with a carriage return.

It is recommended that the infile contains a line of column headings. The first line is treated as a line of headings, if it begins with the string **FID** and is treated as a data row if it does not start with **FID**.

FID (family ID) and **PID** (person ID) are treated as strings, and must not contain blanks or tabs. FID and PID are not allowed to be zero.

Only unrelated individuals and nuclear families are allowed! Larger pedigrees are not supported, half-sibships are not allowed. Sibs without parents must be treated as nuclear families with missing genotypes in the parents, for the purpose of haplotype frequency estimation (of the parental generation). However, **we do not recommend to use FAMHAP to test for association using sibships without parental genotype data.**

The program assumes that the file contains only individuals or nuclear families. The columns **MO** (mother) and **FA** (father) may be integer or string variables. They are used less stringent than in the original linkage format: if MO and FA are both strings different from “0”, the person is treated as a child of the nuclear family, if both MO and FA are equal to “0”, the person is treated as a parent of the nuclear family or as an independent individual, if the family has just one member. Note that for case-control data, each person must have a unique FID and that MO and FA must be coded “0”.

AFF=1 stands for unaffected, AFF=2 for affected. SEX=1 stands for male, SEX=2 for female. 0 stands for unknown. Unknown affection status or sex are allowed, but with some program options such individuals/families are excluded from analysis. The affection status may be replaced by **quantitative traits**. In this case, missing data is codes as “-“ and the option “q” has to be used.

For each marker **two alleles** are required. Missing alleles are coded as 0. It is expected that either both alleles of a genotype are missing or that the genotype is complete.

Alleles are coded as integers (1,2, 244,247, …) or as basepairs (A,C,G,T, a,c,g ,t) or A/B.

The program determines the number of markers from the first line of the infile. If subsequent lines have more columns, the program stops! (Note that this happens, for instance, when there are blanks in your family IDs or person IDs). If there are incomplete rows, the program stops as well.

Family members are identified via the FID. They do not have to be in rows followed by each other.

Note: The name of your inputfile must not contain whitespaces!

4. The Command Line and General Options

If you are in the directory where the famhap18.exe is located, the general form of the command is

./famhap18 pedfile name2 <options> (UNIX) or
famhap18dos32 pedfile name2 <options> (DOS command prompt).

With this command the main outputfile with the name **name2.out** is produced. If you use **auto** as outputfilename (**famhap18 pedfile auto),** your outputfile is called **pedfile.out**. If you use the **GUI**, this is the default.

Analysis can be restricted to specific individuals/families and markers with the following options. The order of occurrence is irrelevant, unless explicitly specified.

male (female)

Only male (female) individuals are kept for association testing. Females (males) are treated as having unknown disease status, i.e., they are not used for association testing.

malecase (femalecase)

Only male (female) affected individuals /offspring are kept for association testing. Female (male) cases are treated as having unknown disease status, i.e., they are not used for association testing. Female (male) controls are kept. The options **malecase/femalecase** are more reasonable than **male/female** since the considered sample size is larger and since for complex diseases normal and healthy controls do not differ very much. Furthermore, the existence of real differences between allele frequencies in male controls and female controls remains to be shown.

u60,u70,u80,u90,u95,u99,u100

With **u60** only family units with a typing ratio of at least 60% (for the selected SNPs) are considered. Families with multiple offspring are kept if at least one subtrio (father, mother, one of the children) meets the typing criterion. The other options work analogously. Use **u100** to allow only fully genotyped family units.

P, PP

Enhanced/detailed output is produced.

2,3,37

Natural numbers can be used to select certain SNPs.

Example 1 (command line):

./famhap18 pedfile auto u60 2 3 37 PP

SNPs 2, 3 and 37 will be analyzed. Only individuals with a typing ratio of at least 60% are considered. Detailed output is produced and written to the main outputfile **pedfile.out**.

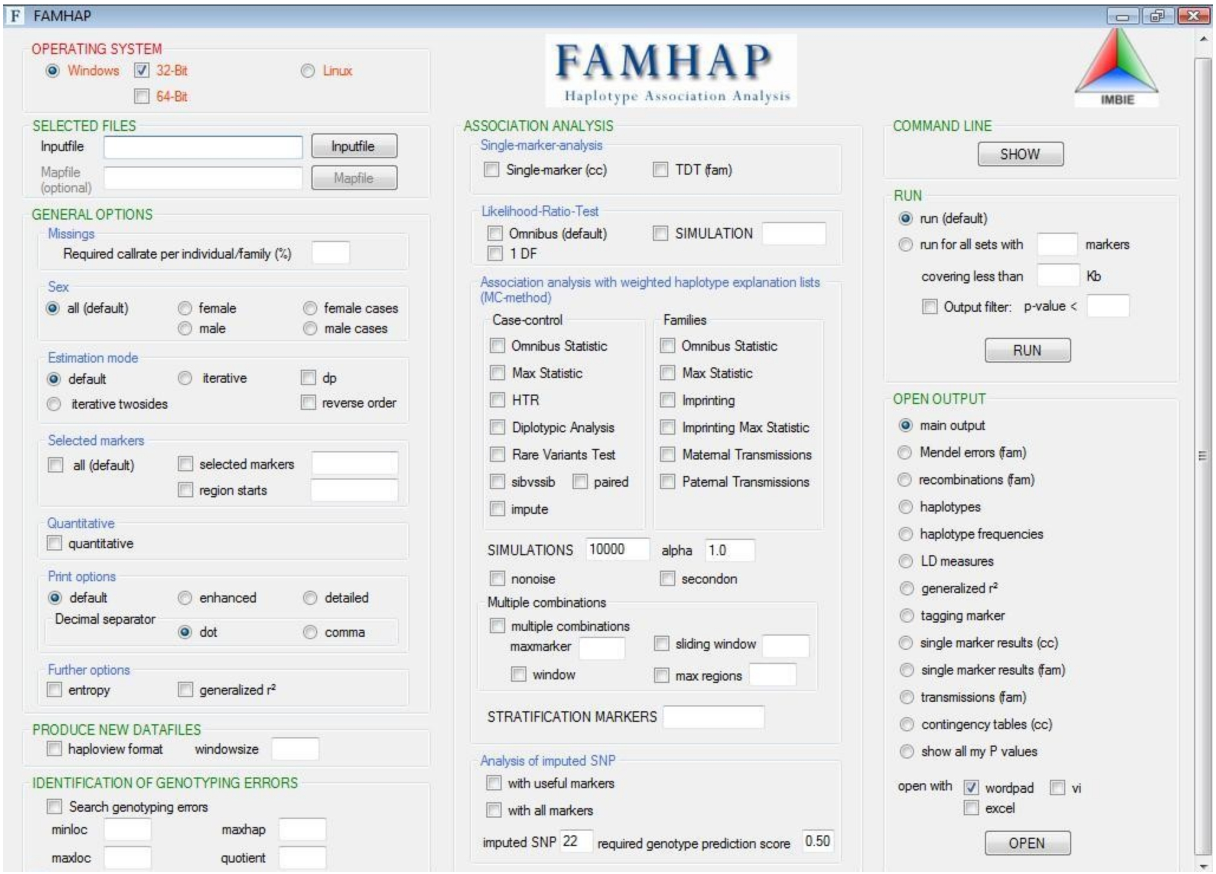
5. The GUI

FAMHAP can be operated with the graphical user interface **famhapGUI.exe**. On a unix machine, run the command **mono famhapGUI.exe**.

Data files and options are selected with the **GUI** and famhap18.exe is called with the **RUN** button.

Example 1 (section 4) can be conducted with the **GUI** as follows:

Select Files->Inputfile: pedfile
General Options->Missings: 60
General Options->Selected Markers: 2;3;37
General Options->Print Options->Detailed
RUN



F FAMHAP

FAMHAP
Haplotype Association Analysis
IMBIE

OPERATING SYSTEM
● Windows ☑ 32-Bit ○ Linux
□ 64-Bit

SELECTED FILES
Inputfile [Inputfile]
Mapfile (optional) [Mapfile]

GENERAL OPTIONS
Missings
Required callrate per individual/family (%)
Sex
● all (default) ○ female ○ female cases
○ male ○ male cases
Estimation mode
● default ○ iterative □ dp
○ iterative twosides □ reverse order
Selected markers
□ all (default) □ selected markers
□ region starts
Quantitative
□ quantitative
Print options
● default ○ enhanced ○ detailed
Decimal separator ● dot ○ comma
Further options
□ entropy □ generalized r²

PRODUCE NEW DATAFILES
□ haploview format windowsize

IDENTIFICATION OF GENOTYPING ERRORS
□ Search genotyping errors
minloc maxhap
maxloc quotient

ASSOCIATION ANALYSIS
Single-marker-analysis
□ Single-marker (cc) □ TDT (fam)
Likelihood-Ratio-Test
□ Omnibus (default) □ SIMULATION
□ 1 DF
Association analysis with weighted haplotype explanation lists (MC-method)
Case-control
□ Omnibus Statistic
□ Max Statistic
□ HTR
□ Diplotypic Analysis
□ Rare Variants Test
□ sibvssib □ paired
□ impute
Families
□ Omnibus Statistic
□ Max Statistic
□ Imprinting
□ Imprinting Max Statistic
□ Maternal Transmissions
□ Paternal Transmissions
SIMULATIONS 10000 alpha 1.0
□ nonoise □ secondon
Multiple combinations
□ multiple combinations
maxmarker □ sliding window
□ window □ max regions
STRATIFICATION MARKERS
Analysis of imputed SNP
□ with useful markers
□ with all markers
imputed SNP 22 required genotype prediction score 0.50

COMMAND LINE
[SHOW]

RUN
● run (default)
○ run for all sets with markers
covering less than Kb
□ Output filter: p-value <
[RUN]

OPEN OUTPUT
● main output
○ Mendel errors (fam)
○ recombinations (fam)
○ haplotypes
○ haplotype frequencies
○ LD measures
○ generalized r²
○ tagging marker
○ single marker results (cc)
○ single marker results (fam)
○ transmissions (fam)
○ contingency tables (cc)
○ show all my P values
open with ☑ wordpad □ vi
□ excel
[OPEN]

The GUI.

The file short_docu_famhap18.xls contains a table where you can find the GUI options together with the corresponding command line options!

Hint: With the **SHOW** button it is possible to check the command line**.**

6. Overview of Major Outputfiles

By default, all outputfiles are written to the folder where the inputfile lies.

infile.out
is the main outputfile.

***_MENDELERRORS.txt**
contains a table of Mendelian errors.

***_families_with_recos.txt**
contains IDs of families with at least one recombination.

***_FREQ.txt**
contains the haplotype frequency estimates.

***_H0_HAPLOTYPES.txt**
contains haplotypes per individual/family.

***_LDmeasuresALL.xt / *.gm**
contain LD measures and a map file for usage with the GOLD program (*Abecasis GR and COOKSON WOC. GOLD-Graphical Overview of Linkage Disequilibrium. Bioinformatics 2000 16:182-183***)**

***_TAG.txt**
indicates tagging markers. See section 8.1.

***_SINGLEMARKER.txt**
contains various single-marker P-values for case-control data (option **singlecc**). Odds ratios with confidence intervals are provided, as well.

***_SINGLEMARKER_FAM.txt**
contains various single-marker P-values for family data (option **tdt**). This option produces P-values for linkage when multiple affected children are present. For association testing use the option **haptdt** (section 12.1).

***_Contingency_Tables.txt**
contains contingency tables for case control data, see section 11.

***_TNT_Tables.txt**
contains transmission/nontransmission tables, see section 12.

***_Pvalues.txt**

stores all P values (from multiple runs), together with the chosen analysis options that were computed for a specific inputfile. Have a look!

Hint and warning: Most files are tab-separated! With the GUI, you can open the files with Microsoft Excel®, but cross-validate by looking at the files with wordpad!! Depending on your computer system, the automatic formatting of Excel may lead to errors!

Hint: With the option **comma**, the decimal separator in most outputfiles is set to be the comma. Note that this option may increase running time.

7. Single Marker Analysis

Depending on available computer storage, it is possible to conduct single-marker analysis of ten thousands of markers with a single run. With the following options all markers of the inputfile are analyzed.

Case-Control Data:

With the unix command

```
./famhap18 pedfile auto singlecc
```

single marker analysis of SNP case-control data is conducted. The results of the single marker analysis are summarized in a file called ***SINGLEMARKER.txt**. The output should be self-explanatory. “Ca” denotes cases, “Co” controls. “A”,”B” are alleles, “AA”, “AB”, “BB” are genotypes. **P_BB** is the P-value for testing “BB” vs other genotypes, for instance. **OR_A** is the odds ratio for allele “A”. **left_A**, **right_A** are the limits of the respective 95% confidence interval. MR stands for missing rate. The main outputfile provides a few additional values.

GUI:
Select Files->Inputfile
Association Analysis->Single-Marker-Analysis->Single-Marker (cc)
RUN

Open Output->single marker results (cc)

Nuclear Family Data

With the option **tdt** a file called *_**SINGLEMARKER_FAM.txt** is produced that contains single-marker TDT *(Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506-516*) and **PAT** results. The PAT (paternal asymmetry test) is a test for imprinting and association (*Weinberg CR Methods for Detection of Parent-Origin Effects in Genetic Studies of Case-Parent Triads. Am J Hum Genet 65:229-235*), see also 12.2. When nuclear families with more than one affected child are analyzed, the results in the file ***_SINGLEMARKER_FAM.txt** are results for testing the null hypothesis “**no linkage**“ or “no linkage or no imprinting“, respectively. Use the options of **section 12** to test for **association**. Note that in the file ***_SINGLEMARKER_FAM.txt** alleles are recoded to be integers starting from 1. This is different from the usual behaviour of FAMHAP. Usually, FAMHAP recodes alleles only internally, and the output coding of alleles is identical to the coding in the inputfile.

GUI:
Select Files->Inputfile
Association Analysis->Single-Marker-Analysis->TDT
RUN

Open Output->single marker results (fam)

8. Haplotype Frequency Estimation

FAMHAP determines maximum-likelihood haplotype frequencies estimates from unrelated individuals or nuclear families with an arbitrary number of children with the EM-algorithm (***Becker T, Knapp M Maximum-Likelihood Estimation of Haplotype Frequencies in Nuclear Families. Genet Epidemiol 27: 21-32***). The frequencies are the frequencies in the founders, i.e., those of the parents of the nuclear families and/or the individuals (single-person families). The estimation makes use of the child information in the nuclear families. Mixed samples with both nuclear families and individuals are allowed with the haplotype frequency estimation, but not with all testing procedures. Sibships can be treated as nuclear families with parents with all genotypes missing. Note, however, that frequencies are still estimated with respect to the parental generation. **It is not recommended to use sibships without parents for association testing with FAMHAP**.

8.1. Haplotype Frequency Estimation, Reconstruction, LD, Tagging

With the standard command **./famhap18 pedfile auto** haplotype frequencies are computed. Various call rate thresholds per family unit can be chosen. Markers can be selected for estimation, cf. section 4. By default, all markers of the inputfile are used and missing data is allowed, only family units without genotypes will be removed from analysis.

Example 2:

```
./famhap18 pedfile auto 3 5 6 11 12 17 u80
```

Haplotypes are estimated for SNPs 3,5,6,11,12,17. Family units with less than 80% call rate are removed from analysis.

Specifying the selected SNPs via their numbers can become tedious when many markers shall be selected. With the GUI, expressions as, for instance, **1;2;10-50;53** are allowed. Mind to separate the terms with ";".

GUI:
Select Files->Inputfile pedfile
General Options->Missings->Required callrate per individual/family 80
General Options->Selected SNPs 1;2;10-50;53
RUN

62 SNP markers define the maximum haplotype length, with the Windows32/dos32 version 30 SNPs is the limit. When more than 20 markers shall be analyzed, chose the option **it** (section 8.2) in addition.

By default, only haplotypes with a frequency above 1% are printed. With the **P** option (**GUI:General Options->Print Options->enhanced**), an enhanced output is produced, in particular all haplotype frequencies different form zero are printed. With **PP** (**GUI:General Options->Print Options->detailed**) the output becomes even more detailed.

The file infile.out is the main outputfile and very comprehensive and, in general, self-explanatory. For quick access or usage with other programs, some additional outputfiles are produced.

File with Haplotype Frequency Estimates

The file ***_FREQ.txt** lists all haplotypes with a frequency estimate different from zero.

1 1 2 1 1 1	HaploiD:	8	Freq	0.024405
1 1 2 2 1 1	HaploiD:	12	Freq	0.139406
1 1 2 2 2 1	HaploiD:	14	Freq	0.003619
1 2 2 2 1 1	HaploiD:	28	Freq	0.003478
2 1 1 1 2 1	HaploiD:	34	Freq	0.647588
2 1 1 2 2 1	HaploiD:	38	Freq	0.024405
2 1 2 1 1 2	HaploiD:	41	Freq	0.003486
2 1 2 2 1 2	HaploiD:	45	Freq	0.045323
2 2 1 1 2 1	HaploiD:	50	Freq	0.097853
2 2 1 2 2 1	HaploiD:	54	Freq	0.006957
2 2 2 2 1 2	HaploiD:	61	Freq	0.001739
2 2 2 2 2 2	HaploiD:	63	Freq	0.001739

Table 1: Haplotype frequencies estimated from the file ***nuclearfamilies*** *with the command* ***”./famhap18 nuclearfamilies auto“.*** *The first columns contains the haplotype allele sequence, the third column contains a haplotype ID for retrieval. The last column contains the corresponding frequency.*

File with Reconstructed Haplotypes

Although the main goal of FAMHAP is to estimate haplotype frequencies and to use them for association testing, it is also possible to reconstruct haplotypes. This can be done by assigning to each individual or each nuclear family its most likely haplotype explanation. Note that for nuclear families, the most likely haplotype explanation refers to the family as a whole. In particular, it is possible that the transmission pattern of the four inferred parental haplotypes is ambiguous, i.e., you will not always get a single most likely haplotype explanation of the offsprings. Also note that the tests FAMHAP provides (following sections) are not based on the most likely explanation, but on likelihood weighted lists of haplotype explanations.

By default, the most likely haplotype explanation of each family is written to the files ***H0_HAPLOTYPES.txt** and ***H0_HAPLOTYPESinrows.txt.** The files have different format, choose the one you prefer. With the option **P** option (**GUI: General Options->Print Options->enhanced**), all haplotype explanations with a likelihood weight >=0.05 are printed, with the **PP** option (**GUI: General Options->Print Options->detailed**), all haplotype explanations with a likelihood weight different from zero are printed.

ALLELESEQUENCE	FID	PID	HAPLOID	LIKELIHOOD_WEIGHT
2 1 1 1 2 1	1019	1	34	0.987185
1 1 2 2 1 1	1019	1	12	0.987185
2 1 1 1 2 1	1019	2	34	0.987185
1 1 2 2 1 1	1019	2	12	0.987185
2 1 1 1 2 1	1019	1	34	0.012815
1 1 2 2 1 1	1019	1	12	0.012815
2 1 1 1 2 1	1019	2	34	0.012815
1 1 2 2 2 1	1019	2	14	0.012815

Table 2: Haplotypes corresponding to the file ***nuclearfamilies*** *with the command "****./famhap18 nuclearfamilies auto PP****". Excerpt from the file* ***nuclearfamilies_H0_HAPLOTYPESinrows.txt****. The first four rows refer to the first possible haplotype explanation of the family with family id 1019. The first row refers to the father's transmitted haplotype, the second to the father's non-transmitted haplotype, the third row refers to the mother's transmitted haplotype, and the 4th row refers to the mother's non-transmitted haplotype. Transmission is with respect to the first* ***affected*** *child of the family, if there is an affected child in the family, and to the first child otherwise. The likelihood weight of the first haplotype explanation is 0.987185. There is a second haplotype explanation for family 1019 with conditional likelihood weight 0.01281. It can be found in the four following rows.*

LD measures

By default, a file *LDmeasures.xt is produced wich contains the LD measures D' and r² for all marker pairs. The LD measures are computed from the frequencies of the n-marker haplotypes frequencies by restricting them to 2-marker haplotype frequencies. The file can be opened with the GOLD program by G. Abecasis (*"Abecasis GR and COOKSON WOC. GOLD-Graphical Overview of Linkage Disequilibrium. Bioinformatics 2000 16:182-183"*) in order to obtain a visual LD map.

entropy

With this option, the normalized entropy difference (*Nothnagel M, Fürst R, Rohde K. Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. Hum Hered, 54:186-198)* is computed. The entropy is a multi-marker LD measure.

GUI:
…
General Options->Further Options->entropy
RUN

proxies

In the command line, this option has to occur before the selected SNPs! A generalized r² measure is computed. In contrast to the known situation where r² refers to two SNPs, the generalized r² is computed for LD between a marker set (its haplotypes) and a single marker.

-1	-1	4	5	6	approximated marker:	2	0.98
-1	-1	4	-1	6	approximated marker:	3	0.94
2	3	-1	5	6	approximated marker:	4	0.215
2	**-1**	**4**	**-1**	**6**	**approximated marker:**	**5**	**0.98**
2	3	4	5	-1	approximated marker:	6	0.94

Table 3: File **casecontrol_PROXIES.txt** produced with command "./**famhap18 casecontrol auto u proxies 2 3 4 5 6**". Consider the line in bold: marker 5 and the marker set {2,4,6} have a generalized r^2 of 0.98. Marker 2 is omitted (-1) since the generalized r^2 with marker 5 does not improve when it is included in the set. For more than 12 markers, such selection is not carried out because of computer running time.

GUI:
…
General Options->Further Options->proxies
RUN

Tagging

By default, a file called ***TAG.txt** is produced. Its first row refers to pairwise tagging at a cut-off of r^2=1, the second row to a cut-off of r^2=0.95, the third row to r^2=0.8 and finally the last row refers to tagging at a cut-off of r^2=0.5. The file is read as follows: if a marker is a tagging marker, its number is written to the file. If a marker is tagged by another marker its number is replaced "0".

For up to 15 markers the set of tagging markers is computing by considering all possible marker subsets. Thus, it is guaranteed that the number of tagging markers is minimized. When more than 15 markers are considered, tagging markers are determined with a greedy algorithm, not necessarily resulting in a minimal set.

1	2	3	4	5	6
1	2	0	4	5	6
1	2	0	4	5	6
0	2	0	0	5	6

Table 4: According to the first row, all 6 markers have to be used at a cut-off of r^2=1. At r^2=0.95 and r^2=0.8, markers 1,2,4,5 and 6 form a minimal set of tagging markers. Markers 2, 5, and 6 are sufficient to guarantee pairwise tagging at r^2=0.5 (4th row).

8.2. Haplotype Frequency Estimation with Progressive Extension Technique

it (progressive extension technique, one direction)

./famhap18 infilename auto it

With this option, haplotype frequencies are computed in a marker-iterative mode (***Becker T, Knapp M (2004) Maximum-Likelihood Estimation of Haplotype Frequencies in Nuclear Families. Genet Epidemiol 27: 21-32***). It is the technique described for unrelated individuals by *David Clayton (SNPHAP)*. The technique can be viewed as a fast method to obtain good approximations of the true frequency estimates. In many cases the results with and without the

it option are identical. If you want to analyze more than 20 SNPs (or respectively less when multi-allelic markers are included), this option is obligatory. The reason is, that without the **it** option, it would be necessary to allocate storage for all theoretically possible haplotypes (2^n for n SNPs). Note that the results with the **it** option may depend on the order of the markers.

Hint: The **it** option can be useful also for less than 20 SNP markers. It often allows faster computation and reduces the required computer storage**.**

GUI:
…
General Options->Estimation Mode->it
RUN

With the additional option **ro** the marker order is reversed.

General Options->Estimation Mode->ro

itt (progressive extension technique in two directions, followed by estimation with inferred haplotype parameter set)

./famhap18 infilename auto itt

With this option, haplotype frequency estimates are obtained in three steps.

Step 1: FAMHAP internally calculates the results one would get with
./famhap18 infilename auto it

Step 2: FAMHAP internally calculates the results one would get with
./famhap18 infilename auto it ro
i.e., the progressive extension technique is applied in reverse order.

Step 3: Both steps yield a set of haplotype parameters which have a frequency estimate different from zero. The union of the haplotype set form step 1 and the haplotype set form step 2 is taken. From this union, the final haplotype frequencies are estimated. Not all theoretically possible haplotype explanations are allowed in step 3: only haplotype explanations consisting exclusively of haplotypes from the union set are allowed.

GUI:
…
General Options->Estimation Mode->itt
RUN

8.3. Haplotype Frequency Estimation According to Affection Status

dp (haplotype frequency estimation with respect to affection status)

The program uses a doubled parameter set of haplotype frequencies, one set for cases and/or transmitted haplotypes and one set for controls and/or non-transmitted haplotypes. Unaffected children of nuclear family are only used to infer and weight the possible haplotype explanations of the family. When a nuclear family has more than one affected child, the transmission status of haplotypes is defined with respect to the first affected child of the family.

.

When the **dp** option is used, two files with LD measures are written, one for cases and another one for the (pseudo-)controls. Entropy will be computed for cases and controls separately. Haplotype reconstruction is also conducted using the affection status. As a consequence, the results most not be used as input for other association analysis software when **dp** was chosen.

GUI: General Options->Estimation Mode->dp

9. Identification of Genotyping Errors Using Haplotypes

By default, the program checks all marker of the infile for Mendelian errors. If there are Mendelian errors, the program writes them to the file ***MENDELERRORS.txt**. The program proceeds with the haplotype frequency estimation even when it has encountered Mendelian errors. In families with Mendelian errors, the offspring genotype is set to be missing. Note that this may lead to bias [*Mitchell AA, Cutler DJ, Chakravati A (2003) Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. Am J Hum Genet 72: 598-610*]. Families with recombinations are excluded from further analysis. Families with recombinations are listed in the outputfile and the file ***_families_with_recos.txt**. There is the possibility to systematically screen the data for potential genotyping errors. With this method, genotypes which lead to Mendelian errors, recombinations or unlikely haplotype explanations are set to be missing and a new data file is produced. Such practice may lead to an inflated type I error rate in subsequent analysis, an issue which still has to be investigated.

The error search routine can be started as follows:

```
./famhap18 pedfile auto  ed maxhap 50 quotient 1000 minloc 5 maxloc 10
```

ed (systematic search for potential genotyping errors)

With the **ed** option the program searches systematically for potential genotyping errors. For this purpose, at a given marker either all genotypes for all family members or the genotype of just one family member are set to zero. The outputfile lists all constellations under which recombinations disappear, or under which the families (individuals) new most likely haplotype explanation is more likely by the factor **quotient** (see below) than the most likely haplotype explanation according to the original genotyping.

maxhap <m>

The systematic search is conducted for all sets of neighboring markers for which the ML-estimation determines no more than <m> haplotypes with a frequency greater than zero. The default is maxhap=50.

maxloc <l> minloc <k>

The systematic search is conducted for all sets of neighboring markers of size less than or equal to <l> and greater than or equal to <k>.

quotient <r>

The program stores all constellations for which treating a genotype as missing, leads to a most likely haplotype explanation, which is more likely than the original most likely haplotype explanation at least by the factor <r>. The default is 10000.

With the **ed** option a file named *__corrdata_** is produced in which the potentially erroneous are set to be missing.

Hint: The error search routing can be used to produce a file without Mendelian errors and recombinations. If an extremely high value is chosen for the **quotient** parameter, no data manipulation is performed, except for the treatment of Medelian errors and recombinations. If the input data contains double recombinants, it may be necessary to run **ed** on the file ***_corrdata file.**

```
FID                   PID     MARKER     LEFT  RIGHT   QUOTIENT
10026                 all     3          2     4       4.198904e+03
ftold:  1 1 2 2.233236e-02 ftnew:  1 2 2 1.930168e-01
fntold: 1 2 1 2.346756e-02 fntnew: 1 1 1 6.595580e-01
mtold:  1 1 2 2.233236e-02 mtnew:  1 2 2 1.930168e-01
mntold: 1 1 1 6.595580e-01 mntnew: 1 1 1 6.595580e-01
```

Table 5: Typical output of the genotyping error search routine

FID: Family with potential genotyping error.
PID: Person with potential error. “all” means that all family members where set to be missing, i.e., one of the members has a potential mistake.
MARKER: The number of the marker with a potential error.
LEFT RIGHT: Denotes the extension of the haplotype, here from marker 2 to marker 4.
QUOTINET: Increase of likelihood of the family.
ftold: Father’s transmitted haplotype according to the most likely haplotype explanation of the original data. Corresponding frequency.
ftnew: Father’s transmitted haplotype according to the most likely haplotype explanation of the modified data. Corresponding frequency.
fnt: As before, but non-transmitted
mt, mnt: As before, but referring to mother.

GUI:
Identification of Genotyping Errors->Search genotyping errors
Identification of Genotyping Errors->minloc, maxloc, maxhap, quotient

10. Likelihood Ratio Tests

In general, it is not a valid strategy to estimate haplotype frequencies from a sample, construct a contingency table from these estimates and treat the values of that table as observed units. The cell counts are not observed but were estimated, and therefore have a higher variance. Instead, the likelihood L(S) of a sample S can be used to construct a likelihood ratio test.

lrtest (likelihood-ratio test)

The obvious likelihood ratio test is performed. The test statistic is 2* (ln L(cases) + ln L(controls) – ln L(cases and controls)) which follows approximately a chi-square distribution with [number of haplotype parameters-1] degrees of freedom. For family data, cases are the first affected child of each nuclear families and the controls are the nontransmitted haplotypes. Transmission is with respect to the first affected child. As before, mixed samples are allowed. The chi-square approximation may fail with sparse cell counts. On the other hand, when many parameters are equal to zero, the test can be too conservative. My feeling is, that [number of haplotype parameters different from zero-1] is the best possible guess for the degrees of freedom (DF). **However, we recommend to use the permutational tests of sections 11 and 12 in the context of haplotype analysis for gene-based analysis.**

The LR-test can be applied to case-control data and nuclear family data and to data mixed of both structures. Note, however, that even for families the likelihood ratio test is not robust with respect to population stratification. Therefore, the **LR-test is not the recommeded analysis option for family data** FAMHAP provides. Instead, a robust test for association for general nuclear families is described in section 12.

GUI:
Association Analysis->Likelihood-Ratio Test->OMNIBUS

lrtest SIMULATIONS <n> (permutational analogue of likelihood ratio test for case-control data)

A problem of the likelihood ratio test is that the chi-square approximation may be poor with many degrees of freedom, for instance, when several haplotypes have a low frequency. A permutational analogue of the LR-test is implemented to overcome this problem. In each permutation replicate, the disease status of the individuals is randomly permuted, such that the ratio of cases and controls is kept constant. For each simulated data set the likelihood ratio statistic is computed, (in particular a new haplotype frequency estimation is needed in each permutation replicate), and the P-value is calculated from the number of times that the chi-square of the replicated data is greater than or equal to the chi-square of the real data. Following the **SIMULATIONS** option, the number <n> of permutation replicates can be specified. The default is n=10000. The permutational likelihood ratio test is implemented only for case-control data. For families, use the options of section 12.

Haplotype	HaploID	Freq_AFF	Freq_NOTAFF
1 1 1	0	0.402047	0.369782
1 2 1	2	0.019930	0.004972
2 2 1	6	0.457290	0.473364
2 2 2	7	0.116942	0.148346

(NUMBER OF PARAMETERS DIFFERENT FROM ZERO)-1 =	5
Likelihood Ratio Test Chi-Square Data:	18.58133
P-value according to chisquare-distribution:	0.00229952
SIMULATED p AT MARKER COMBINATION:	0.00640000000.

Table 6: Output obtained with the command **"./famhap18 casecontrol u 20 21 22 lrtest SIMULATIONS 10000"**. There are 5 degrees of freedom. The permutational analogue of the

likelihood ratio test shows that the P value computation with the asymptotic distribution is not exact enough.

Association Analysis->Likelihood-Ratio Test->OMNIBUS
Association Analysis->Likelihood-Ratio Test->SIMULATIONS

oneDF (likelihood ratio test for a single haplotype)

With this option, the program computes the chi-square for a likelihood ratio test with 1 d.f. for haplotypes with a relevant difference in frequency between cases and (pseudo-) controls (pre-test a very liberal level). An additional run of the EM-algorithm is required for each haplotype of interest. For k SNPs, we have 2^k*2 parameters with the **dp** option. For the test we calculate frequencies under the restriction that the frequencies for the haplotype of interest are the same in cases and controls. Thus, we have only 2^k *2-1 parameters, and from the likelihoods we get a likelihood ratio test with 1 DF. The P-values have to be Bonferroni-corrected by the number of haplotypes! Even for family data, the test is not robust against stratification or non-random distributions of missing genotypes. It is not a TDT-like test, but more in the spirit of a HHRR. Some caution is necessary when the haplotype of interest has a low frequency in at least one group. The chi-square approximation may fail then. A permutational analogue of this test is not implemented. However, the options **haptdtmax** and **hapccmax** provide tests which go in that direction, see sections 11 and 12, respectively.

```
ChiSquare 1 DF for haplotype 0 is 2.714896.
P-value for haplotype 0 according to chisquare-distribution: 0.0994155

ChiSquare 1 DF for haplotype 2 is 11.840926.
P-value for haplotype 2 according to chisquare-distribution: 0.00057943

ChiSquare 1 DF for haplotype 7 is 5.236858.
P-value for haplotype 7 according to chisquare-distribution: 0.0221132
```

Table 7: Output obtained with the command **“./famhap18 casecontrol 20 21 22 oneDF”**. Hapotype 2 is the haplotype with haploID 2 from table 6. The Bonferroni-corrected P-value is 0.00348.

Association Analysis->Likelihood-Ratio Test->1 DF

11. Association Analysis with Weighted Haplotype Explanations Lists (Case-Control Data)

In this section, several Monte-Carlo simulation based tests are presented. Haplotype frequencies are estimated only once, from the joint case-control sample. Weighted haplotype explanations lists (WHLs) are assigned and are used to construct contingency tables. In each replication of the MC simulations, affection status of individuals is randomly permuted, such that the ratio of cases and controls is kept constant. For quantitative traits, the measurements are randomly distributed over all individuals in each permutation replicate. P values are always computed as **s/n**, where **n** is the number of permutation replicates, and where **s** is the number of permutation replicates leading to a test statistic higher than or equal to that of the real data.

SIMULATIONS <n>

With this option the number of permutation replicates can be specified. The default is n=10000.

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->SIMULATIONS

11.1. Haplotype Analysis

hapcc (RECOMMENDED OPTION for Case-Control Data!)

This option does not require a new maximum-likelihood estimation in each permutation replicate (***Becker T, Cichon S, Jönson E, Knapp M (2005) Multiple testing in the context of haplotype analysis revisited: application to case-control data. Ann Hum Genet* 69:*1-10***) As a consequence, it is much faster than the permutational analogue of the likelihood ratio test. Firstly, haplotype frequencies are estimated form the compound sample of cases and controls. Each individual is assigned its list of possible haplotype explanations and each of these explanations is assigned its conditional likelihood given the individual's genotype and given the estimated haplotype frequencies. From the weighted lists a contingency table of haplotype counts in cases and controls, respectively, is constructed and the usual chi-square statistic for such tables is calculated. As this table is constructed and not observed, significance has to be obtained via Monte Carlo simulations, as described at the beginning of this section.

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Case-Control->Omnibus Statistic
hapccmax

These options can be used instead of **hapcc**. With **hapccmax** the test statistic refers to the best haplotype, i.e., the haplotype with the highest chi-square value when comparing the haplotype to all other haplotypes via a 2x2 table. The resulting P-value is corrected for the number of different haplotypes at the marker combination via the Monte-Carlo simulation procedure. I prefer **hapcc** instead of **hapccmax.**

Haplotype	Cases	Controls	FreqCa	FreqCo	OR
1 1 1	486.4	633.5	0.459	0.42	1.17
2 1 1	422.6	629.5	0.399	0.417	0.93
2 2 2	145.9	243.5	0.138	0.161	0.83
1 2 2	2	3.5	0.002	0.002	0.82
2 2 1	1	0	0.001	0	--
1 1 2	1	0	0.001	0	--
2 1 2	1.1	0	0.001	0	--

```
Hapcc Statistic: 9.280
HapccMax Statistic: 3.911
MARKERCOMBI: 1  MARKER: 67 68 69 p: 0.05018000000 hapcc
MARKERCOMBI: 1  MARKER: 67 68 69 p: 0.10037000000 hapccmax
```

Table 8: Output with "./famhap18 casecontrol.fhp auto itt 67 68 69 hapcc hapccmax SIMULATIONS 100000".

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Case-Control->Max Statistic

11.2. Haplotype Trend Regression

htr (haplotype trend regression)

With the **htr** option, the permutational analogue of the haplotype trend regression test proposed by Zaykin et al. (*Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 71:337--348*) is conducted. It is computationally very intense. Therefore, **hapcc** is the recommended FAMHAP option.

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Case-Control->HTR

htr q *(haplotype trend regression for quantitative traits)*

Haplotype trend regression is applied to quantitative traits. In the inputfile, the row with affection status must contain the quantitative measurements. Missing data is coded "-".

General Options-> Quantitative
Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Case-Control->HTR

11.3. Diplotype Analysis

Diplotypes, i.e., pairs of haplotypes belonging to an individual, are considered rather than haplotypes. **dipcc** is the analogous option to **hapcc**.

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Case-Control->Diplotypic Analysis

11.4. Simultaneous Analysis of Unlinked Regions

As an example consider the following command:

```
./famhap18 infile outfile region 1 2 3 region 4  region 5 6 hapcc allcombi maxmarker 1 maxregions 2
```

region

The **region** option specifies the start of a region of tightly linked markers. In the example statement above, markers 1, 2 and 3 are supposed to be from the same LD region. The next **region** option indicates the end of the first region and the start of a new region, which is unlinked to the first. The third **region** option indicates the end of the second region and the start of the third region. Thus, in our example we have three unlinked regions:

Region 1: markers 1, 2 and 3
Region 2: marker 4

Region 3: markers 5 and 6

For each region, the respective haplotype frequencies are estimated by the program.

The **hapcc** option is recommended for the analysis of unlinked regions. The analysis is now based on a chi-square test statistic for the following contingency table. The rows of the contingency table represent cases and controls, respectively; the columns do not longer correspond to a haplotype of a chosen marker combination. Instead, the columns belong to a configuration of the form h_1-h_2-….-h_m, where m is the number of unlinked regions, and where h_i is an allele or haplotype of region i. The contributions of each individual to the columns are computed in a natural way and add up to 2.

With the **GUI** the region option is selected in a slightly different way:

General Options-> Selected SNPs: 1-6
General Options-> Region Starts: 1;4;5

11.5 Treating Rare Haplotypes

nonoise

With this option, haplotypes which occur less than five times in cases and controls together, do not contribute to the test statistic. The option can be used together with all the methods of sections 11 and 12. Note, however, that using nonoise is not necessarily a great advantage, since the problem of rare haplotypes and increased number of degrees of freedom is already accounted for via the Monte Carlo simulations.

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->nonoise

11.6 Casepair vs Controlpair

sibvssib

With this option, paired cases can be tested against paired controls. In the simulation procedure, the affection status is permuted for both sibs of a pair simultaneously. In this way, linkage is accounted for. The **sibvssib** option requires **special care**: all individuals must be member of a sibpair, the sibs must follow each other in the infile, the sibs must have **different(!)** family IDs and call rate restrictions must not be set (required call rate 0%).

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Case-Control->sibvssib

11.7 Treating Discordant Sibpairs

paired

With this option, discordant sibpairs can be tested for association in a simple manner. In the simulation procedure, the affection status is permuted within the discordant sibpair. In this way, linkage is accounted for. The **paired** option requires **special care**: all individuals must belong to a discordant sibpair, the sibs must follow each other in the infile, the sibs must have **different(!)** family IDs and call rate restrictions must not be set (required call rate 0%).

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Case-Control->paired

12. Association Analysis with Weighted Haplotype Explanations Lists (Family Data)

SIMULATIONS <n>

With this option the number of permutation replicates can be specified. The default is n=10000.

12.1. Testing for Association (HAP-TDT)

haptdt **(RECOMMENDED OPTION for Family Data!)**

A haplotype-based test for nuclear family data, which is robust against population stratification has been proposed by Hongyu Zhao *(Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, Sun F, Kidd KK (2000) Transmission/Disequilibrium Tests Using Multiple Tightly Linked Markers. Am J Hum Genet 67:936-946*) and has been extended to general nuclear families *(Knapp and Becker. Family-based association analysis with tightly linked markers. Hum Hered 2003. 56:2-9*). This test is based on Monte-Carlo simulations, in which the set of transmitted and non-transmitted genotypes is randomly permuted in each permutation replicate. Details can be found in the cited literature. With the option **haptdt**, the haplotype-based test can be run for general nuclear families. This test uses only families which are fully genotyped. However, it makes use of the haplotype frequencies estimated before. All families are included into the frequency estimation, but for the subsequent testing method only fully typed families are used. If you have families, where both parents and at least one child are fully typed, the fully typed affected children will be incorporated.

Note: Within the simulation procedure, transmitted and non-transmitted genotypes are permuted either for both parents or no parent. According to our opinion, it is not possible to allow permutation of transmission status of just one parent when weighted lists of haplotype explanations are considered. (Hint: To see this, consider a trio which is heterozygous for two SNPs for all family members and consider all permutation schemes). As a consequence of the construction principle of the permutated families, the **resulting test is not equivalent to the TDT in the case of trios and a single marker**.

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Families->Omnibus Statistic

haptdtmax

With this option, the chi-square statistic of the marker combination is replaced with the maximum of the chi-squares of the single haplotypes (maxTDT statistic). The program yields

P-values, which are corrected for considering different haplotypes via the simulations procedure.

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Families->Max Statistic

12.2. Association and Imprinting (PAT)

FAMHAP provides a haplotype-based test for imprinting and the option to restrict association analysis to paternal or maternal transmissions.

PAT

With this option a haplotype-based version of the parental asymmetry test (*Weinberg CR Methods for Detection of Parent-Origin Effects in Genetic Studies of Case-Parent Triads. Am J Hum Genet 65:229-235*) is computed. Note that a direct impact of either the father's or the mother's genotype on the child's affections status may lead to invalid evidence for imprinting.

```
Haplotype             P           M

2 1 2 2 1 2         5.00        2.00
2 1 1 1 2 1        12.50       33.50
2 2 1 1 2 1        10.00        5.00
1 1 2 2 1 1        14.50        5.50
2 1 1 2 2 1         2.00        1.00
1 1 2 1 1 1         2.00        0.00
2 1 2 1 1 2         2.00        0.00
2 2 1 2 2 1         0.00        1.00

MARKERCOMBI: 1  MARKER: 1 2 3 4 5 6 p: 0.03010000000 Default Statistic
```

Table 9: Output obtained with **"./famhap18 nuclearfamilies auto PAT".** Among heterozygous children who carry haplotype 2 1 2 2 1 2, for instance, 5 received that haplotype from the father (P) and 2 received that haplotype from the mother (M). The P value corresponding to the whole parent-of-origin table is 0.0301.

Note: The number of counts of paternally and maternally inherited alleles among heterozygous obtained with the **PAT** option may differ from the number of counts listed in the single marker analysis file ***SINGLEMARKER_FAM.txt**. The reason is that with **PAT** half units are counted when parental origin is ambiguous. This treatment does not affect the validity of the test since P-values are computed via Monte Carlo simulations. Furthermore, when the **allcombi** option (section 13) is used, parental origin can sometimes be inferred by considering neighbouring markers. This may lead to further differences with the ***SINGLEMARKER_FAM.txt** file.

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Families->Imprinting

PATmax

PAT with maximum statistic.

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Families->Imprinting Max Statistic

haptdt imppat / haptdt impmat

The haptdt option can be restricted to analysis of maternal (**haptdt imppat)** or paternal (**haptdt impmat**) transmissions. With the following command only transmissions from the mother are relevant for the test.

./famhap18 infilename outfilename haptdt imppat

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Families->Omnibus Statistic
Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Families->Maternal Transmissions (=imppat!)

12.3. Tests for Linkage

linkage

The family-based tests from 12.1. and 12.2. are carried out as tests for linkage. Differences from the association test occur, if there are nuclear families with more than one affected child. With the **linkage** option, the transmission/non-transmission status of the genotypes is permuted independently for each child of the family. The option is available only on the command line.

13. Testing Mulitple Marker Combinations, Global P values

13.1 Testing Mulitple Marker Combinations, Global P values

FAMHAP provides options that allow the analysis of multiple marker combinations in a single run of the program. In addition, the program yields a corrected P-value for the **global null hypothesis** that **none of the marker combinations tested is associated with the phenotype**. This global P-value accounts for the multiple testing but also for the dependence of the tests due to linkage disequilibrium (LD). The principle is described for family data in ***Becker T, Knapp M (2004), A powerful strategy to account for multiple testing in the context of haplotype analysis. Am J Hum Genet:561-70***. The correction procedure can be used together with all tests from sections 11 and 12.

allcombi

All marker combinations (=selections of marker subsets) are tested. Can be applied to family and case-control data.

Note: If **allcombi** is applied to familiy data (**haptdt**), at each marker combination only those family subtrios contribute to the test statistic of the combination, which are fully genotyped at the marker combination. (In nuclear families in which the parents are fully genotyped, all transmissions to fully typed children are counted, even if there are additional not fully genotyped children). The test statistics depend on the haplotype frequency estimates which belong to all markers. Only families which can be incorporated into this estimation can contribute to the statistics of the marker combinations. Therefore, the haplotype frequency estimation always makes use of missing data together with the **haptdt** option, because it is possible that at some of the marker combinations genotyping is complete, but missing data does is not used for the association tests themselves. In this way**, robustness against population stratification** is guaranteed.

```
3 MARKERCOMBINATIONS

MARKER COMBINATION: 1 LOCI: 2

Test Statistic: 2.880

very rough p-value guess for linkage: 0.089686

Haplotype       T           NT

1             31.00       19.00
2             19.00       31.00


MARKER COMBINATION: 2 LOCI: 1

Test Statistic: 7.053

very rough p-value guess for linkage: 0.00791179

Haplotype       T           NT

2             49.00       26.00
1             26.00       49.00


MARKER COMBINATION: 3 LOCI: 1 2

Test Statistic: 15.226

very rough p-value guess for linkage: 0.00163297

Haplotype       T           NT

2 1           70.00       37.00
1 2            0.00        2.00
1 1           26.00       47.00
2 2           19.00       29.00

MARKERCOMBI: 1  MARKER:   2 p: 0.16480000000 haptdt
MARKERCOMBI: 2  MARKER: 1   p: 0.02290000000 haptdt
MARKERCOMBI: 3  MARKER: 1 2 p: 0.00860000000 haptdt

MAXMARKER: 2 WINDOW: 0 BEST P VALUE: 0.008600. Markercombination 3 OMNIBUS
Statistic.
GLOBAL_P 0.015000000000.
```

Table 10: Output obtained with **"./famhap18 nuclearfamilies auto 1 2 allcombi haptdt alpha 0.10"**. Three marker combinations were tested. The best raw P-value (0.0086) is obtained for MARKERCOMBI 3 which considers haplotypes referring to markers 1 and 2. The very rough P-value

guesses are indeed rough. They are only included to get a notion about how many SIMULATIONS might be necessary.
For haplotype 2 1, for instance, there are 70 transmissions and 37 non-transmissions from heterozygous parents.
The Global P-value is 0.015 (adjusted for multiple testing).

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Mulitple Combinations->Multiple combinations

The **allcombi** option can be combined with the following options:

window

Only marker combinations that do not leave out in between markers are tested. I never use this option because I do not think that it makes sense to consider only neighbouring markers.

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Mulitple Combinations->window

slidingwindow <n>

Only marker combinations of size n that do not leave out in between markers are tested. My comment to the **window** option applies again.

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Mulitple Combinations->slidingwindow: n

maxmarker <n> (RECOMMENDED OPTION: maxmarker 2)

Only marker combinations with less than or equal to <n> markers are tested. **I recommend "allcombi maxmarker 2" as a standard analysis strategy.** For large samples higher maxmarker values might be useful.

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Mulitple Combinations->maxmarker <n>

maxregions <n>

Only marker combinations which include less than or equal to <n> regions are tested, c.f. 11.4. Note that **maxmarker** and **window** now refer to each region, i.e., all marker combinations which satisfy the criteria imposed by **maxmarker** and **window** at each region are tested

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Multiple Combinations->max regions

13.2. Reduce Required Computer Storage

alpha <r>

Choose 0 <= r <= 1. The tests are carried out as a test of size <r>, i.e., if the P-value is greater than <r>, the program does not compute the P value, but provides simply that information.

Hint: The alpha option can be used to overcome computer storage restrictions. When many marker combinations and/or a high value of **SIMULATIONS** are chosen, the **alpha** option can be extremely helpful. In addition, the **it** option should be used.

13.3. Improve Exactness of Computation of Global P-value

secondon

Improved Global P-value computation. Ties are broken by consideration of the second best P value.

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->secondon

13.4. MC-based Association Testing of Imputed Genotypes

With the (already known) command **./famhap18 infile auto hapcc allcombi maxmarker 1 impute** you get, besides the global P-value, a P-value for each marker. For these single-marker tests, all individuals are used, even if the individual is not genotyped at the marker. The information comes from the haplotypes of the WHLs and the allelic status at the marker in question. In other words, missing genotypes are imputed.

In general, all family units contribute to the haplotype frequency estimation, in particular, trios or individuals with unknown affection status. Based on the frequencies and WHLs, allelic status at each marker can be inferred (possibly with uncertainty) for each individual. Only individuals with affection status 1 (control) or 2 (case) contribute to association testing. In the extreme case, when a marker is genotyped only for the families or individuals with unknown disease status, the association test for the marker depends only on imputed genotypes.

A description of the required inputfile can be found in section 15 where a quick alternative for association testing of imputed genotypes is introduced.

It is not recommended to use impute when the genotyping rate differs strongly between cases and controls.

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->impute

14. Conditional Analysis

Many of the test procedures may be carried out on stratified data sets, in order to test for causality of a certain marker or even marker combination.

Conditional Analysis of Nuclear Family Data

MUTE <n>   MUTEALLELE   <k>

Only transmissions from parents which are homozygous at marker n are counted. The **MUTEALLELE** option is not obligatory. If **MUTEALLELE <k>** is chosen, only transmissions from parents are counted which are homozygous for allele k at marker n. The idea is more or less that described in "***Homozygous parent affected sib pair method for detecting disease predisposing variants: Application to insulin dependent diabetes mellitus (1993)*** *Wendy P. Robinson, Jose Barbosa, Steven S. Rich, Glenys Thomson Genet Epidemiol 10:273-88"*. The **MUTE** option is available only on the command line. A drawback of the option is that many parents do not contribute to the conditional statistic. Therefore, FAMHAP implements an alternative method. Consider the following command:

./famhap18 nuclearfamilies auto strat 4 2 haptdt (RECOMMENDED OPTION!)

With this option, marker 2 is tested conditional on marker 4. Haplotype phase is accounted for! It is not required that families are homozygous at the stratification markers. P-values are obtained via MC-simulations. For the permutation replicates, the transmission ratio at the conditioning marker (marker haplotypes) is kept. The test statistic is explained in the following example table. The **strat** option can be combined with **allcombi**. Multiple conditioning markers are specified as a follows: ….**strat 1 strat 2 strat 45…**

Haplotype		T	NT		
1	1	82	42	T_1=7.094	$T=T_1+T_2$=8.790
2	1	18	24		
1	2	33	61	T_2=1.695	
2	2	1	7		

Table 11: The TNT_Table for markers 2 and 4 is shown (file **nuclearfamilies**). Haplotype 1-1 has a T/NT-ratio of 82:42, while haplotype 2-1 has a T/NT-ratio of 18:24. Thus, the T/NT-ratio is not only determined by the allelic status at the last allele, but modified by the first allele. The corresponding 2x2 table (82,42;18,24) leads to a test statistic T_1=7.094. T_2 is computed for haplotypes 1-2 and 2-2 in the same way. The final test statistic is $T=T_1+T_2$. With the command ./**famhap18 nuclearfamilies auto strat 4 2 haptdt** the P-value corresponding to **T** is obtained with MC simulations. The results is P=0.02999.

General Options->Selected markers 2
Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->STRATIFICATION MARKERS 4

15. Imputing (Case-Control Data)

With the command

./famhap18 infile proxies auto it fimpute 7 gpsbench 0.50

marker 7 is imputed and tested for association using all markers of the inputfile (or all selected markers) that are genotyped with a relevant portion in cases and controls and the training data (HapMap trios t.ex.).

With the command

./famhap18 infile proxies auto it fimpute 7 gpsbench 0.50 with_all_markers

Marker 7 is imputed and tested for association using all markers of the inputfile (or all selected markers) even if they are not genotyped (with a relevant portion) in cases and controls. This should be reasonable only in very special cases. Therefore, the with_all_markers option is not recommended.

The **gpsbench** option is used to exclude individuals whose genotype has an imputation score of less than 0.50. The definition of the genotype score will follow in a respective publication. With the **fimpute** option, P-values are computed using the chi²-distribution. This leads to a slightly conservative test. The recommendation is to use **fimpute** to conduct a pre-test and to validate the result thereafter as described in 13.4. The optional **proxies** option provides a measure how well marker 7 is approximated by markers 1-6, i.e.; you can judge if imputing makes sense at all for your data. The measure is called **generalized r²**.

To test an imputed marker you need an inputfile (**standard FAMHAP input format**) of the following form: you have **cases** and **controls** (**study sample**), nuclear **families** (**training sample**) and/or individuals (training sample) with unknown disease status (AFF=0). The training sample is expected to be typed for all markers, while your study sample is typed only for some markers. In the example, marker 7 has missing genotypes (0 0) for all cases and controls. It is forbidden that there is true genotype data for marker 7 either only in cases or only in controls. In such a case, imputing leads to strongly inflated typed I error! **Association testing of an imputed marker should only be done if the marker is NOT genotyped both in cases and controls. I strongly recommend not to merge different case-control data sets with partially overlapping marker panels. It is clear that such practice leads to inflated type I error.** Inflated type I error will not occur when genotyping rate in cases and controls is equal; equal to zero, for instance.

With the **fimpute** option you can impute and test only one marker at a time.

General Options->Selected SNPs 1-6
General Options->Further Options->proxies
Association Analysis-> Analysis of Imputed SNPs->Analysis of imputed SNPs
Association Analysis-> Analysis of Imputed SNPs->Imputed SNP 7
Association Analysis-> Analysis of Imputed SNPs->required genotype prediction score 0.50

Hint: If you use the fimpute option, the imputed genotypes with prediction score are written to the file *_impgeno.

16. Testing for Excess of Rare Haplotypes in Cases

The goal of this feature is to detect association in the presence of multiple, rare disease variants from within on gene. The basic idea is that in this situation one would expect an access of rare haplotypes in cases. This can be tested with the following framework:

I. Estimate haplotype frequencies from the compound sample of cases and controls and assign to each individual the list of its possible haplotype explanations with respective likelihood weights (WHLs).

II. For each *x*, $0 < x < 1$, consider the class $H_{<=x}$ of haplotypes with a frequency less than or equal to *x* and the class $H_{>x}$ of haplotypes with a frequency greater than *x*. Furthermore, consider the 2x2 contingency table T_x whose rows correspond to the two classes, whose columns correspond to case/control status and whose cell counts are determined using the haplotype explanation lists from I. Compute the corresponding test statistic t_x for 2x2 tables. Note that only a finite number of values for *x* has to be considered.

III. Define $t:=max\ (t_x)$

IV. The distribution of *t* is determined using Monte-Carlo simulations. Within each permutation replicate *i* of these simulations, case/control status is randomly permuted such that the ratio of cases to controls is kept constant. For the simulated data, $t_i=max\ (t_{x,i})$ is computed from the contingency tables $T_{x,i}$, which are constructed using the weights from I. Finally, the *P* value can be computed as $P=\#\{i \mid t_i >=t\}$.

diversityhap

The procedure described above is conducted.

Note: It is recommended to use the u100 option with diversity mapping, since different missing rates for cases and controls may lead to a bias.

```
diversity P 0.017100
diversity border  0.003957 case heterogeneity
```

Table 12 Output obtained with **"./famhap18 casecontrol 1 2 3 4 5 6 7 8 9 10 u100 itt diversityhap"**. The P-value is 0.0171 and the cut off x is defined by a haplotype frequency of 0.003957. The statement "case heterogeneity" means that haplotypes below the cut off frequency are more frequent among cases than among controls.

Association Analysis-> Association Analysis with Weighted Haplotype Explanationlists->Case-Control->Rare Variants Test

17. GWAS: Running FAMHAP Repeatedly (Featured!)

For the analysis of GWAS it is desirable to run the same command repeatedly. An obvious application would be to conduct the haplotype LR-test (**lrtest**) for all pairs of SNPs less than 30 KB apart. This task can be performed with the **GUI**. It is assumed that you have an inputfile with data of one chromosome and an additional mapfile. Each row of the mapfile corresponds to a marker of the inputfile, in the same order of occurrence. The number of columns of the **mapfile** is flexible. The **last column** must contain the physical position of the marker in base pairs, preceding rows are ignored.

To conduct the chromosome-wide two-marker haplotype analysis select the following options with the **GUI**:

Select Files->Inputfile-> inputfile
Select Files->Mapfile-> mapfile
Association Analysis->Likelihood-Ratio-Test->OMNIBUS
RUN->run command for all sets with 2 markers covering less than 30 KB
RUN

The RUN button now starts the program runFamhapdos.exe which starts famhap18dos32.exe (or famhap18dos64.exe if you are working on a 64bit machine) for all pairs of SNPs meeting the KB cirterion. All P-values are written to the file *_Pvalues.txt.

Note: With this option, the only accessible outputfile is the file **inputfile_Pvalues.txt.** Each row corresponds to one **famhap18dos32** call (=one test = one P-value). In order to improve running time for inputfiles with many markers, **runFamhapdos** first produces new data files from the **inputfile**. The files contain windows of 100 marks with an overlap that is set depending on the chosen distance (30 KB in the above example). **runFamhap** then calls **famhap18dos32** for the subfiles and marker combinations. The results are nevertheless written to the file **inputfile_Pvalues.txt** and the numbering of markers found there refers to the numbering in the main infile.
Due to the overlap of files, a small portion of P-values maybe computed twice.

A row of the file **inputfile_Pvalues.txt** could look like this:

```
inputfile_301_405 LR-TEST haplotype_ID OMNIBUS P 0.528 addCORRFACTOR 1 u bothsexes 385 397
```

In this example, runFamhap has produced the file inputfile_301_405 which contains markers 301-405 of the inputfile. The likelihood-ratio test results in a P-value of 0.528. Missing data has been used (u) and analysis was not restricted according to sex (bothsexes). The P-value belongs to markers 385 and 397 of the inputfile, although actually markers 85 and 87 of the file inputfile_301_405 were analyzed.

It is, of course, possible to run the runFamhapdos.exe directly from the command line.

DOS:

The syntax is

```
runFamhapdos famhappath infilepath mapfilepath <n> <m> outputfile
      <famhapoptions>.
```

<n> is the size of the marker sets that shall be considered and <m> is the marker distance in KB.

Examples of famhappath:

a) "C:\folder1\famhap18dos32.exe"

b) Folders with blanks: a folder with name "new folder" translates to "newfol~1" (six first letters, not counting the blank), example "C:\newfol~1\famhap18dos64.exe" (if you use the 64version). Mind to hyphenate the path! ("path").

The same rules hold for the **infilepath** and the **mappath**.

A complete command:

```
runFamhapdos "C:\myfolder\mysubfolder\famhap18windows\famhap18dos32.exe" "C:\
inputfile" "C:\newfol~1\mapfile" 2 30 auto lrtest male
```

Unix:

Compile gcc runFamhapLinux.c –o runFamhapLinux –O3 and use as described above, mind to replace "\" with "/".

```
./runFamhapLinux "~/myfolder/mysubfolder/famhapGUI/famhap18Linux"
"~/inputfile" "~/myfolder/mapfile" 2 30 auto lrtest male
```

The functionality can be combined with most FAMHAP options. The **region** option is not allowed. To run the single-marker-analysis options (**singlecc**, **tdt**), the **fimpute** option or the error search routine (**ed**) repeatedly does not make sense.

18. Further Options

#define **MAXEX** 100000
Families which have more than MAXEX haplotype explanation are not used for haplotype frequency estimation. On a unix machine a much higher MAXEX can be used. In order to increase MAXEX, the #define statement in the source code file **famhap18.c** has to be changed and FAMHAP has to be recompiled.

haploview auto

The data of the inputfile is written to a file called ***.haploview**, but with alleles recoded and without column headings. The new file can be analyzed with HAPLOVIEW, FBAT, PLINK and UNPHASED, for instance.

General Options->Produce New Datafiles->haploview format

haploview maxloc <n>

Works like **haploview** , but produces a set of files with <n> neighbouring markers each, in a sliding window fashion. The first file will contain markers 1,2, ..n, the second file will contain markers 2,3,…,n+1 and so on.

General Options->Produce New Datafiles->haploview format
General Options->Produce New Datafiles->windowsize: n

apl

produces a file *_APL.txt that contains the pseudo-controls (nuclear families) that are allowed when conditioning on the sufficient statistics (Horvarth et al. Genet Epidemiol 26:61-69). Please contact us for an explanation of the outputfile. The command is not available on the GUI, because much computer storage is necessary and the option should be run on a unix machine.

19. Citation

Becker T, Knapp M (2004) Maximum-Likelihood Estimation of Haplotype Frequencies in Nuclear Families. *Genet Epidemiol* 27: 21-32 [MAIN REFERENCE: Haplotyping]

Herold C, Becker T (2008) Genetic association analysis with FAMHAP: a major program update. Bioinformatics [MAIN REFERENCE: famhap18, GUI and runFamhap]

Becker T, Knapp M (2004) A powerful strategy to account for multiple testing in the context of haplotype analysis. *Am J Hum Genet 75:561-70* **[Citation: Family-Based Haplotype Association Analysis, MC-Method to Account for Multiple Testing]**

Becker T, Cichon S, Jönson E, Knapp M (2005) Multiple testing in the context of haplotype analysis revisited: application to case-control data. *Ann Hum Genet* 69: *1-10* **[Citation: Case-Control Haplotype Association Analysis, MC-Method to Account for Multiple testing]**

Becker T, Schumacher J, Cichon S, Baur MP, Knapp M (2005) Haplotype Interaction Analysis of Unlinked Regions *Genet Epidemiol 29: 313-322* **[Citation: Simultaneous Analysis of Haplotypes from Different Chromosomes]**

Becker T, Valentonyte R, Croucher PJP, Strauch K, Schreiber S, Hampe J, Knapp M (2006) Identification of probable genotyping errors by consideration of haplotypes. *Eur J Hum Genet 14: 450-8.* **[Citation: Identification of Genotyping Errors]**

Becker T, Baur MP, Knapp M (2006) Detection of Parent-of-Origin Effects Using Haplotype Analysis. *Hum Hered 62:64-76* **[Citation: Imprinting and Association]**

Becker T, Flaquer A, Brockschmidt FF, Herold C, Steffens M (200x) Evaluation of Potential Power Gain with Imputed Genotypes in Genome-Wide Association Studies. *Hum Hered* [Citation: Imputing]