

# **HAPLO-IHP: A Program for Haplotype Reconstruction Using Identified Haplotypes and Haplotype Patterns**

**Yun Joo Yoo, Jianming Tang, Richard Kaslow, Kui Zhang**

## **1. HAPLO-IHP**

HAPLO-IHP is a program to infer haplotypes from unrelated individuals using a set of identified haplotypes and haplotypes patterns from previous studies. It has been shown that the program is effective for genotypes with a substantial amount of missing data. Please refer to Yoo et al. (2007) for the detailed description of methods used in the program. HAPLO-IHP is currently only applicable to bi-allelic markers.

## **2. Files Included in the Package**

HAPLO-IHP.pl: the source Perl script file

HAPLO-IHP.exe: the executive file

HAPLO-IHP-Manual.pdf: the manual for HAPLO-IHP

HAPLO-IHP-Geno.txt: an example file for the genotype data

HAPLO-IHP\_Haplits.txt: an example file for the set of haplotypes

HAPLO-IH-Pattern.txt: an example file for the sets of haplotype patterns

out\_freq\_example.txt: an example file for haplotypes and their frequencies;

out\_best\_example.txt: an example for assigned haplotypes

## **3. How to run HAPLO-IHP**

HaploIHP.exe runs on Microsoft Windows 95, 98, NT, 2000, ME, and XP.

HaploIHP.pl runs on Windows, Linux, and UNIX system if a Perl interpreter is installed.

If you want a Perl interpreter installed, please refer to <http://www.perl.org/get.html> to see how to obtain and use the Perl interpreter in your system.

To run HAPLO-IHP, you can either run the script HaploIHP.pl (if you already installed a Perl interpreter in you system.) or HaploIHP.exe with the options followed by the specified value for each option in the command line:

**HaploIHP.pl -d [genotype data file] -i [initial haplotype file] -s [haplotype pattern**

**file] -t [threshold] -iter [maximum number of iterations] -o1 [frequency output file] -o2 [haplotype assignment output file] -best [output file format (all, one)]**

(If you do not have a Perl interpreter, replace “HaploIHP.pl” with “HaploIHP”. The command line to run the HaploIHP.pl script file for a specific Perl interpreter may require certain command “perl” before you type the script file (e.g.: perl HaploIHP.pl -d example.txt ).

The order of options can be arbitrary. However, you must specify the value for each option right after that option identifier with a space between them. The option “-d” is the only one required by HAPLO-IHP and the others are optional. If you do not specify a value for an operational option, the program will use its default value. The brief description and the default value for each option are listed in Table 1. Using the data files provided in the package, we can infer their haplotypes based on the following command:

**Example 1. Command line example**

**HaploIHP.pl -d HaploIHPsample.txt -i haplist.txt -s sub.txt -t 0.0001 -iter 200 -o1 out\_freq.txt -o2 out\_best.txt -best one**

(If you do not have the Perl interpreter, replace “HaploIHP.pl” with “HaploIHP”).

**Table 1 : The description and the default value for each option**

<b>Option</b>	<b>Description</b>	<b>Default value</b>
-d	[genotype data file]: The name of file for genotype data. This is required by HAPLO-IHP.	No default value
-i	[initial haplotype file]: The name of file for a set of identified haplotypes.	No default value.
-s	[haplotype pattern file]: The name of file for the sets of identified haplotype patterns	No default value
-o1	[frequency output file]: The name of file for estimated haplotypes and their frequencies	[genotype data file]_freq.txt
-o2	[haplotype assignment output file]: The name of file for compatible haplotype pairs and their posterior probabilities for each individual	[genotype data file]_best.txt
-t	[threshold]: The threshold for the convergence and haplotype frequencies. Only haplotypes with frequency greater than this threshold are retained in the final results.	0.00001

-iter	[maximum number of iterations]: The maximum number of iteration that will be performed in HAPLO-IHP	300
-best	[output file format]: There are two formats for the output file for compatible haplotype pairs and their posterior probabilities for each individual. If the value of this option is “all”, then the output file will contain all possible haplotype pairs and their posterior probabilities for each individual. If the value is “one”, then the output file will only contain a most likely haplotype pair and its posterior probability for each individual.	All

#### 4. How to prepare input files (options -d, -i, -s)

HAPLO-IHP can have three different types of input file specified by options -d, -i, -s. These files are [genotype data file], [initial haplotype file], and [haplotype pattern file]. Only [genotype data file] is required by the program and the options -i or -s can be omitted. However, the program might run slower, especially for the data with a large amount of missing.

##### [genotype data file] : -d

This is the file for genotypes. Two alleles at a bi-allelic locus are denoted by 0 and 1, respectively. Missing allele is denoted as ‘?’. If your genotype is present-absent genotype as those described in our paper (Joo et al., 2007) and is denoted by one digit (0 – absence, 1 – presence, and ? – missing) at each locus, you can convert 1 to 1 ?, 0 to 0 0, and ? to ? ? to comply the file format required by HAPLO-IHP. The detailed format of genotype data file is as follows:

**First raw:** the list of marker names in this genotype data file

**Second raw ...:** the individual ID and genotype data

**First column:** the individual ID (an arbitrary character string)

**Second column ...:** Two columns for two alleles at a marker locus. The order of genotypes does not matter. The columns should be separated by the space or the tab.

##### Example 2: [genotype dataset file]

	2DL1		2DL2		2DS1		2DL3		3DS1	
E001	1	0	0	0	1	1	?	?	0	1
E002	0	?	0	1	1	0	1	0	0	0
E003	0	1	1	?	1	1	0	0	1	0
...										

**[initial haplotype file]: -i**

This is the file for a set of identified haplotypes. This list of haplotypes is checked for the compatibility with genotype data in prior to run the greedy algorithm in HAPLO-IHP. Each haplotype is listed in one row as a combination of allele symbols of 0 and 1. The missing allele is not allowed in the file. The first column is the list of marker names and must match the names given in [genotype dataset file]. The detailed format of [initial haplotype file] is as follows:

**First row:** the marker names which match the names given in [genotype dataset file]

**Second row ...:** haplotype and its name

**First column:** the haplotype name (a character string, but can not be a string starting with 'H' or 'C' followed by numbers. This type of string will be use in HAPLO-IHP)

**Second column ...:** each column is for an allele at a marker locus for the haplotype. The columns should be separated by the space or the tab.

**Example 3: [initial haplotype file]**

	2DL1	2DL2	2DS1	2DL3	3DS1
A1	1	0	1	1	0
A2	1	0	0	1	1
A3	1	1	1	0	1
...					

**[haplotype pattern file]: -s**

This is the file for the sets of haplotype patterns. Each set of haplotype patterns consists of a header line to indicate the involved markers and a set of haplotype patterns allowed in the program. For example, if you have a set of haplotype patterns for markers 2DS2 and 2DL2 and have observed that they are in complete negative LD. In other words, you only have two types of haplotypes '1 0' and '0 1' allowed for 2DS2 and 2DL2 but do not

have other two haplotypes ( '0 0' and '1 1'). This set of haplotype patterns can be specified in the file as follows:

```
2DS2 2DL2 2
1 0
0 1
```

In the first row, the first two are the names of two markers and the last number is the number of haplotype patterns allowed in the program for these two markers. The subsequent rows specify the haplotype patterns.

The detailed format of [haplotype pattern file] is as follows;

**First section:** the first set of haplotype patterns

**First row:** the header line

**First column ...:** the name of markers involved

**Last column:** the number of haplotype patterns that are allowed

**Second row ...:** each haplotype pattern is specified in each row;

**Second section ..:** other sets of haplotype patterns

**Example 4: [haplotype pattern file]**

```
3DL3 1
1
2DS2 2DL2
1 1
0 0
3DS1 3DL1 3
0 1
1 0
0 0
```

**Example 5: [haplotype pattern file]**

```
3DL2 2DL4 1
1 1
2DS2 2DL2 2DL3 2
1 1 0
0 0 1
3DS1 3DL1 2
0 1
1 0
```

It is worth noting that all identified haplotypes in initial haplotype file must be compatible with one haplotype pattern in each set of haplotype patterns. For example, the haplotype (1 1 0 1 0 1 0 1) for the markers 3DL2, 2DL1, 2DS2, 2DL2, 2DL3, 2DL4, 3DS1, 3DL1 contradicts with the haplotype pattern in example 5 because the haplotype pattern (1 0 1) for the markers 2DS2, 2DL2, and 2DL3 is not allowed. So this haplotype cannot be used together with the haplotype patterns in example 5. The program checks the contradiction between haplotypes in [initial haplotype file] and haplotype patterns in [haplotype pattern file]. The program will stop and reports an error message when a contradiction is found.

### 5. The Formats and Interpretations of Output Files (-o1, -o2, -best)

Two output files are generated from HAPLO-IHP. Their names can be specified with options -o1 and -o2. Otherwise the program will assign default names to them (See **Table 1**). The output file for individuals' haplotypes have two formats specified with the option -best.

#### [frequency output file] : -o1

This is the file for the estimated haplotypes and their frequencies. A row consists of the name of haplotype, the estimated frequency, and the configuration of the haplotype. The haplotypes given in [initial haplotype file] are listed first using same haplotype names given in [initial haplotype file]. The haplotypes that are not in [initial haplotype file] are given names such as 'H\*' or 'C\*', where '\*' is a number generated during the computation. 'H\*' haplotypes are the haplotypes that can resolve an individual together with a haplotype in [initial haplotypes file]. 'C\*' haplotypes are the haplotypes that can resolve an individual with a haplotype that is not in [initial haplotype file]. An example of [frequency output file] is given in Example 6.

#### Example 6: [frequency output file]

Hap	Freq	3DL3	2DS2	2DL2	2DL3	2DL5B	2DL1	2DL4	3DS1	3DL1	2DL5A	2DS3	2DS5	2DS1	3DL2
1	0.5823	1	0	0	1	0	1	1	0	1	0	0	0	0	1
11	0.0315	1	1	1	0	1	1	1	1	0	0	1	0	0	1
12	0.0153	1	0	0	1	1	1	1	1	0	0	1	0	1	1
16	0.1054	1	1	1	0	0	1	0	0	1	0	0	0	0	1
H1	0.0001	1	1	1	0	0	0	0	0	1	1	0	1	1	1
H10	0.0073	1	1	1	0	0	1	0	0	1	1	0	0	0	1

H12 0.0054 1 1 1 0 0 1 0 0 1 1 1 1 1 1

**[haplotype assignment output file]: -o2**

This is the file for individuals' haplotypes. The haplotype pairs assigned for each individual are reported along with their posterior probabilities. If the value of '-best' option is specified as 'all' (default value), all compatible pairs are produced in this output file. These pairs are displayed in an ascending order according to their posterior probabilities so the first haplotype pair has the highest posterior probability. If the value of '-best' option is specified as 'one', the haplotype pair that has highest posterior probability for each individual is produced in the output file. In the file, the numbers in the second columns are the number of compatible haplotype pairs for each individual. Only the haplotype names are given and the haplotypes can be reconstructed from [frequency output file]. Two examples of [haplotype assignment output file] are given in Example 7 and example 8.

**Example 7 : [frequency output file] when option '-best' is 'all'**

ID	Match#	Pair1	Prob1	Pair2	Prob2	Pair3	Prob3	...
ID1	3	1,1	0.9422	1,H8609	0.0563	H8577,1	0.0015	
ID2	1	H14769,1	1.0000					
ID3	3	1,1	0.9422	1<H8609	0.0563	H8577,1	0.0015	
ID4	1	8,1	1.0000					
ID5	1	1,5	1.0000					
ID6	1	1,H14519	1.0000					

**Example 8 : [frequency output file] when option '-best' is 'one'**

ID	Match#	Pair1	Prob1	...
ID1	3	1,1	0.9422	
ID2	1	H14769,1	1.0000	
ID3	3	1,1	0.9422	
ID4	1	8,1	1.0000	
ID5	1	1,5	1.0000	
ID6	1	1,H14519	1.0000	

**6. Computation Criteria [-t, -iter]**

**[threshold]: -t**

This is a convergence threshold for the EM algorithm. The default value is set as  $10^{-5}$ . Also this threshold is served to discard haplotypes with estimated frequency less this threshold in the EM algorithm.

**[maximum iteration limit]: -iter**

This is a maximum number of iterations allowed in the greedy algorithm and the EM algorithm. The default value is set to 300. You can check the number of iterations performed for the greedy algorithm and the EM algorithm from the standard output (the output on the screen). This default value is large enough in most situations. However, if you have observed that the iterations have reached this number when the program stops, you can increase this number to make sure that the algorithm converges before the iteration reaches the maximum number.

## **7. Remarks**

In this section, we highlight some possible solutions for the problems that may be encountered in running HAPLO-IHP.

HAPLO-IHP can only handle bi-allelic markers and the two alleles at each marker locus should be specified as 1 and 0. You cannot use other allele symbols such as A, G, C, T or other numeric or character symbols.

The first lines in three input files only list the marker names in the file and the order of these names is the order that data are entered. These first lines are not the variable headers. If you have a data file with the first line as column names (variable header), you need to modify the first line. Otherwise, the program will report an error message.

While running the HAPLO-IHP, some important information appears in the screen such as the number of individuals resolved by initial haplotypes set, extended set (first step, and second step), the resolution status with each iteration of greedy algorithm and the number of iterations for the EM algorithm. You may want to monitor this information to check how well your initial sets of haplotypes and haplotype patterns contribute to the inference.

Both [initial haplotype file] and [haplotype pattern file] can help to reduce the computing time. Without one of these input files, the program will run much longer in the presence

of a large of portion of missing data. You can monitor the computing progress through the screen output.

## **8. Contact information**

This program and related materials can be downloaded through following website:

<http://www.soph.uab.edu/Statgenetics/People/KZhang/HAPLO-IHP>.

Bugs and comments should be addressed to:

Kui Zhang

Section of Statistical Genetics

Department of Biostatistics

University of Alabama at Birmingham

1600 University blvd.

[kzhang@ms.soph.uab.edu](mailto:kzhang@ms.soph.uab.edu)

Yun Joo Yoo

Section of Statistical Genetics

Department of Biostatistics

University of Alabama at Birmingham

1600 University blvd.

[YYoo@ms.soph.uab.edu](mailto:YYoo@ms.soph.uab.edu)

or

## **9. Program History**

- February 14, 2007: the program is released.

## **10. References**

Yoo Y, Kaslow R, Tang J, Zhang K. 2007. Haplotype inference for present-absent genotype data for clustered genes using identified haplotypes and haplotype patterns. Manuscript in preparation.