

March 24, 2000

# **TRIMHAP**

## **Version 1.2**

**Software for the trimmed-haplotype test  
of linkage disequilibrium**

# **User Manual**

Charles J. MacLean, Rory B. Martin\* and Huan Wang

Virginia Institute for Psychiatric and Behavioral Genetics  
at Virginia Commonwealth University  
Richmond, VA

Web site for software: <http://vipbg.psi.vcu.edu/trimhap>

\* Current affiliation: Millennium Pharmaceuticals, Cambridge MA

TRIMHAP is a fortran software package that implements the trimmed-haplotype mapping techniques discussed in MacLean et al. [2000]. An earlier version of the program, under the name HAL, was written by Rory Martin at Virginia Commonwealth University.

The latest version of the package can be obtained from web site <http://www.vipbg.vcu.edu/trimhap>. For installation, please refer to the text file INSTALL distributed with the TRIMHAP package.

TRIMHAP can only be understood by reference to MacLean et al. [2000].

## INPUT

TRIMHAP requires haplotype data from general pedigrees. It is based on four user-supplied input files:

- haplotype file
- locus control file
- sex file
- TRIMHAP control file, parameters.dat

The file parameters.dat contains input values for parameters controlling the TRIMHAP analysis. Since many parameters can be input using default values, we suggest that a copy of the example parameters.dat file be used as a template when preparing new analyses, with users changing values only as necessary. The order of parameters as well as the line number on which they appear is significant, and should not be altered.

```
# Debug output level (-1=none, 0=minimal, 1=medium, 2=full)
0
```

Controls the amount of intermediate output generated by TRIMHAP written to file debug.out. Values higher than 0 will cause a significant increase in run-time and may also create large debug.out files, but this output is valuable for tracing errors in TRIMHAP and in input files. A value of -1 suppresses results from all haplotype specific analyses from being written to results.out. [Suggested value for ordinary running: 0].

### Marker map parameters

```
# Total number of marker loci
10
#
# Map (in cM) for marker loci
1 0.2 2 1.5 3 0.2 4 0.2 5 0.2 6 0.2 7 0.2 8 0.2 9 0.2 10
```

Defines the map of marker loci used in the analysis. The format is

```
<marker index> <distance> <marker index> <distance> ... <marker index>
```

as in GENEHUNTER. Inter-marker distances are in cM and must be greater than zero. The number of marker loci must be consistent with the locus control file.

```
# Region of map to be analyzed (In cM)
-20.0 20.0
```

Specifies the part of the map in which TRIMHAP scans for evidence of a disease locus. The user may specify a subset of the map for analysis. For example, 0.0 0.2 would limit the search for the disease locus to the interval between the first two markers in the example above. A large range outside the limits of the marker map, such as -20.0 20.0, simply specifies testing of the entire region. This parameter range affects only test locations for the disease locus, and not the marker loci tested.

```
# Number of test-points per marker-marker subinterval
1
```

Only statistic 3, employing trimming probabilities, is affected by this parameter, because the precise position of the disease locus is used in the trimming probability calculation (see # Which statistic, below and also refer to MacLean et al. [2000] ). For statistics 1 and 2, any position between two markers produces the same result, so this parameter should be set to the minimum, 1. Even with statistic 3, the information about position is probably not useful until the final stages of analysis. Suggested value: 1

```
# Off-end parameter (in cM)
0.0
```

As in GENEHUNTER the parameter controls how far testing continues past the left-hand and right-hand limits of the map. A value 0.0 confines positions for the putative disease susceptibility locus to marker-marker intervals within the putative ancestral haplotype. This is usually best, but tests of single markers require an off-end parameter greater than zero.

### **Ancestral marker parameters**

```
# Number of feasible marker loci
10
#
# List of feasible marker loci
1 2 3 4 5 6 7 8 9 10
```

A marker locus is termed 'feasible' if it can be included in the putative ancestral founder haplotype during analysis. All markers may be defined as feasible, but the user may wish to concentrate on a

particular subset of marker loci; e.g., specifying `2 4 5` limits candidates for ancestral loci to these markers.

This parameter can be used in conjunction with the `Region analyzed` parameter to determine exactly the tests to be performed.

```
# Number of loci in ancestral haplotype
2
```

TRIMHAP tests for linkage disequilibrium using ancestral founder haplotypes of a fixed size. Computational effort is proportional to the number of possible haplotypes, which in turn is proportional to the product of the number of admissible alleles at each marker of the combination. This product rises rapidly with the number of ancestral loci.

```
# Maximum spread of ancestral loci (in cM)
1.0
```

Only marker combinations that have a total span less than or equal to the specified maximum are analyzed. For the marker map above, marker pair 2-3 would give an ancestral haplotype spanning 1.5 cM, which would be rejected. However, other pairs, such as 1-2, or 3-4, would be accepted. Since it is impossible to detect linkage disequilibrium using combinations of markers that span too wide a region, suppression simply avoids spurious, random associations, and it also spares useless computational time. If the parameter `Maximum spread` is set greater than the maximum marker-marker subinterval in the map, all combinations of consecutive feasible marker loci are considered for analysis.

```
# Is this a screen? (True/false)
True
```

A screen is the normal first step of the search for a disease susceptibility locus. Normally, users would begin with a single-locus screen and work their way through analyses involving two loci, three loci, and so on. Since computation time increases radically with haplotype size, screens of more than five markers are probably not practical.

In a screen, all feasible marker loci are employed serially. With three-marker ancestral haplotypes, TRIMHAP would test all haplotypes of markers 1-2-3, 2-3-4, 3-4-5,...8-9-10. For one- or two-marker haplotypes, the screen simply tests all admissible locations in the chromosomal region, producing a map analogous to a multipoint linkage map. However, for haplotypes of three or more markers, conflicts arise, such as 1-2-D-3 and 2-D-3-4, so that the user must make choices.

```
# Ancestral loci (put value=0 to leave locus free to vary)
0 0
```

Ancestral loci must correspond to the number of loci specified above. If the input values are zero, all consecutive loci in the `list of feasible marker loci` above are considered serially. To test specific hypotheses, the value of one or more components may be fixed. Fixing loci save computer time, but it disturbs empirical p-values.

```
# Ancestral haplotype (put value=0 to leave allele free to vary)
0 0
```

Specifies the alleles comprising the ancestral founder haplotype, and must correspond to the number of loci specified above. Normally, the input values are zero, in which case all alleles are considered. Values may be fixed to test specific hypotheses; it probably makes sense to do this only if the ancestral loci have also been fixed.

If `screen` is set to `true`, both parameters `ancestral loci` and `ancestral haplotype` are irrelevant, but they must appear anyway. Just set them to all zeros.

### **Haplotype history parameters**

```
# Number of generations since ancestral mutation event
200
```

Defines the number of generations since the ancestral founder mutation was introduced to the population [MacLean et al. 2000]. This value is only used for calculation of statistic 3, employing trimming probabilities, and has no effect upon statistics 1 or 2.

```
# Mutation rate per generation
1.0e-5
```

The rate is assumed the same for all markers in the study. The mutation rate is accumulated over the number of generations since ancestral founder.

```
# Genotyping error rate per marker
.05
```

The error rate is assumed the same for all markers in the study, and can be estimated from the haplotype input data [MacLean et al. 2000]. Genotype errors are usually recorded zero or blank. TRIMHAP adds the accumulated mutation rate and the error rate for the appropriate number of markers for each trimmed-haplotype category. This value is used only for calculation of statistic 3, employing trimming probabilities, and has no effect upon statistics 1 or 2.

### **Study size parameters**

```
# No analysis, just list & count feasible configurations
false
```

May be set to `true` if the user wishes to quickly estimate the computational complexity of a given analysis. In this case, TRIMHAP will output a list of ancestral haplotypes to be tested, but no analysis will take place.

```
# Total number of pedigrees to be analyzed
100
```

The number of pedigrees may be set to a smaller number (say, 2 or 3) for debugging purposes. Any value equal to or larger than the actual number of pedigrees in the haplotype file will cause all pedigrees to be analyzed.

```
# Number of replications for calculation of significance levels
1000
```

The number of replicate samples generated by permutation bootstrapping. If the value exceeds the array size parameter `max_replicates` declared in `half` (currently 10,000), a warning will be written to `debug.out` and the analysis will terminate. In this case, re-compilation with a larger array size is necessary. Computational effort is proportional to the number of replications. [Suggested value: 1000].

```
# Minimum frequency for allele to be in ancestral haplotype
0.1
```

This limits the alleles considered for selection in the ancestral founder haplotype. Only alleles having a frequency larger than the given value are considered for selection. In general, the analysis will be quicker if relatively infrequent alleles are excluded, but potentially interesting haplotypes may be skipped with too coarse a filter.

```
# Minimum number of trans. to aff. indiv's for test sample
2
```

Controls the selection of the test subsample of haplotypes. In a given pedigree, only founder haplotypes transmitted to at least this number of affected individuals are included in the test subsample.

```
# Maximum trans. to aff. indivs for controls (neg for all)
1
```

Controls the selection of the control subsample of haplotypes. In a given pedigree, only founder haplotypes transmitted to no more than this number of affected individuals are included in the control subsample. If the value is negative, all pedigree-founder haplotypes are included in the control subsample. Several assignment schemes are discussed in MacLean et al. [2000].

```
# Use HBPPL (true) or multinomial class frequencies? (false)
true
```

Haplotype Based Posterior Probability of Linkage may be used to measure the within-pedigree relationship between haplotypes and affection [MacLean et al. 2000]. The alternative is to score every haplotype as 1.0. Numerical simulations indicate that significantly more power is achieved when appropriate HBPPL values are used, but they depend on a realistic segregation model. See Martin et al. [1999] for further details.

```
# Maximum number of D alleles among a pedigree's founders
4
```

When calculating prior and likelihood terms for HBPPL for a given pedigree, high-risk disease alleles D are allocated to a subset of pedigree founder haplotypes in every possible disease genotype configuration. Since it is highly unlikely that there are a large number of independent copies of the disease mutation D segregating in a pedigree, computational effort can be reduced with little or no significant loss of information if calculations are truncated at some point. The value may be set to a large number (say, 20) to disable this option. [Suggested value: 4]

```
# Maximum null alleles allowable in founder haplotype
4
```

Eliminates degenerate pedigree-founder haplotypes from analysis. Pedigree-founder haplotypes having a large number of missing alleles should be excluded from analysis because they interfere with the TRIMHAP permutation scheme for constructing significance levels. [Suggested value: 4]

### **Heterogeneity**

```
# alpha: proportion of linked founder haps
1.0
```

Specifies the locus heterogeneity parameter in the sample being studied. Note that the frequency is per haplotype, not per pedigree as in the linkage admixture model. Therefore, if linkage analysis estimates that 20% of families segregate a particular disease susceptibility locus, and there are an average of four independent haplotypes per pedigree, then  $\alpha = .05$ . On the other hand, we may assume that the linked haplotypes will be concentrated in the test subsample. Thus, if only one third of sample haplotypes fall into the test subsample,  $\alpha$  should be increased to .15.

Since computational effort increases significantly for  $\alpha < 1$ , due to the fact that pedigree likelihoods must be recalculated assuming  $\alpha = .5$ , a value of  $\alpha = 1$  together with a concomitantly reduced value of  $\gamma$  is more efficient.

```
# gamma: proportion of alpha descended from given ancestor
0.2
```

Allelic heterogeneity. The proportion of haplotypes descended from a given ancestral founder haplotype is given by the product  $\alpha \times \gamma$ . Note that this product affects only statistic 3, employing trimming probabilities. If we artificially set  $\alpha = 1$  to decrease run-time,  $\gamma$  should be reduced accordingly so that the product is held at the appropriate value. For example, a sample in which 10% of haplotypes are ancestrally-derived could be described using  $\alpha = 0.5$  and  $\gamma = 0.2$  or  $\alpha = 1$  and  $\gamma = 0.1$ , but the latter would be much more efficient for TRIMHAP.

### **Trimmed-haplotype statistic**

```
# Which statistic: 1 = est, 2 = regress, 3 = trim prob
```

More than one formulation of the trimmed-haplotype test is possible. The most familiar form for a likelihood ratio test uses category frequencies estimated from the data themselves. We call this general purpose LR value statistic 1. Statistic 1 does not explicitly employ the model on which trimmed-haplotype analysis is based. To exploit the model of the trimming process, we may employ it to generate the expected value under the alternative hypothesis in a value called statistic 3. Using the trimming probability defined in MacLean et al. [2000], we calculate a category similarity score that measures similarity between haplotypes in each category and their putative ancestor. In cases where we do not have enough information to perform the trimming probability calculations, we may perform a regression analysis between the category position in the trimmed-haplotype table and the proportion of test versus control subsamples. See Martin et al. [1999] for full details.

### Input files

```
# Filenames: haplotypes / allele frequencies
haplo.dump
dom.loc
#
# Is there a sex file?
True
#
# sex filename, if any. If not, just skip.
families.pre
```

Input files should be listed in the order indicated, one name per line. The haplotype file is assumed to be in the format of a Genehunter haplo.dump output file [Kruglyak et al. 1996]. The locus control file contains marker allele frequencies and penetrances for the disease locus model. It can be created by the Preplink utility (cf. [Terwilliger and Ott 1994]). The sex file is in Pre-madeup format [Terwilliger and Ott 1994] and is used by TRIMHAP to determine the sex of pedigree members when sex-specific penetrances are used, since this information is not contained in the Genehunter haplotype file. If HBPL is not calculated, or if penetrances are equal for both sexes, families.pre may be skipped.

```
# Seed for random number generator
2001
```

A starting value between 1 and 30,000 is required for the pseudo-random number generator. We suggest users change this value for each separate analysis.

### Array Sizes

The following array sizes are declared in TRIMHAP near the head of the main routine, hal.f

- `max_indiv = 40`: number of individuals in a single pedigree
- `max_loci = 25`: number of marker loci

- `max_alleles = 30`: number of alleles at a marker locus
- `max_ped = 1000`: number of pedigrees in the sample
- `max_ad_haps = 4000`: number of pedigree-founder haplotypes
- `max_replicates = 10000`: number of permutation replicates
- `max_depth = 4`: number of generations in a pedigree
- `max_dis_pos = 10 * max_loci`: number of test-points for the disease locus

Alternative values may be defined by the user, as needed.

## Running TRIMHAP

To run TRIMHAP, type in the program name without arguments, at the command prompt:

```
trimhap
```

## OUTPUT

Four output files contain the results of a TRIMHAP analysis. They are:

- `results.out`: output for each haplotype-specific analysis plus scan-wise summary
- `best.out`: allele names and locations, sorted by significance level.
- `map.out`: map-wise significance levels
- `debug.out`: intermediate output with warning and error messages

### Haplotype specific statistics: results.out file

The `results.out` file contains full details of results from each specific ancestral haplotype analyzed by TRIMHAP. There may be many thousands of such haplotype specific analyses, so in general the volume of output contained in the `results.out` file will be overwhelming. For this reason and because of the obvious problem interpreting statistical significance given the multiple tests being conducted, we recommend using the map-wise statistics summarized in the `map.out` file. However, once a region has been identified as being of unusual significance with respect to disease location, the user may wish to examine the output contained in `results.out` for details concerning specific ancestral loci and alleles that may be involved. The output for each haplotype specific analysis is as follows.

	Disease position	Ancestral loci	Ancestral haps
Scan to date:	1 / 1	1 / 1	1 / 1
Current disease position:		1 / 1	1 / 1
Current anc loci:			1 / 1

The first four lines of each haplotype-specific analysis contain a summary of information about disease location and ancestral loci and alleles, and the number of analyses performed. These values are cumulative over an entire test run.

```

          1   2   3   4   5   6   7   8   9  10  11
Ancestral:  -----X-----
Loci        =1= =2= =3=
Hap         =5= =4= =1=
Allele freqs .25 .20 .15

```

The current hypothesis being tested is represented pictorially as a map, with the location of the disease gene given as --X--. The first ancestral locus refers to marker locus 2, while the second refers to marker locus 3, and so on. The notation =1= means that the first locus was fixed by the user, not free to vary over the test. The same convention is used for haplotypes.

The next few lines of output summarize TRIMHAP control parameters.

```

Disease      min      Span      min      max
position     freq      observ  max      test     control
0.1250      .100      0.500   8.0000   2        1

duration     alpha     gamma
200          1.00     0.20

mutation rate per generation = 1.000E-05
total error+mutation rate =   2.001E-03

```

```

Number of replicates (feas/ total) = 1000 / 1000

```

The disease position is 0.125cM to the right of the first marker locus, the minimum admissible allele frequency set by the user is 0.100. The current ancestral loci span, 0.500 cM, is less than the maximum admissible spread of ancestral loci set by the user. The minimum number of transmissions for a test haplotype and the maximum number of transmissions for a control haplotype set by the user are listed. The number of generations since introduction of the mutation, the locus heterogeneity parameter, alpha, and the allelic heterogeneity parameter, gamma, all set by the user are listed, as well as mutation rate and genotyping error rate. Finally, the number of permutation replicates that produced feasible haplotypes is shown together with the total replications specified by the user.

The next block of output for the haplotype specific analysis summarizes the trimmed-haplotype table.

lh	rh	t	c	t	c	trim	contrib	non-t	excess
1	2	4	3	.0192	.0156	.1058	1.36	3.2	0.8
1	1	12	9	.0577	.0469	.0828	1.20	9.8	2.2
0	2	6	13	.0288	.0677	.0730	-0.40	-9.4	-8.1
1	0	40	37	.1923	.1927	.1821	0.01	-3.9	-0.1
0	1	25	28	.1202	.1458	.1258	0.85	-8.2	-5.3
0	0	121	102	.5817	.5313	.4306	-2.24	0.0	10.5
400		208	192						

Going from left to right, we have the number of alleles in common with the founder haplotype on the

left- and right-hand sides of the current disease locus test-location. Next are the number of haplotypes in the test and control subsamples, and the corresponding frequencies. Theoretical haplotype trimming probabilities provide the basis for the contribution to statistic 3, only. The excess category frequency of test over control subsamples, adjusted to the total number of haplotypes in the test subsample, is used to demonstrate the familiar fit of test subsample data to the null hypothesis, although TRIMHAP does not employ this method to calculate the significance value (see MacLean et al. [2000] for a full description of empirical significance levels) . Finally the estimated category frequencies for the test sub-sample against the trimming probability, to demonstrate the fit of statistic 3 to the alternative hypothesis. At the bottom of the table, we see that 400 pedigree-founder haplotypes are admissible for the current combination of ancestral loci. This may vary slightly from test to test, due to missing data and ambiguities in tracing within-pedigree inheritance. These locally-admissible haplotypes are comprised of 208 test haplotypes and 192 controls. Note that in this case the test and control subsamples are disjoint, but this need not be so.

```

statistic used:      3
observed:      raw,  normed =      -7.151,      0.147
replicates:    mean, std dev =      -7.645,      3.360
empirical p-value =      0.3880

```

Output includes estimated significance level, raw and normed values for the statistic used, and the mean and standard deviation in the permutation replications. See MacLean et al. [2000] for a full description.

### **Scan-wise statistics: results.out file**

At the very end of the results.out output file, significance levels are given for scan-wise results.

```

Summary over all scanned configurations:
Number of replicates =      1000
Maximum stat(trim prob) =      0.147      empirical p-val =      0.4740

```

The statistic represents compound hypothesis testing evidence for a disease locus located anywhere within the region analyzed. See MacLean et al. [2000] for further details.

## Summary of all calculations: best.out file

```
# Created on Tue Nov 30 13:51:04 1999
# File: best.out
Statistic calculated 3

total haps tested 1532
rank config stat empir disease
p_val locat ancestral founder
1 2 3 4 5 . . .
-----
1 1025 2.245 0.0240 2 3 1 5
2 47 1.896 0.0380 1 2 3 1
3 115 1.804 0.0490 2 2 3 1
4 1027 1.635 0.0660 3 3 1 5
5 214 1.433 0.0810 1 4 2 1
6 13 1.403 0.0870 2 1 3 1
7 820 1.289 0.1060 2 1 4 3
8 1026 1.160 0.1250 3 3 2 5
. . .
```

Every haplotype-specific test of a TRIMHAP run is listed in the best.out file, sorted by the value of the user-chosen statistic. Note that the corresponding empirical p-values are haplotype-specific values, not adjusted for multiple testing.

The haplotype is identified by its position in the map ( markers 2-3-4 for the first entry), its allele numbers ( alleles 3-1-5 in the first entry), and the location of the putative disease susceptibility locus within it. The entry under disease locat refers to the interval following the marker number given.

The entry config indicates the position of the detailed analysis in the results.out file, where it corresponds to Scan to date: Ancestral haps.

## Map.out file

A list of map-wise significance levels are listed in the map.out file, one for each test-point for the disease locus:

```
# File: map.out
# Created on Tue Nov 30 13:16:55 1999
# Statistic used 3

point cM stat
1 0.125 0.3880
2 0.375 0.5840
. . .
```

The index of the test-point for the disease locus, the location in cM of the test-point from the start of the map, and map-wise significance levels are listed. Each map-wise significance level is a compound hypothesis of multiple haplotype specific analyses, with small values suggesting that the

disease locus may be located nearby. Data in the map.out file may be plotted using the interactive plotting program Gnuplot. This package is commonly installed on many unix (and non-unix) systems, but if not, it is freely available via anonymous ftp at <ftp://ftp.gnuplot.vt.edu/pub/gnuplot/gnuplot-3.7.tar.gz>.

### **Intermediate output, warnings, and errors: debug.out file**

The debug.out file contains intermediate output from TRIMHAP, together with any warning or error messages which are generated. The first section at the head of the file is an echo of the parameters as read from parameters.dat. There is no need to inspect debug.out unless TRIMHAP terminates prematurely due to a fatal error. In this case, a short explanatory message should appear at the end of debug.out. Note: in our experience, most errors are caused by improperly-formatted input data in the parameters.dat file. In case of problems, users should carefully check data in parameters.dat versus the echo of this data printed at the start of debug.out.

### **Bugs**

Although we have made every effort to test this package, bugs in the code may still exist. Users can reach the authors using the contact information given below, preferably by email. In reporting bugs, please include details concerning computer hardware and operating system version, together with input and output files if possible. It is our experience that most failures result from faulty inputs. Before you report a bug, we ask that you check to ensure the failure is not due to wrongly-formatted input data.

Contact address for scientific questions:

Charles MacLean  
Virginia Institute for Psychiatric and Behavioral Genetics  
Virginia Commonwealth University  
Box 980126, Richmond, VA 23298  
email: [cmaclean@bara.psi.vcu.edu](mailto:cmaclean@bara.psi.vcu.edu)

Contact address for computer questions:

Huan Wang  
Virginia Institute for Psychiatric and Behavioral Genetics  
Virginia Commonwealth University  
Box 980126, Richmond, VA 23298  
email: [huwang@hsc.vcu.edu](mailto:huwang@hsc.vcu.edu)

## References

- MacLean CJ, Martin RB, Sham PC, Wang H, Straub RE, Kendler KS (2000) The trimmed-haplotype test for linkage disequilibrium. **American Journal of Human Genetics** 66:1062-1075
- Martin RB, MacLean CJ, Sham PC, Straub RE, Kendler KS (1999) Tests for linkage disequilibrium: haplotypes, multiplex pedigrees, and complex traits. **Human Heredity** Under review
- Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. **American Journal of Human Genetics** 58: 1347-1363
- Terwilliger JD, Ott J (1994) **Handbook of Human Genetic Linkage**. Baltimore, John Hopkins University Press