

NAME

hapstat – statistical analysis of haplotype-disease association

SYNOPSIS

hapstat *data spec* [-o *result*] [-h, --hardy-weinberg=*HW*] [-i, --iterations=*N*] [-m, --model=*M*] [-t, --tolerance=*TOL*]

DESCRIPTION

hapstat is a command-line program for the statistical analysis of haplotype-disease association in case-control studies. **hapstat** allows the user to estimate or test haplotype effects and haplotype-environment interactions by maximizing the (observed-data) likelihood that properly accounts for phase uncertainty and study design.

The first argument to **hapstat**, *data*, must be the name of the input file or directory. If *data* is a directory, **hapstat** will process all files residing in that directory.

The second argument, *spec*, is a required file specifying the variables **hapstat** should use in its analysis. The format of this file is described in the SPECIFICATION FILE section.

Any options should follow the mandatory arguments. The **mh** options will override any variables defined in the *spec* file.

This program provides a subset of the functionality available in the HAPSTAT 2.0 software interface:

<http://www.bios.unc.edu/~lin/hapstat/download>

Refer to the HAPSTAT 2.0 documentation for any details not discussed here.

OPTIONS**-o result**

Save the output of **hapstat** to *result*. If *result* is a directory, the file *data.out* is saved in the specified directory, where *data* is the name of the input file. The default behavior is to save the output to *data.out* in the working directory.

Note: If *data* is a directory and *result* is **not** a directory, the option is ignored and the default behavior applies.

The following will override values set in the *spec* file.

-h, --hardy-weinberg=HW

Set the Hardy-Weinberg assumption *HW*, where *HW* can be HWE or HWD if the population is in Hardy-Weinberg equilibrium or disequilibrium, respectively. The default is *HW*=HWE.

-i, --iterations=N

Set the maximum iterations *N* when estimating haplotype effects. The default is *N*=500.

-m, --model=M

Set the mode of inheritance *M*, where *M* can be additive, dominant, recessive or codominant. The default is *M*=additive.

-t, --tolerance=TOL

Set the error tolerance *TOL* when estimating haplotype effects. The default is *TOL*=0.0001.

SPECIFICATION FILE

This section describes the format for specifying the variables **hapstat** requires for its analysis. A variable is defined using the syntax

VARIABLE = *value1* [*value2* ...]

Variables indicated as required must be defined in this file. Optional variables assume default values if omitted from the file or are not assigned a value.

Input data

Input files should contain text data in a tabular (row-column) format. Each row contains space or tab delimited data specific to an individual. The file must contain one column describing the disease status of the individual and one or more columns for each multi-SNP gene. Optionally, the file may include one or more columns of environmental covariates. Column titles may be specified in the first line of the file.

Note: Define rows and columns to begin at 1.

HEADER = *hdr*

Set *hdr* to 1 if column titles are specified in the first line of the file, otherwise set *hdr* to 0. The default is *hdr* = 1. **Optional.**

STATUS = *column*

Specify the column in the data file corresponding to the disease status. **Required.**

ENVIRONMENT = [*column1* ...]

Specify zero or more columns in the data file corresponding to the environmental covariates. **Optional.**

GENE = *column1* [*column2* ...]

Specify one or more columns in the data file corresponding to the SNP sites for a particular gene. For multiple-gene analysis, provide a definition for each gene.

At least one GENE definition is **required**.

Assumptions

MODEL = *model*

Set the mode of inheritance, where *model* is one of the following: additive, dominant, recessive or codominant. The default is *model* = additive. **Optional.**

HW = *hw*

Set the Hardy-Weinberg assumption, where *hw* is set to HWE or HWD if the population is in Hardy-Weinberg equilibrium or disequilibrium, respectively. The default is *hw* = HWE. **Optional.**

Convergence criteria

TOLERANCE = *tol*, ITERATIONS = *itr*

The EM and Newton-Raphson algorithms to estimate haplotype effects will terminate when the number of iterations exceeds *itr* or the error between successive iterations is less than *tol*. By default, *tol* = 0.0001 and *itr* = 500. **Optional.**

FTOLERANCE = *ftol*, FITERATIONS = *fitr*

The EM algorithm to estimate haplotype frequencies terminates when the number of iterations exceeds *fitr* or the error between successive iterations is less than *ftol*. By default, *ftol* = 0.000001 and *fitr* = 2000. **Optional.**

Effects selection

SAMPLE = *sample*

hapstat can estimate haplotype frequencies of the combined case-control sample or consider cases and controls separately. The value of *sample* must be one of the following: cases, controls or combined. The default is *sample* = controls. **Optional.**

THRESHOLD = *threshold*

Frequencies are estimated over all genes and haplotypes with frequencies no greater than *threshold* in the joint distribution are excluded from computation. The default is *threshold* = 0.001. **Optional.**

EFFECT = *hap1 ... hapN1, cov1 ... covN2*

An effect is defined by $N = N1 + N2$ integer values, where $N1 = \#genes$ and $N2 = \#covariates$.

Genes are indexed by the order in which the GENE variables are defined. The first value *hap1* indicates the haplotype from Gene 1 to be included in the effect. Haplotypes are specified by the decimal equivalent of the binary representation of the haplotype, for example, 01100 is specified as 12. Set *hap1* to -1 if no haplotype from Gene 1 is included in the effect. Set values *hap2 ... hapN1* in the same manner.

Covariates are indexed in the order they are defined in the ENVIRONMENT variable. Set *cov1* to 1 if the covariate in *column1* is included in the effect. Otherwise, set *cov1* to -1. Values *cov2 ... covN2* are set to 1 or -1 in the same manner.

At least one EFFECT definition is **required**.

EXAMPLES

To process all files in the directory *data/* and save the output files in the directory *result/* using the specification file *spec.txt*,

```
$ hapstat data/ spec.txt -o result/
```

To process the file *data/case-control.dat* and save the output to the directory *result/* using the specification file *spec.txt*,

```
$ hapstat data/case-control.dat spec.txt -o result/
```

where the output is saved to *result/case-control.dat.out*. To save the output to the working directory, omit the **-o** option,

```
$ hapstat data/case-control.dat spec.txt
```

To estimate dominant effects under Hardy-Weinberg disequilibrium,

```
$ hapstat data/case-control.dat spec.txt -m dominant -h HWD
```

BUGS

If you encounter problems, have questions or suggestions for improvement, please contact us at <linsoft@bios.unc.edu>.

AUTHORS

Copyright (c) 2005-2008 Tammy Bailey, Danyu Lin and the University of North Carolina at Chapel Hill, Department of Biostatistics, 3101 McGavran-Greenberg CB#7420, Chapel Hill, North Carolina 27599-7420 USA. All rights reserved.

SEE ALSO

Zeng D, Lin DY, Avery CL, North KE, Bray MS (2006). Efficient semiparametric estimation of haplotype-disease associations in case-cohort and nested case-control studies. *Biostatistics*, 7(3):486-502.

Lin DY and Zeng D (2006). Likelihood-Based Inference on Haplotype Effects in Genetic Association Studies. *Journal of the American Statistical Association*, 101:89-104.

Lin DY, Zeng D, Millikan R (2005). Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. *Genetic Epidemiology*, 29:299-312.

<<http://www.bios.unc.edu/~lin/hapstat>>