HAPSTAT

statistical analysis of haplotype-disease association

Input data

HAPSTAT supports data from cross-sectional, longitudinal, case-control, and cohort (including case-cohort and nested case-control) association studies. The cross-sectional and longitudinal designs collect data on a random sample of individuals. In a cross-sectional study, the response variable is measured only once on all study subjects; in a longitudinal study, the response variable is measured repeatedly through time. In the case-control design, data is collected on a sample of diseased individuals or cases and a sample of disease-free individuals or controls. The cohort design follows a sample of at-risk individuals over time and records their times of disease occurrence. The individuals who are withdrawn prematurely or who are disease-free at the end of the cohort study have censored observations, in that their ages at onset are only known to be beyond their durations of follow-up.

File format

HAPSTAT accepts ANSI-encoded text files containing data in a tabular (row-column) format. Each row contains space or tab delimited data specific to an individual. The file must contain one or more columns representing the multi-SNP genotype. Optionally, the file may include one or more columns of environmental covariates. Column titles may be specified in the first line of the file, although not required, and must also be space or tab delimited. The file may contain columns of extraneous data. There are no requirements on the ordering of the columns. Study-dependent data requirements are as follows:

Cross-sectional

: The file must contain one column describing the trait value of the individual.

Longitudinal

: File data must be tab delimited. The file must contain two columns of data per individual providing an identifier unique to that individual and the observed trait value. For each individual, each observation is recorded on a separate row. At least some individuals have more than one observation. Data that is constant over all observations, such as the multi-SNP genotype, need only be specified for one observation. An identifier must be specified for each row of observed data. There are no requirements on the ordering of the rows.

Case-control

: The file must contain one column describing disease status of the individual.

Cohort

: The file must contain two columns of data per individual providing the observation time and event indicator.

Data specification

The multi-SNP genotype is represented by a sequence of the values 0, 1 and 2, corresponding to the number of occurrences of a specific allele at each SNP site. Any value other than 0, 1 or 2 is assumed to indicate missing SNP data. Individuals are allowed to have missing data at all SNP sites. Environmental covariates are represented by decimal values and may not contain missing values. The representation of data not intended

for analysis is unimportant.

In cross-sectional studies, disease-related traits are represented by decimal values. In longitudinal studies, the identifier is represented as a string value and disease-related traits are represented by decimal values. In case-control studies, the disease status is specified by 1 for cases or 0 for controls. In cohort studies, decimal values represent the observation times and a binary event indicator distinguishes between uncensored and censored individuals by the values 1 and 0, respectively.

The file case-control.dat, shown below, contains simulated data for a case-control study of 2000 individuals genotyped at five SNPs, where some SNP values are missing.

Status	Age	Gender	SNP1	SNP2	SNP3	SNP4	SNP5
1	48	0	2	1	0	2	2
1	49	0	1	2	.	2	.
1	40	0	0	2	2	0	.
1	44	1	.	1	1	1	1
1	24	0	1	1	1	1	1
1	48	1	0	2	1	1	.
1	48	1	2	0	0	.	2
1	36	1	0	2	2	0	0
1	48	1	0	2	2	0	0
1	44	1	0	2	1	1	1
1	22	0	.	2	2	0	0
1	43	0	2	0	2	0	0

case-control.dat: Format of case-control data for HAPSTAT input.

The disease status is specified in the first column, titled "Status". The columns "Age" and "Gender" contain environmental covariate data, and the columns SNP1-SNP5 represent the five SNP sites. The '.' character indicates a missing SNP value.

Data import

To open a file in HAPSTAT, select the menu option *File»Open* and choose the study type corresponding to your data from the submenu. Browse to the directory where your data file resides, select your file and click the *Open* button. The HAPSTAT display after importing case-control.dat is shown in Figure 1.1.

You may only have one file open in HAPSTAT at any time. To open a new file, select the menu option *File»Open* or click the icon  on the toolbar. HAPSTAT will prompt you to save your results before closing the current file. You may also close a file via the menu option *File»Close* or the icon  on the toolbar.

Header data

HAPSTAT will attempt to detect if the first line of your file contains column titles or actual data. You can toggle what HAPSTAT decides by checking/unchecking the menu option *Settings»Include header*.

Variable selection

To specify the columns that correspond to the variables HAPSTAT should use for analysis, first click inside the text area of the variable you wish to set in the *Variables* box on the right panel. Then select the columns of data corresponding to that variable by clicking on the column labels on the left panel. Use the toolbar icons  and  to show or hide unselected columns. The selection of variables for the single-gene analysis of case-control.dat is shown in Figure 1.2. After completing your selection, click *Continue* to

proceed.

Multiple genes

Click the  icon on the toolbar to create multiple genes. Figure 1.3 shows the choices of SNP1-SNP3 and SNP4-SNP5 as Gene 1 and Gene 2, respectively.

Edit/clear selection

To change your variable selections after the *Continue* button is clicked, return to the file tab and click the *Edit* button. The *Settings»Clear* menu option will clear the current selection.

Sorting

Right click on the title of the column you wish to sort by and select *Sort ascending* or *Sort descending*. All columns are sorted accordingly.

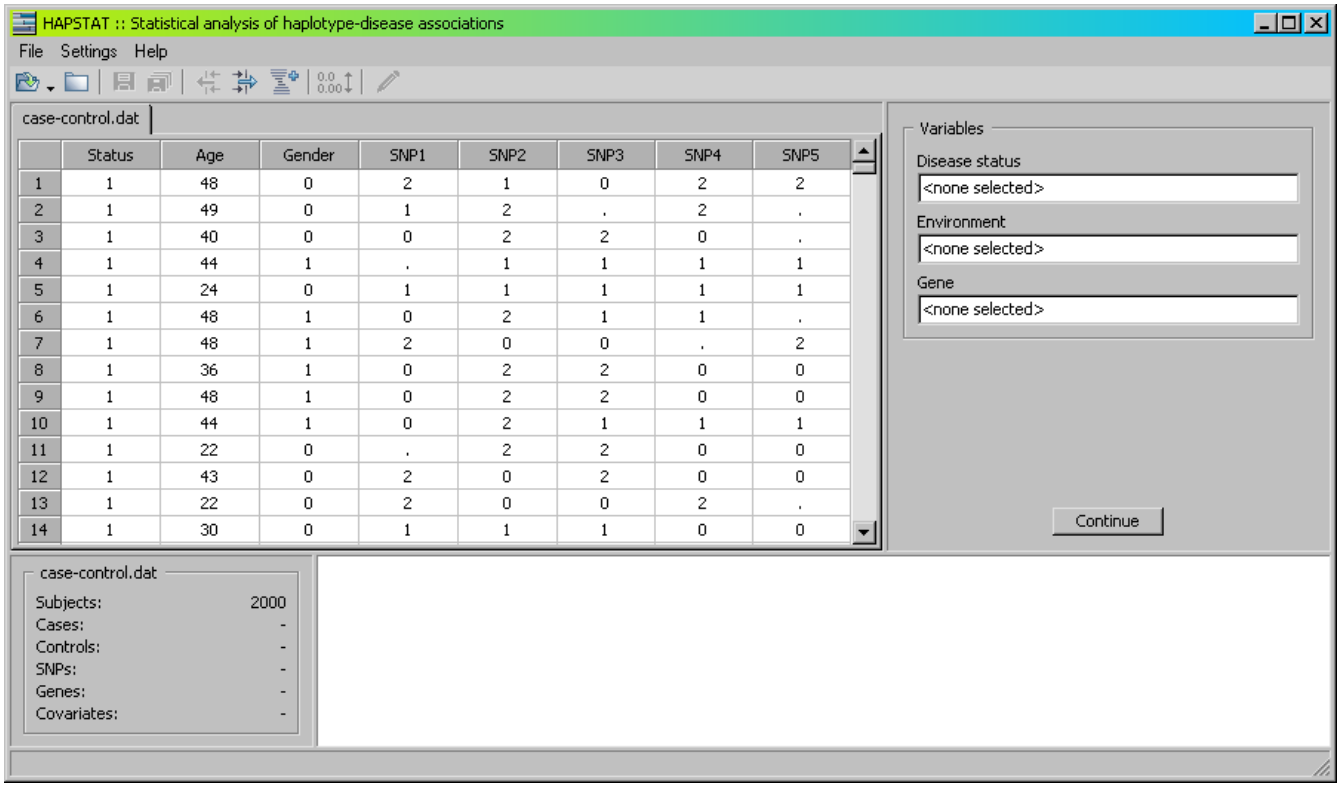

Figure 1.1: Importing a case-control data file.

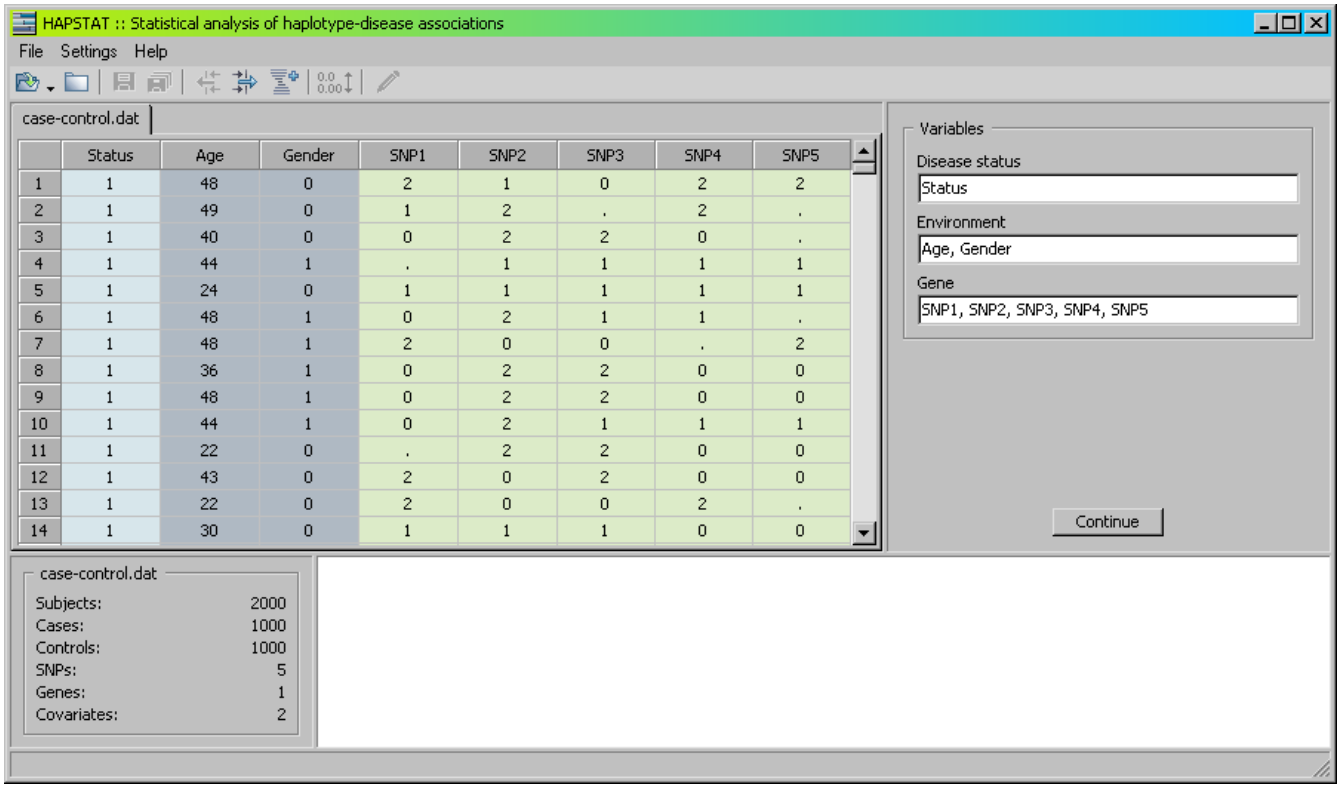

Figure 1.2: Selecting variables for the single-gene analysis.

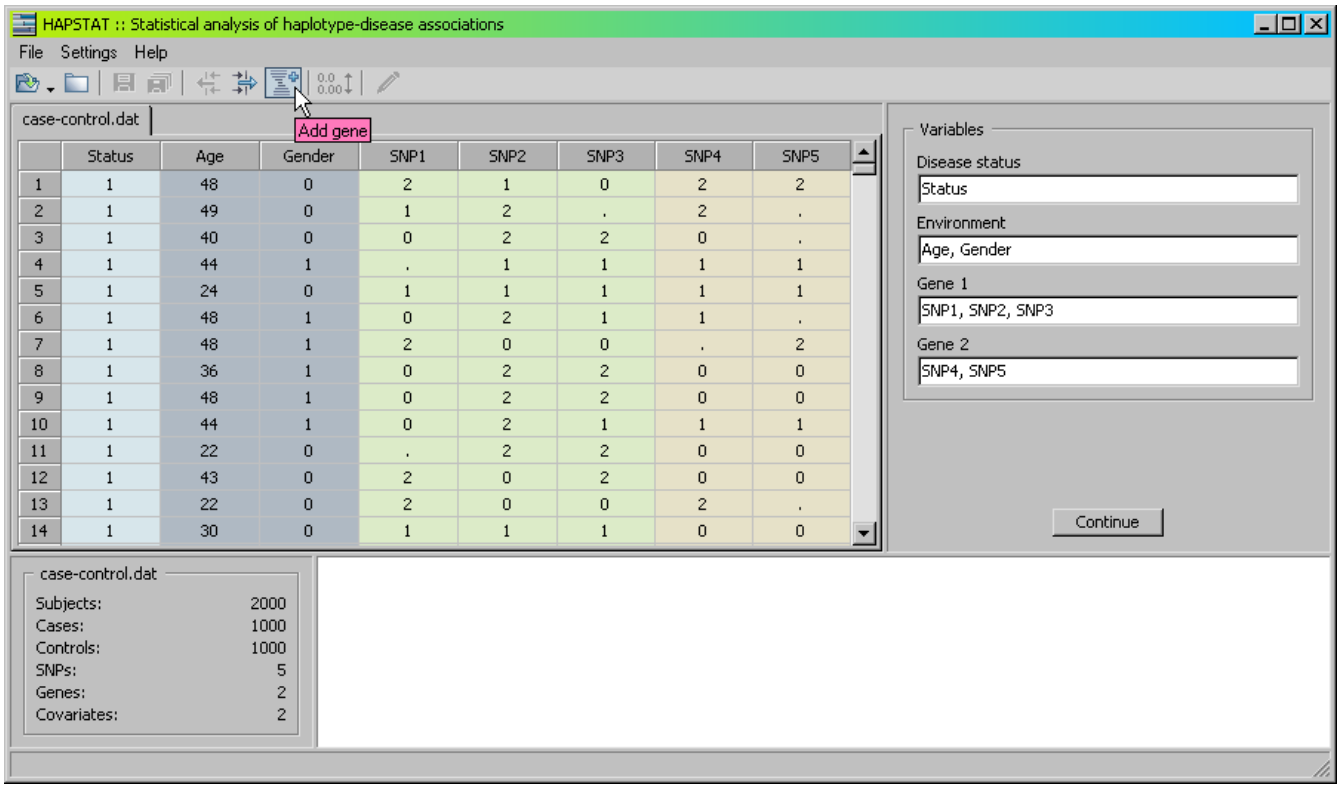

Figure 1.3: Defining multiple genes.

Frequency estimation

Navigation

HAPSTAT estimates the joint haplotype frequencies of all the SNPs included in the analysis. Select the tab labeled *Frequencies* in the left panel; see Figure 2.1. The options available to the user are located in the right panel. After you click on *Calculate*, your results will display on the left.

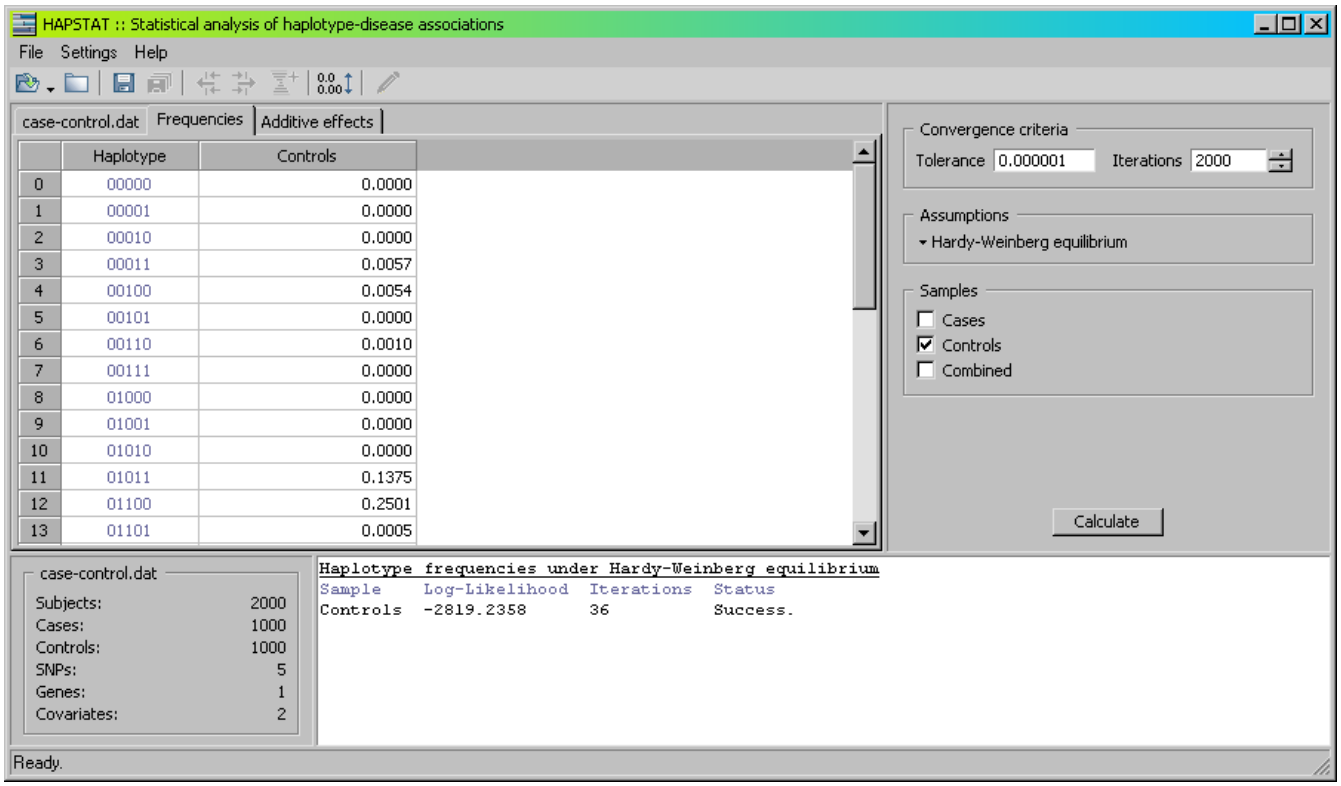

Figure 2.1: Estimating haplotype frequencies under Hardy-Weinberg equilibrium.

Convergence criteria

HAPSTAT uses an EM algorithm to estimate haplotype frequencies. The algorithm terminates when the number of EM iterations exceeds the value specified by *Iterations* or when the change in parameter values between successive iterations satisfies the following inequality:

$$\max_k | \delta_{k,i} | < \varepsilon ,$$

where ε denotes the specified value of *Tolerance*,

$$\delta_{k,i} = \begin{cases} \theta_{k,i} - \theta_{k,i-1} & \text{if } | \theta_{k,i-1} | < 0.01 \\ \dfrac{\theta_{k,i} - \theta_{k,i-1}}{\theta_{k,i-1}} & \text{otherwise} \end{cases}$$

and $\theta_{k,i}$ is the estimate of parameter k at iteration i. By default, $\varepsilon = 10^{-6}$ and the number of iterations is 2000.

Assumptions

Use the dropdown to estimate frequencies assuming the population is in Hardy-Weinberg equilibrium

(default) or disequilibrium. For Hardy-Weinberg disequilibrium, HAPSTAT returns an estimate for the inbreeding coefficient (ρ).

Samples

For cross-sectional and longitudinal studies, HAPSTAT will automatically estimate frequencies based on all individuals. For a case-control study, choose to estimate haplotype frequencies of the combined case-control sample or consider cases and controls separately. The default is the control sample. For a cohort study, check *Cohort* to estimate haplotype frequencies based on all genotyped cohort members. Under case-cohort or nested case-control designs, the genotyped individuals are not representative of the entire cohort. Thus HAPSTAT also estimates haplotype frequencies based on all genotyped controls and a random sample of cases such that the proportion of cases used for estimation is the same as the proportion of controls that are genotyped. This option is referred to as *Subcohort*. Multiple selections are permitted.

Summary

The results of the frequency estimation are summarized in the lower panel of the HAPSTAT display. In the rare event that the computation fails, an error status message is shown. It may then be necessary to increase the maximum iterations or decrease the error tolerance.

Sorting

Right click on the header of the column you wish to sort by and select *Sort ascending* or *Sort descending*. All columns are sorted accordingly.

Filtering

To display frequencies above a certain threshold, right click on the column header and select *Filter*. In the dialog box, specify the desired threshold and frequency sample. Select *Show all* to disable the filter.

Precision

You may change the decimal precision of the displayed frequency values via the menu option *Settings*»*Precision* or the icon  on the toolbar. In the *Precision* dialog box, enter the number of digits to follow the decimal point for fixed notation (default) or the maximum number of significant digits for scientific notation. The default precision is 4.

Saving

To save frequency estimates, select the menu option *File*»*Save* or click the icon  on the toolbar. Navigate to the desired directory and enter a file name or choose an existing one. Overwrite and append options are supported for existing files. Selecting the menu option *File*»*Save All* or the toolbar icon  will save results of all open tabs. HAPSTAT result files are in text format and open with common word processing software.

Effects estimation

HAPSTAT estimates the effects of haplotypes and environmental covariates and haplotype-environment interactions through regression modeling. For quantitative traits, the linear regression model is employed. For binary traits, the logistic regression model is employed, and the regression parameters pertain to the log odds ratios. For age-at-onset data, the Cox proportional hazards model is employed, and the regression parameters pertain to the log hazard ratios. The mode of inheritance can be additive, dominant, recessive or codominant. Under the additive model, having two copies of a causal haplotype has twice the effect on the trait as compared to having a single copy. Under the dominant model, having one or two copies has the same effect. Under the recessive model, only having two copies of the causal haplotype will affect the trait. Under the codominant model, the effect of having two copies can be arbitrarily different from that of having a single copy. In HAPSTAT, the codominant effects are decomposed into additive and recessive components.

Navigation

Estimate haplotype effects by selecting the tab in the left panel labeled *Additive effects*. The options available to the user display in the right panel. The additive genetic model is set by default; changing this setting in the options panel will change the selected tab label accordingly. After you click on *Calculate*, your results will display on the left; see Figure 3.1.

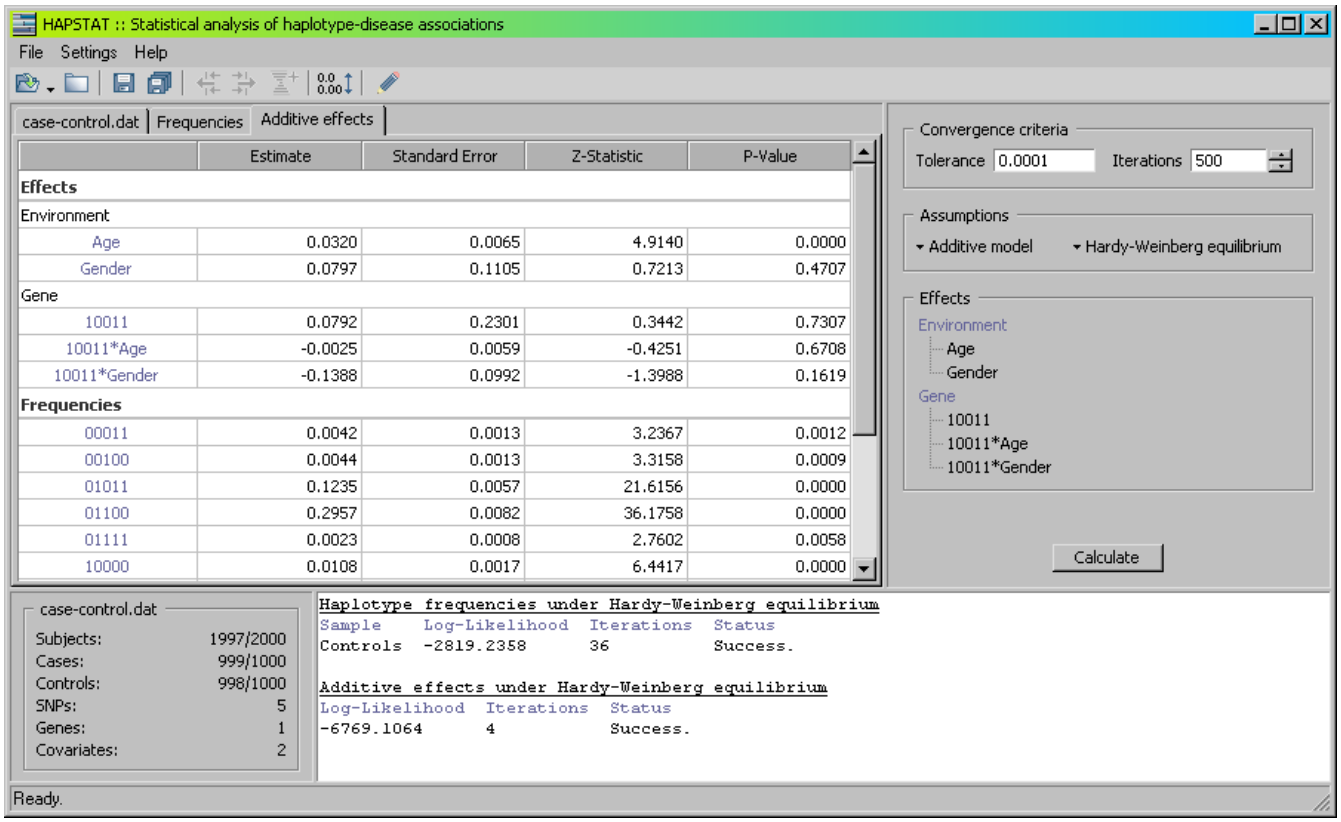

Figure 3.1: Estimating additive haplotype effects in the default setting.

Convergence criteria

HAPSTAT uses the EM and Newton-Raphson algorithms to estimate haplotype effects. The convergence

criteria are the same as those used for the estimation of haplotype frequencies described in the previous section. The maximization is taken over all parameters in the likelihood. The default tolerance is 10^{-4} and the number of iterations is 500.

Assumptions

Select the additive, recessive, dominant or codominant mode of inheritance from the left dropdown. Use the right dropdown to estimate haplotype effects under Hardy-Weinberg equilibrium (default) or disequilibrium. For Hardy-Weinberg disequilibrium, HAPSTAT will return an estimate for the inbreeding coefficient (ρ).

Effects

The box labeled *Effects* is a static display of the main effects and interactions selected for estimation. By default, HAPSTAT selects the haplotype with the highest frequency in the default sample and all covariates, as well as the interactions between them. The selected haplotypes are compared to a reference group, which includes all unselected haplotypes.

Select effects

To change the default selection, click the icon  on the toolbar to activate the *Select effects* dialog, shown in Figure 3.2. The panel labeled *Effects* shows the current selection.

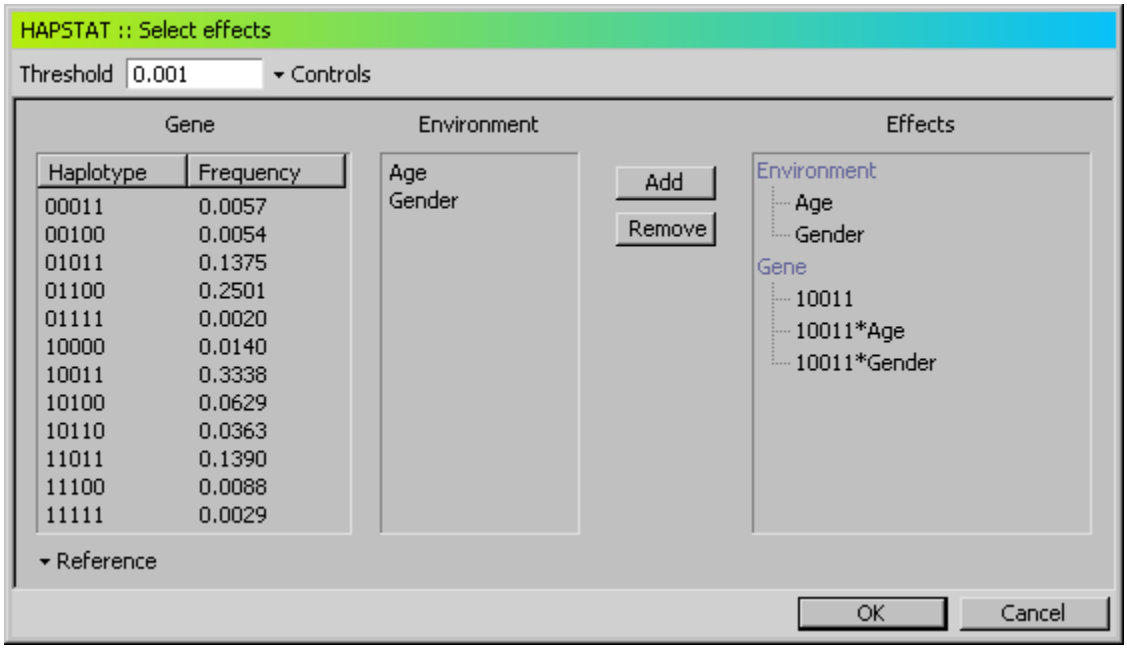

Figure 3.2: The *Select effects* dialog.

The haplotypes whose frequencies are no greater than the value specified by *Threshold* are removed from calculation. For case-control and cohort studies, frequencies are determined by the sample chosen from the adjacent dropdown. The default threshold is given by

$$\max (2/n , 0.001),$$

where n is the total sample size. For case-control studies, the control sample is chosen by default; for cohort studies, the subcohort is the default. The haplotypes above the threshold along with their frequencies are listed in the *Gene* panel. The *Reference* dropdown lists haplotypes whose frequencies are below the threshold

along with those haplotypes that are above the threshold but are not selected for effects estimation. Covariates are listed in the *Environment* panel.

To add a main effect, click on the desired variable in the *Gene* or *Environment* list followed by the *Add* button. The selected variable now appears in the *Effects* panel under the heading *Gene* or *Environment*, respectively. To add an interaction, select the appropriate variables from the *Gene* and *Environment* lists and click the *Add* button. You can select multiple variables from the *Environment* list by using the Shift/Ctrl key. To remove a specific effect from the selection, click on that effect on the *Effects* panel followed by the *Remove* button. Clicking on a heading on the *Effects* panel will remove all associated effects.

In Figure 3.3, we remove the haplotype effect 10011 and the interactions 10011×Age and 10011×Gender by clicking on the *Gene* heading followed by the *Remove* button. Next, we add haplotype effects 01011, 01100 and 11011 and the interactions 01011×Age, 01100×Age, 11011×Age, 01011×Gender, 01100×Gender, 11011×Gender, and 01011×Age×Gender. Figure 3.4 illustrates the addition of interaction 01011×Age×Gender. The results are shown in Figure 3.5.

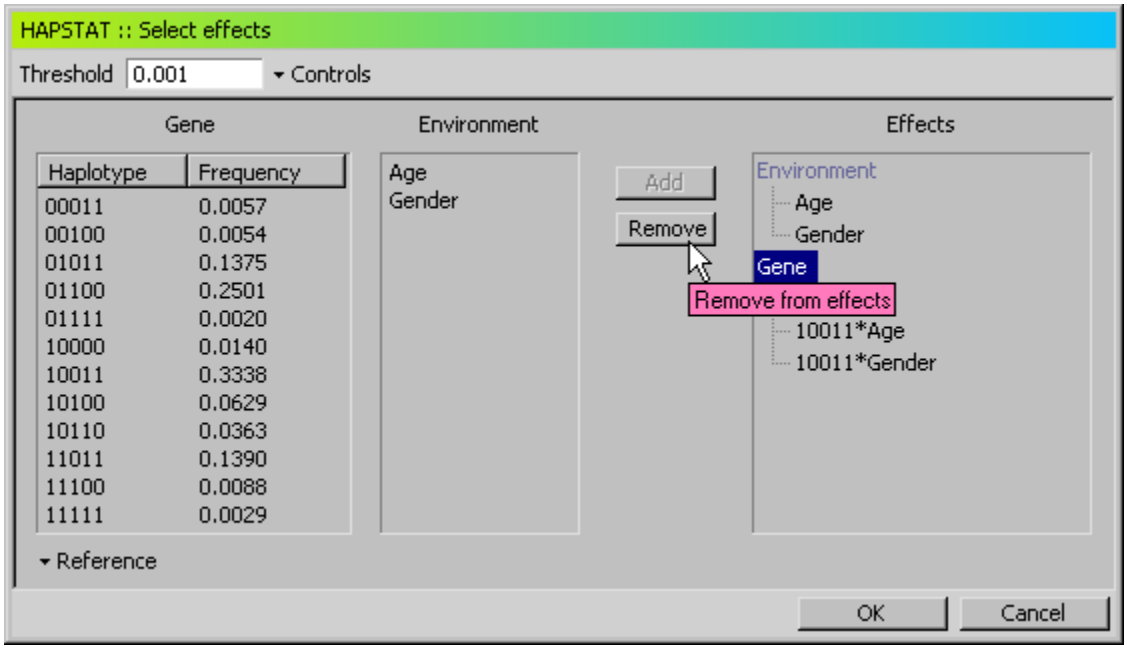

Figure 3.3: Removing effects 10011, 10011×Age and 10011×Gender.

For longitudinal studies, the user can specify both fixed and random effects through the *Select effects* dialog. See the *Longitudinal data* section under *Examples* for further detail.

Multiple genes

Consider the multiple gene selection illustrated in Figure 1.3. Click *Continue* and then click the icon  to activate the *Select effects* dialog. Frequencies are estimated over all genes and haplotypes with frequencies no greater than the *Threshold* in the joint distribution are excluded from computation. For each gene, haplotypes and their frequencies from the marginal distribution are listed in the corresponding *Gene* panel. In the *Select effects* dialog, select haplotype 100 from Gene A and haplotype 11 from Gene B to add the gene-gene interaction 100×11 to the *Effects* selection. Clicking *Calculate* gives the result in Figure 3.6.

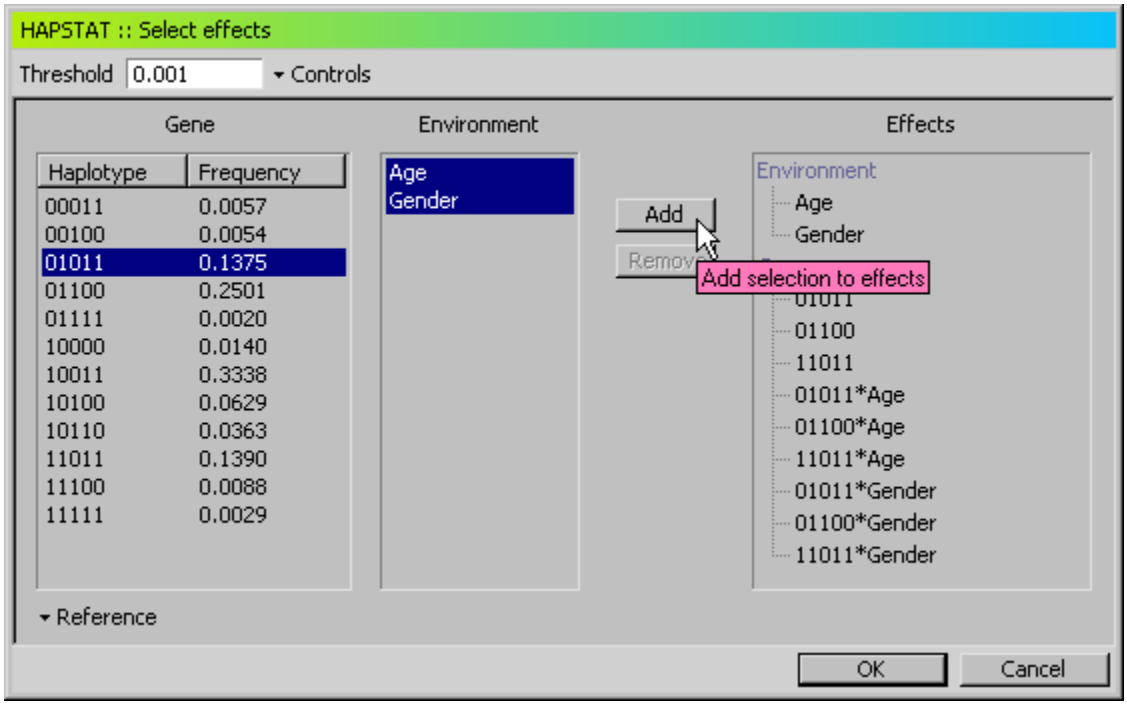

Figure 3.4: Adding interaction 01011×Age×Gender.

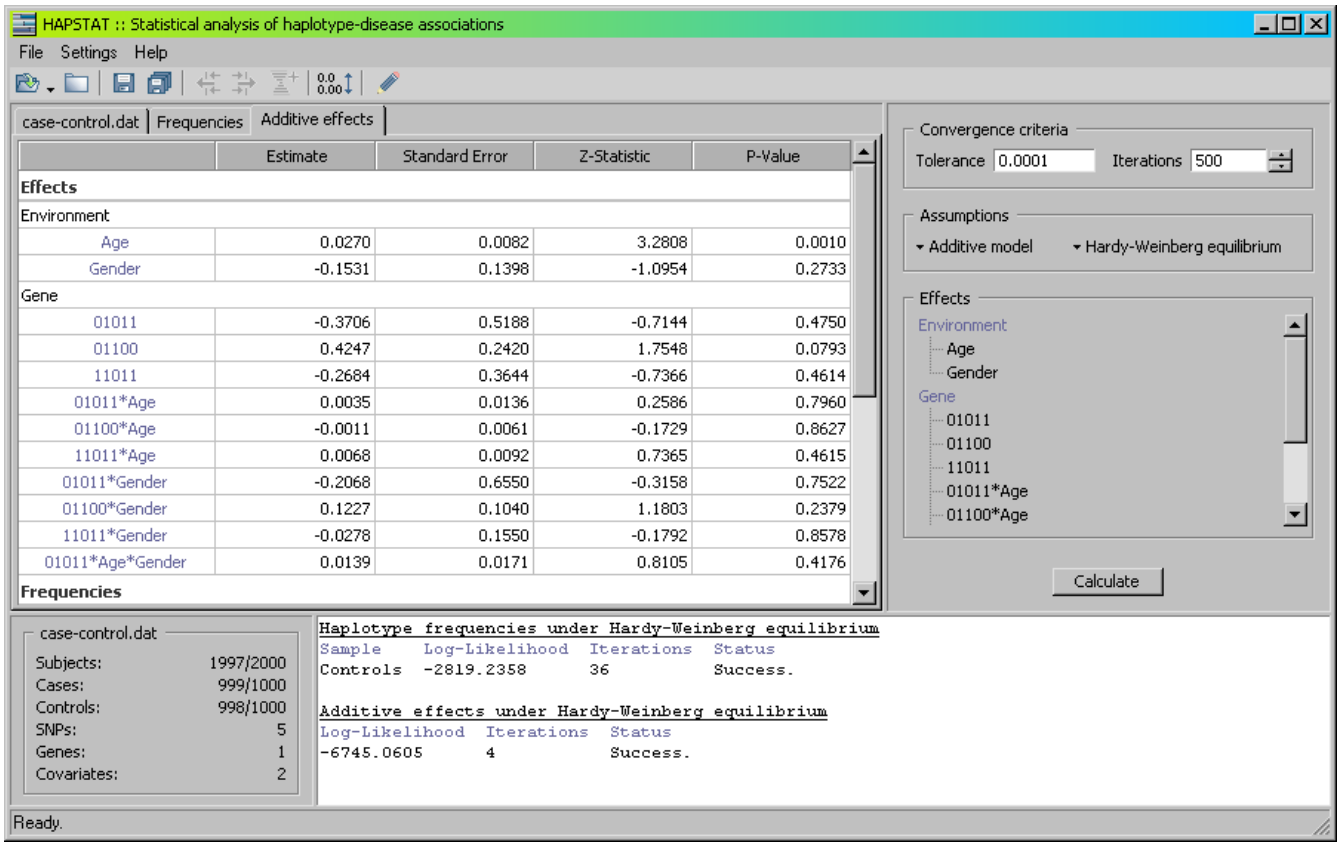

Figure 3.5: Estimating additive haplotype effects after selection of effects.

SNP Analysis

HAPSTAT can be used to analyze the effects of individual SNPs by treating each SNP as a separate gene. By using the linkage disequilibrium information of multiple SNPs to infer the missing SNP values, HAPSTAT provides efficient estimation of SNP effects in the presence of missing data. Figure 3.7 and Figure 3.8 show the estimation results from two models, one including all the five SNPs and one including only SNP4.

Summary

In the left panel, HAPSTAT displays the estimates of regression parameters and their standard errors, together with the Wald statistics and two-sided p-values. The lower panel displays the log-likelihood value(s). You can calculate the likelihood ratio statistic to test a set of parameters by fitting the models with and without the set of parameters of interest.

Precision

You may change the decimal precision of the displayed results via the menu option *Settings»Precision* or the icon  on the toolbar. To change the decimal precision for an individual column, right-click on the column header and select *Precision* from the drop-down menu. In the *Precision* dialog box, enter the number of digits to follow the decimal point for fixed notation (default) or the maximum number of significant digits for scientific notation. The default precision is 4.

Saving

Select the menu option *File»Save* to save the effects estimates. To save both frequency and effect estimates, select the menu option *File»Save All* or click the icon  on the toolbar. The results for the case-control data using the options shown in Figures 3.1-3.8 are given in case-control.out

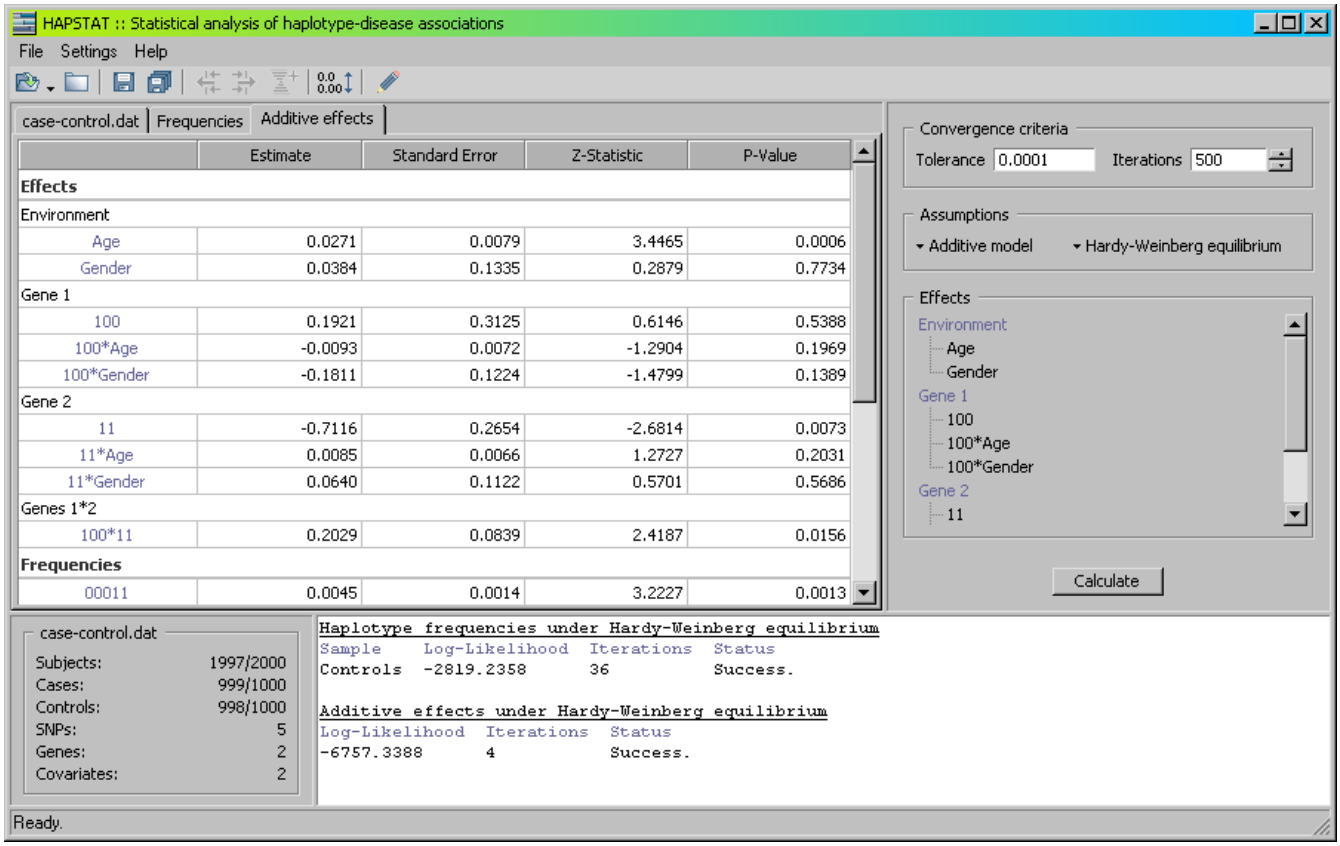

Figure 3.6: Estimating additive haplotype effects of multiple genes.

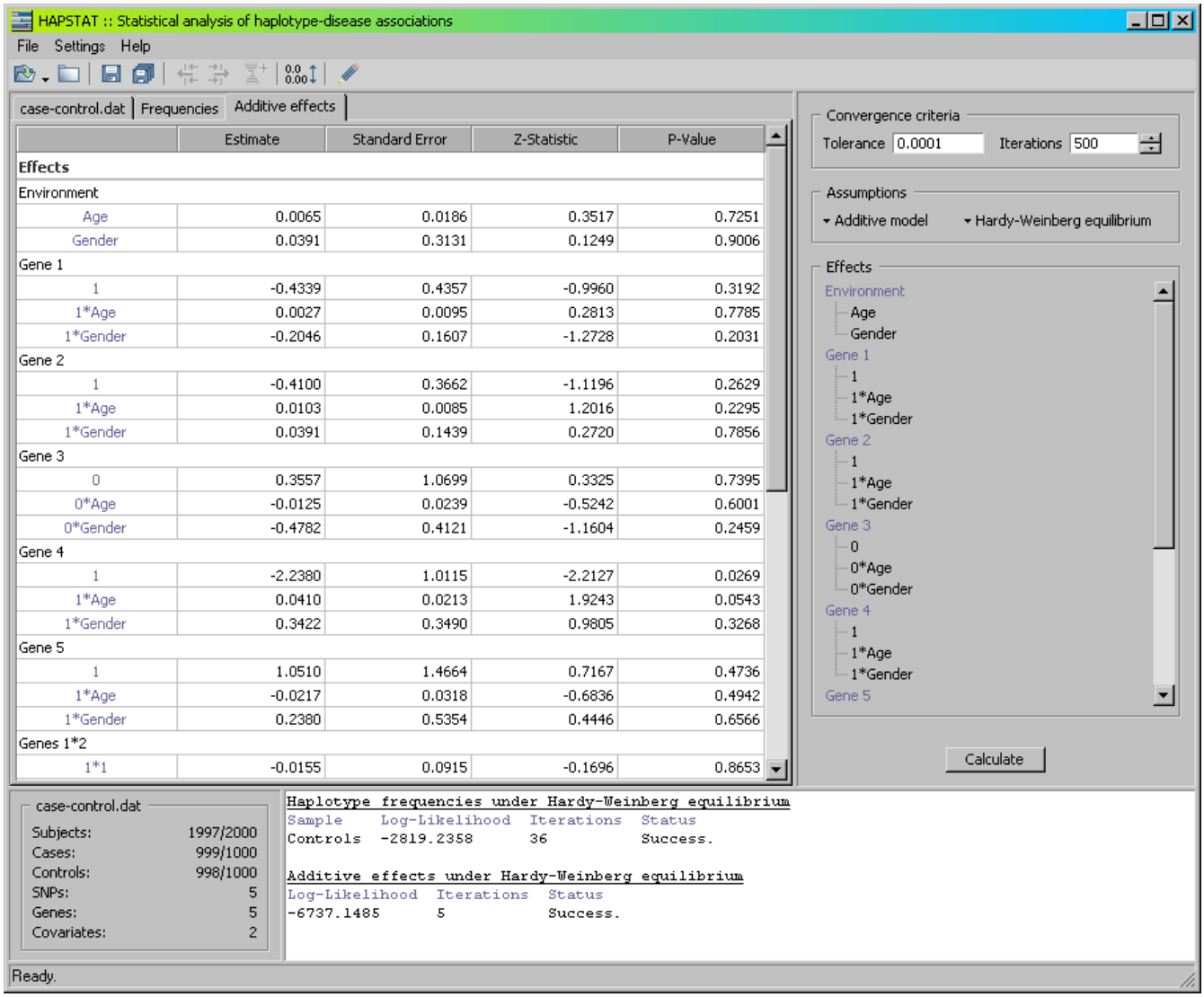

Figure 3.7: Estimating the effects of multiple SNPs.

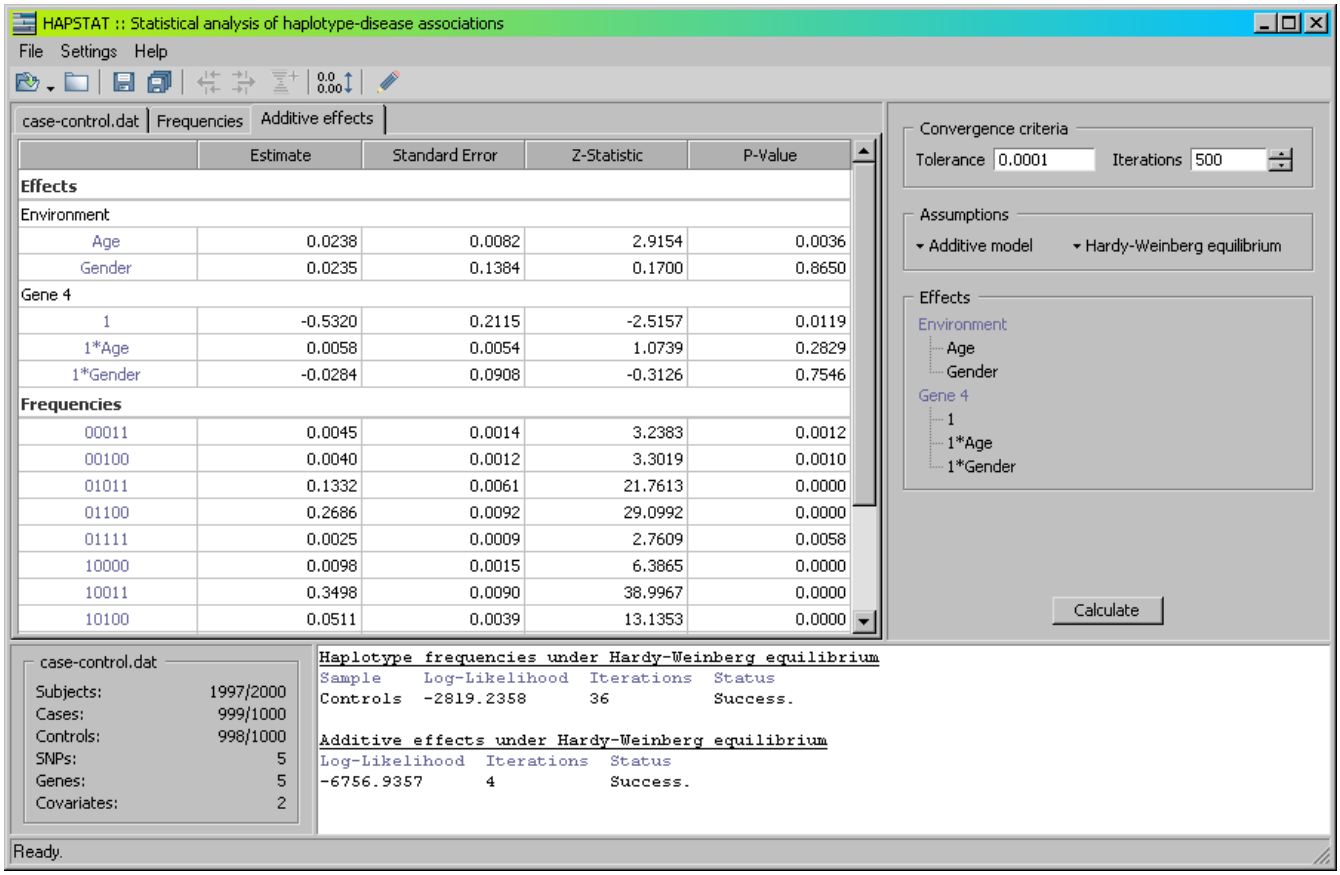

Figure 3.8: Estimating the effects of a single SNP.

Examples

Cohort data

The file cohort.dat, shown below, contains simulated data from a cohort study of 5000 individuals genotyped at five SNPs. The observation time and event indicator are specified in the columns titled "Time" and "Status", respectively. The "Smoking" column contains environmental covariate data, and the columns SNP1-SNP5 represent the five SNP sites. Missing SNP values are indicated by '9'.

Time	Status	Smoking	SNP1	SNP2	SNP3	SNP4	SNP5
1000	0	0	1	0	1	2	1
764	1	1	0	2	2	0	2
1000	0	0	0	2	2	0	1
718	1	1	1	1	1	2	2
1000	0	0	9	1	2	1	2
1000	0	0	0	1	2	1	2
1000	0	1	0	2	2	0	9
160	0	1	2	0	0	9	2
1000	0	0	1	1	1	1	1
313	0	1	9	1	1	1	2
125	0	0	2	0	0	2	2
856	0	0	1	1	1	9	1

cohort.dat: Example cohort data file for HAPSTAT input.

Select the tab labeled *Frequencies* in the left panel. In the right panel, select *Hardy-Weinberg disequilibrium*, check both the *Cohort* and *Subcohort* samples and click on *Calculate*. Your results will display on the left; see Figure 4.1.

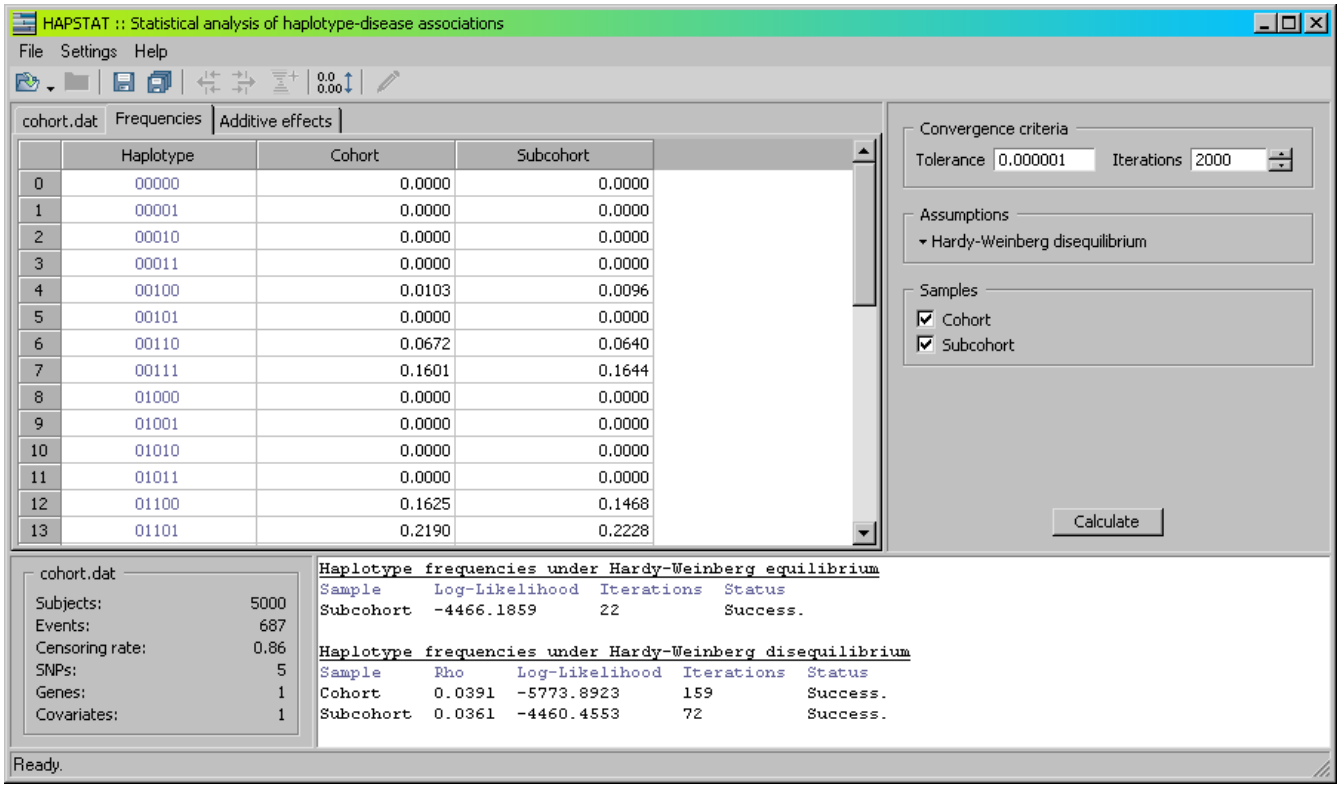

Figure 4.1: Estimating haplotype frequencies under Hardy-Weinberg disequilibrium.

Select the tab labeled *Additive effects*. To estimate dominant effects under Hardy-Weinberg disequilibrium, change the *Assumptions* settings by highlighting the *Dominant* model and *Hardy-Weinberg disequilibrium* from the dropdowns. Click the icon  to activate the *Select effects* dialog. Change *Threshold* to 0.01 and the sample to *Cohort*. Click on *Calculate* to obtain the display in Figure 4.2.

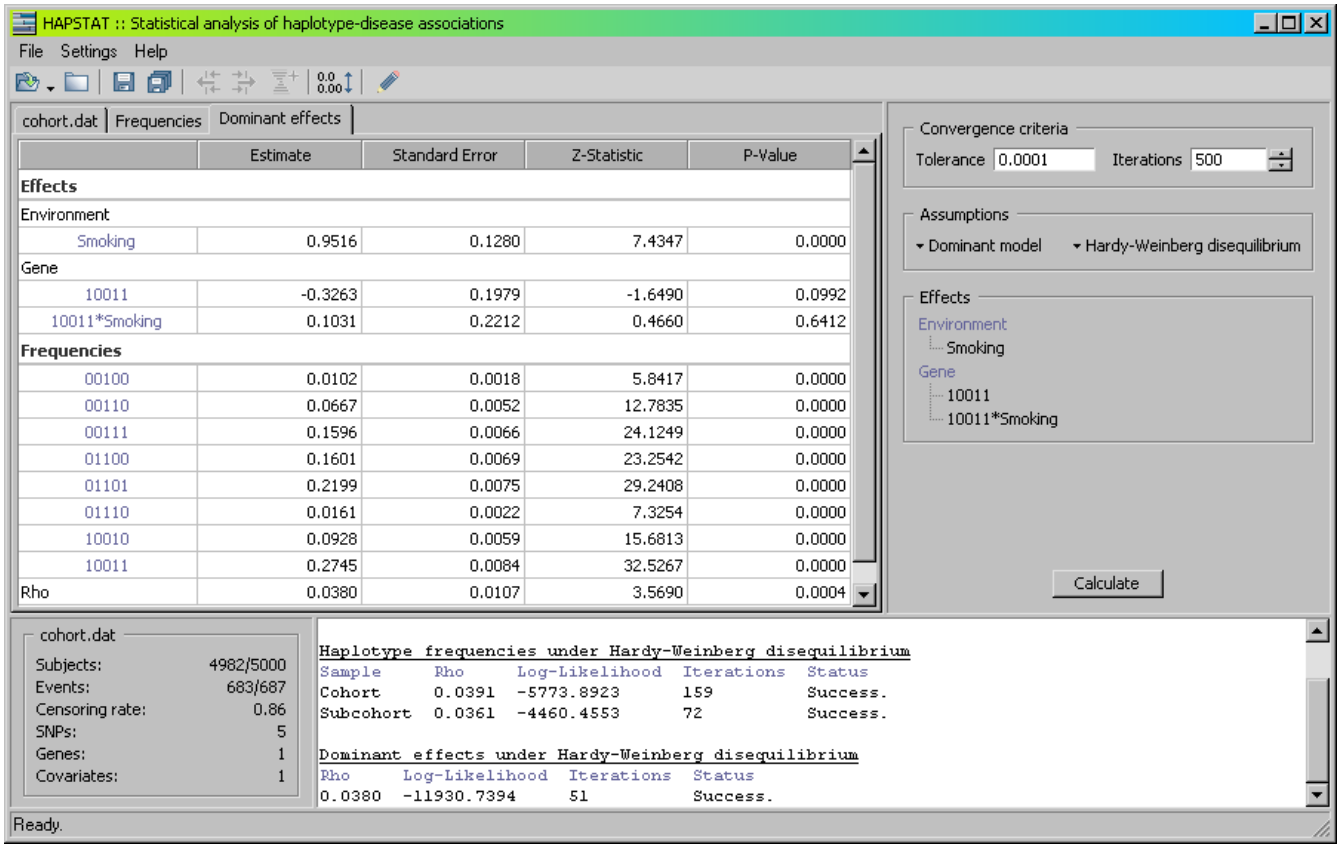

Figure 4.2: Estimating dominant haplotype effects under Hardy-Weinberg disequilibrium.

The results for the cohort data using the options shown in Figures 4.1 and 4.2 are given in cohort.out.

The file cross-sectional.dat, shown below, contains simulated data from a cross-sectional study of 5000 individuals genotyped at six SNPs. Approximately 5% of SNP values are missing. The column titled "Trait" contains disease-related trait data. The columns "Age", "Gender" and "Exposure" contain environmental data and the columns SNP1-SNP6 represent the six SNP sites. Missing SNP values are denoted by '*'.

Trait	Age	Gender	Exposure	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
1.56	22	1	-1.1755	2	2	0	0	0	2
3.61	45	1	0.4119	1	1	1	1	0	2
3.87	54	1	-0.0814	2	2	1	0	0	1
3.48	47	1	-0.1864	1	2	1	0	1	2
6.40	56	1	0.5747	2	0	*	0	2	0
3.96	49	1	0.1212	0	1	2	1	1	2
2.72	28	1	0.6458	1	0	2	2	0	2
4.21	60	0	1.3567	1	1	1	0	2	1
2.32	39	0	-0.9863	0	2	2	0	2	2
2.70	50	0	-1.6120	0	0	2	2	0	2
1.35	17	0	-0.9601	1	0	2	2	0	2
3.14	49	1	-0.5379	1	0	2	2	0	2

cross-sectional.dat: Example cross-sectional data file for HAPSTAT input.

Click the  icon to create two genes, with SNP1-SNP3 as Gene 1 and SNP4-SNP6 as Gene 2. Select the trait data and the three environmental variables and click *Continue*.

Select the tab labeled *Additive effects* in the left panel. Click the toolbar icon  to activate the *Select effects* dialog and add the interactions 001×101, 001×101×Age, 001×101×Gender and 001×101×Exposure. Click on *Calculate* to obtain the display shown in Figure 4.3.

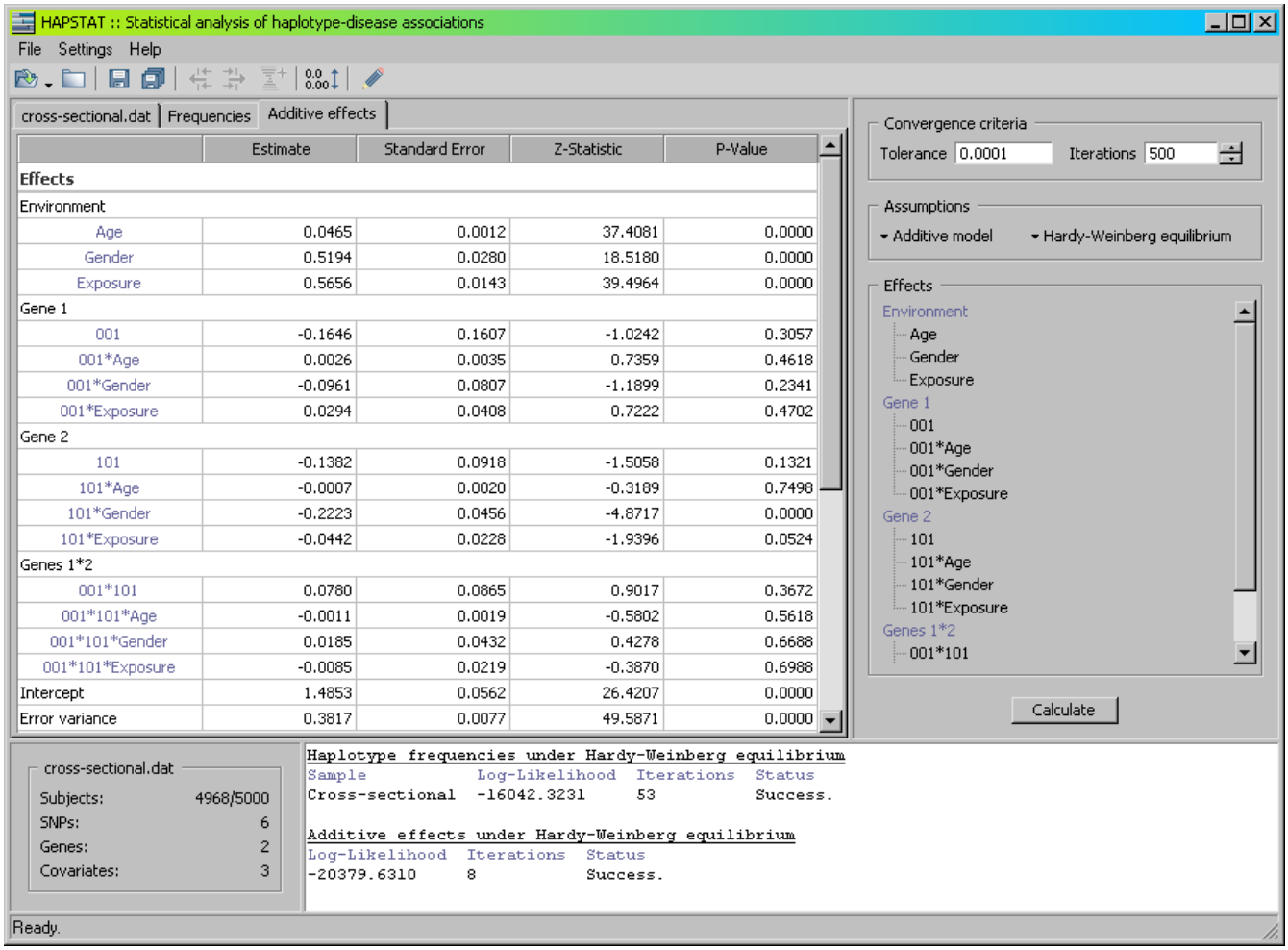

Figure 4.3: Estimating additive haplotype effects under Hardy-Weinberg equilibrium.

Results are provided in the file cross-sectional.out.

The file longitudinal.dat contains simulated data from a longitudinal study of 1000 individuals with a quantitative trait measured at five regular time points and with five SNPs. The column titled "Subject" provides the identifier of the individual. The trait is specified in the column titled "Weight". The "Time" covariate specifies the time point at which the measurement was taken. Columns SNP1-SNP5 represent the five SNP sites.

Subject	Weight	Time	SNP1	SNP2	SNP3	SNP4	SNP5
0001	175	1	1	1	2	1	1
0001	192	2					
0001	165	3					
0001	192	4					
0001	180	5					
0002	170	1	1	1	1	1	1
0002	170	2					
0002	168	3					
0002	195	4					
0002	191	5					
0003	175	1	1	1	1	2	2
0003	186	2					

longitudinal.dat: Example longitudinal data file for HAPSTAT input.

Select the tab labeled *Additive effects* and click on *Calculate* to obtain the result shown in Figure 4.4. Click the icon  to activate the *Select effects* dialog and select the tab labeled *Random effects*. Add the "Time" covariate to the random effects selection as shown in Figure 4.5. Click on *Calculate* to obtain the display in Figure 4.6.

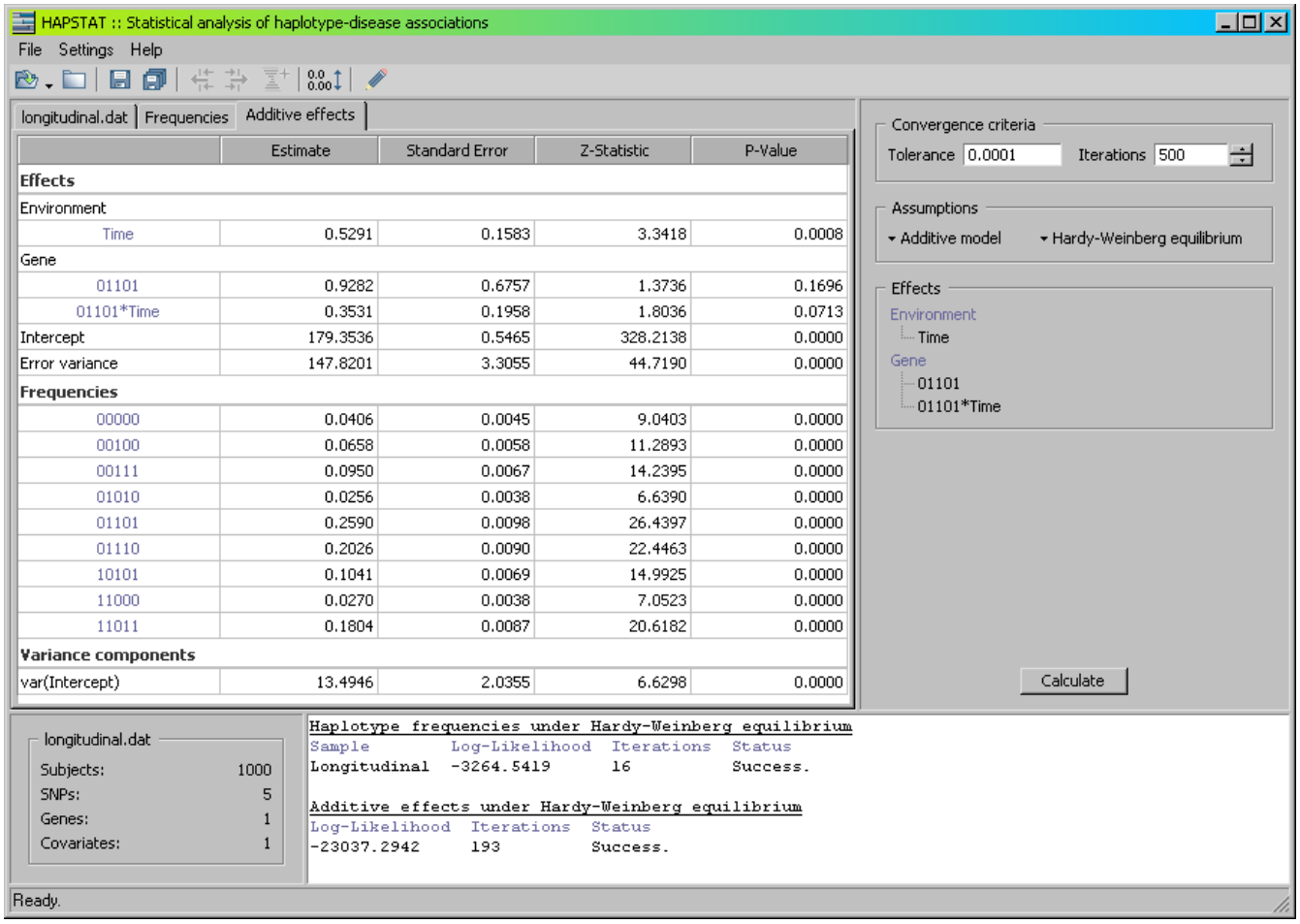

Figure 4.4: Estimating additive haplotype effects under Hardy-Weinberg equilibrium.

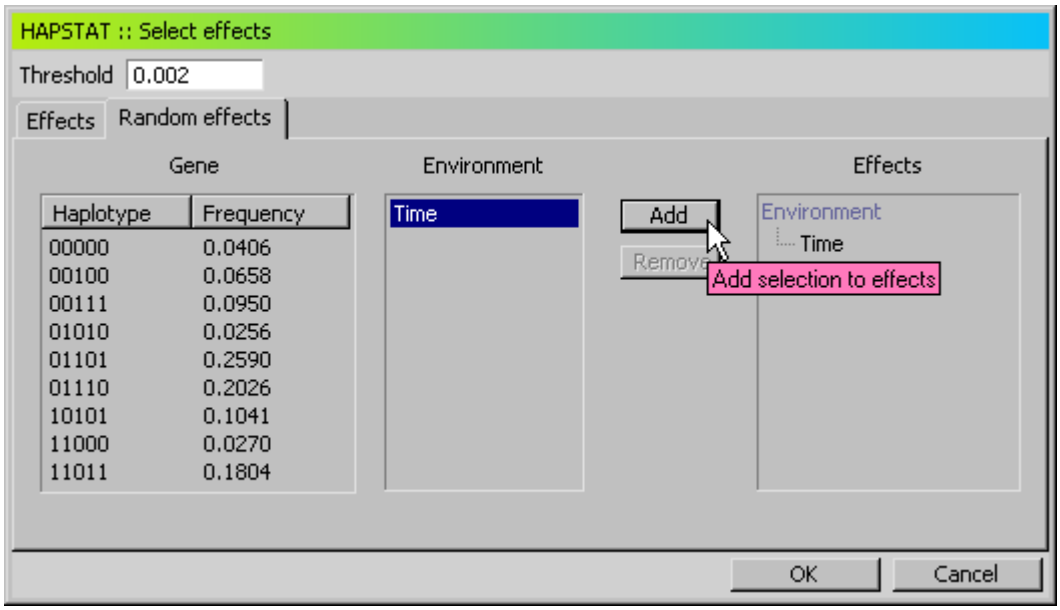

Figure 4.5: Adding the "Time" covariate to the random effects selection.

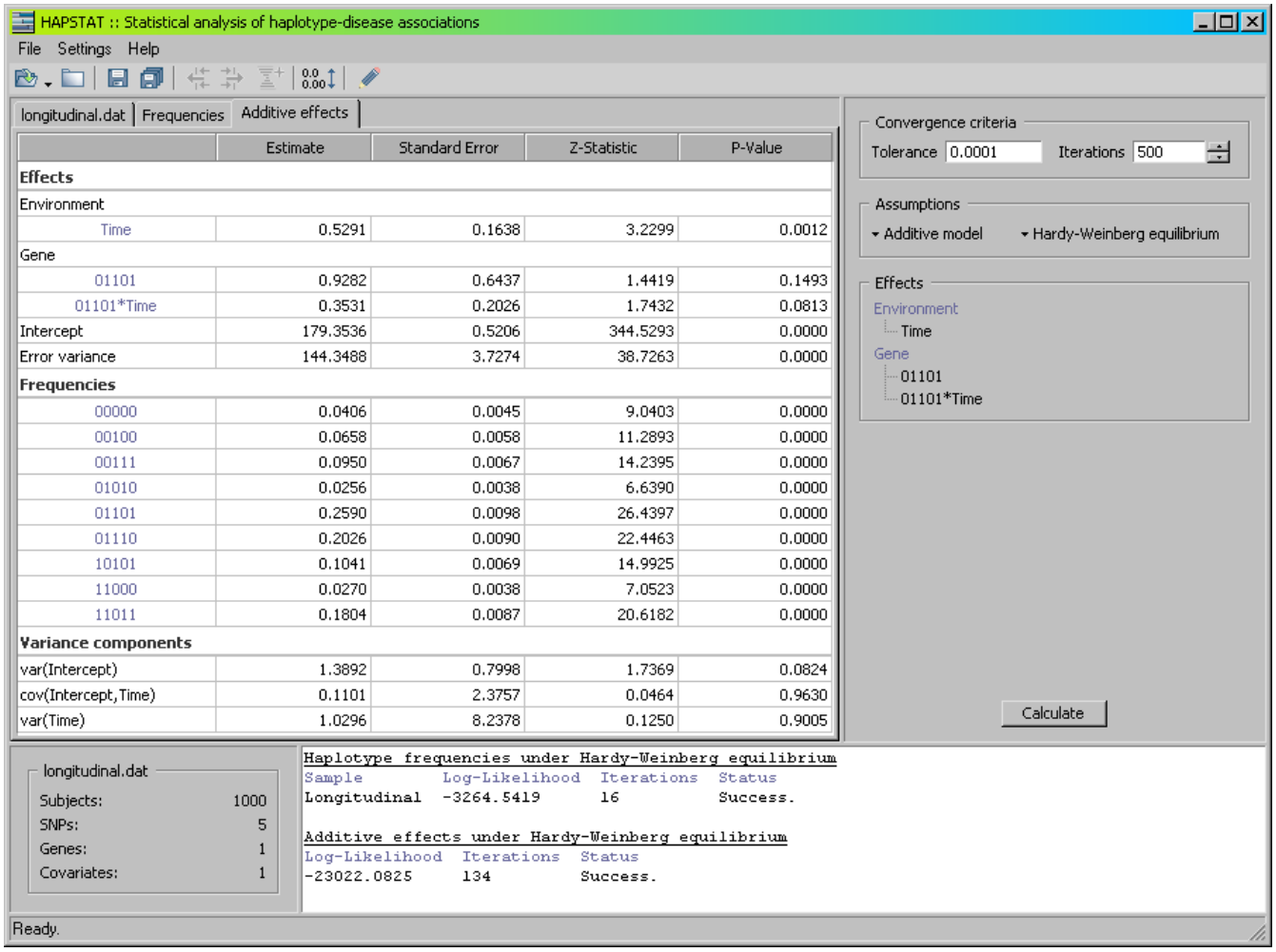

Figure 4.6: Estimating additive haplotype effects under Hardy-Weinberg equilibrium.

Results are provided in the file longitudinal.out.