Input data

HAPSTAT supports data from cross-sectional, case-control and cohort (including case-cohort and nested case-control) association studies. The cross-sectional design collects data on a random sample of individuals. 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 The individuals who are withdrawn prematurely or who are disease-free at the end of the cohort study have censored observations, in that their ages at onset are only known to be beyond their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 column describing either the trait value of the individual in the cross-sectional study or the disease status of the individual in the case-control study. The cohort study requires two columns of data per individual providing the observation time and event indicator. The file must also 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 column order.

Data specification

In cross-sectional studies, 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 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 other value 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 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.

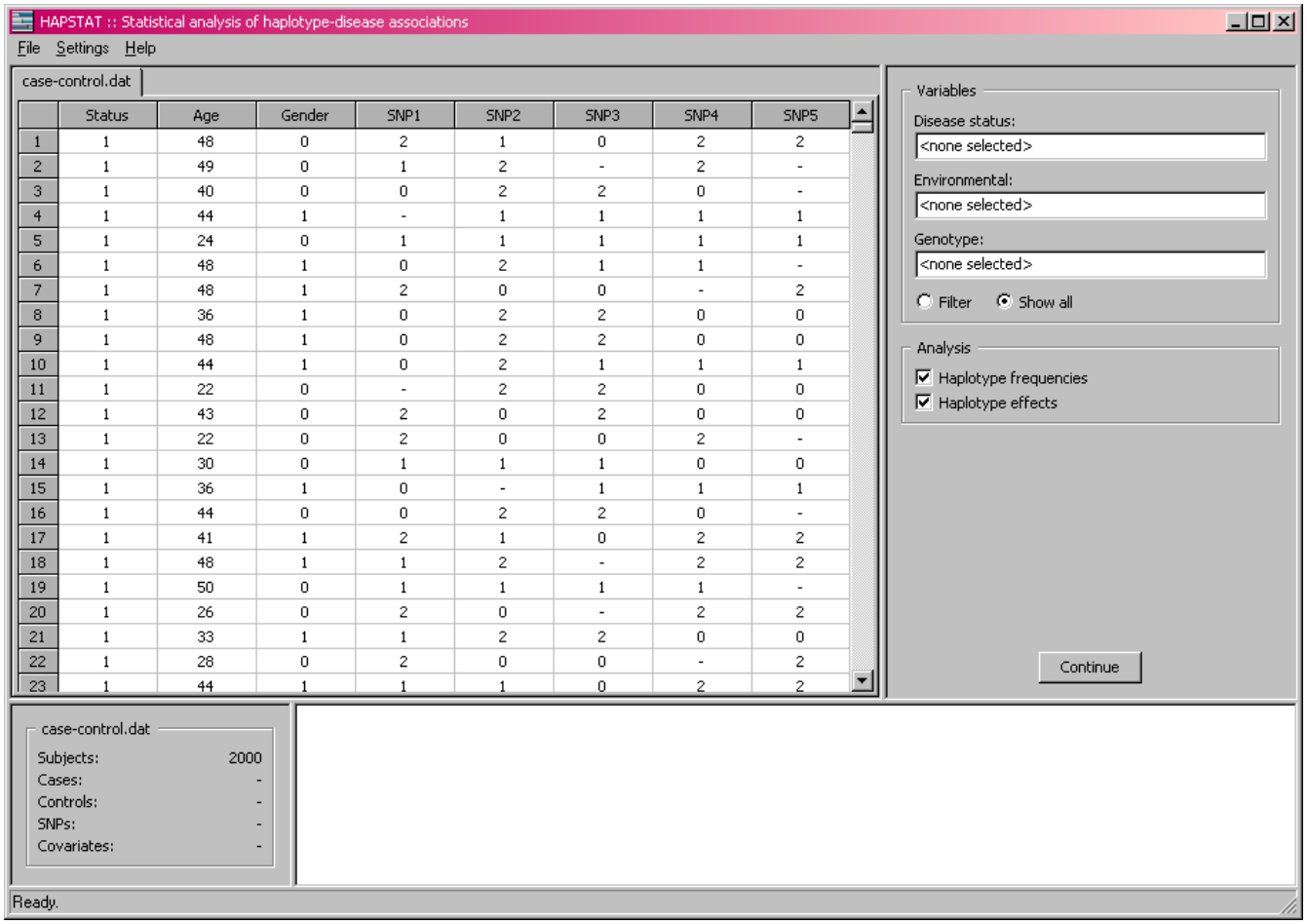
Status	Age	Gender	SNP1	SNP2	SNP3	SNP4	SNP5
1	48	0	2	1	0	2	2
1	49	0	1	2	.	2	.
1	40	0	0	2	2	0	.
1	44	1	.	1	1	1	1
1	24	0	1	1	1	1	1
1	48	1	0	2	1	1	.
1	48	1	2	0	0	.	2
1	36	1	0	2	2	0	0
1	48	1	0	2	2	0	0

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.



	Status	Age	Gender	SNP1	SNP2	SNP3	SNP4	SNP5
1	1	48	0	2	1	0	2	2
2	1	49	0	1	2	-	2	-
3	1	40	0	0	2	2	0	-
4	1	44	1	-	1	1	1	1
5	1	24	0	1	1	1	1	1
6	1	48	1	0	2	1	1	-
7	1	48	1	2	0	0	-	2
8	1	36	1	0	2	2	0	0
9	1	48	1	0	2	2	0	0
10	1	44	1	0	2	1	1	1
11	1	22	0	-	2	2	0	0
12	1	43	0	2	0	2	0	0
13	1	22	0	2	0	0	2	-
14	1	30	0	1	1	1	0	0
15	1	36	1	0	-	1	1	1
16	1	44	0	0	2	2	0	-
17	1	41	1	2	1	0	2	2
18	1	48	1	1	2	-	2	2
19	1	50	0	1	1	1	1	-
20	1	26	0	2	0	-	2	2
21	1	33	1	1	2	2	0	0
22	1	28	0	2	0	0	-	2
23	1	44	1	1	1	0	2	2

Figure 1.1: Importing a case-control data file.

You may only have one file open in HAPSTAT at any time, so you must close the current file by selecting the menu option *File»Close* before opening another.

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 in your file which 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. You can

show or hide unselected columns by toggling the *Filter* and *Show all* radio buttons. The variables chosen from case-control.dat are shown in Figure 1.2.

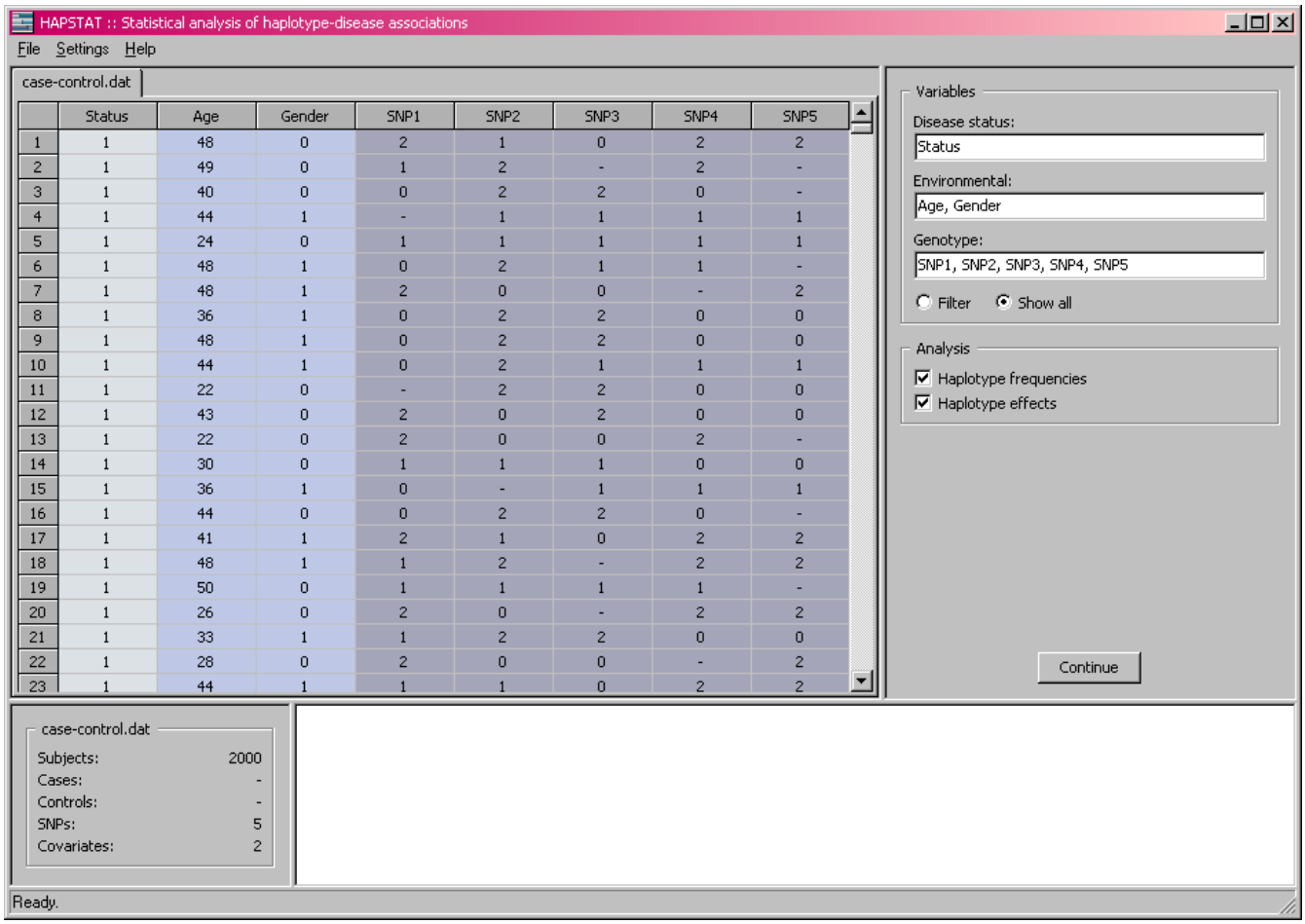


Figure 1.2: Selecting case-control variables.

Choose to estimate haplotype frequencies and/or haplotype effects from the *Analysis* box on the right panel and then click *Continue*.

Edit/clear selection

To change your variable selections after the *Continue* button is clicked, return to the file tab and click the *Edit* button. Select *Settings*»*Clear* to clear all variables.

Sorting

To sort data, right click on the title of the column you wish to sort by and select *Sort ascending* or *Sort descending*. All columns will be sorted accordingly.

Frequency estimation

Navigation

Select the tab labeled *Frequencies* in the left panel; see Figure 1.3. The options available to the user are located in the right panel. After you click on *Calculate*, your results will display on the left.

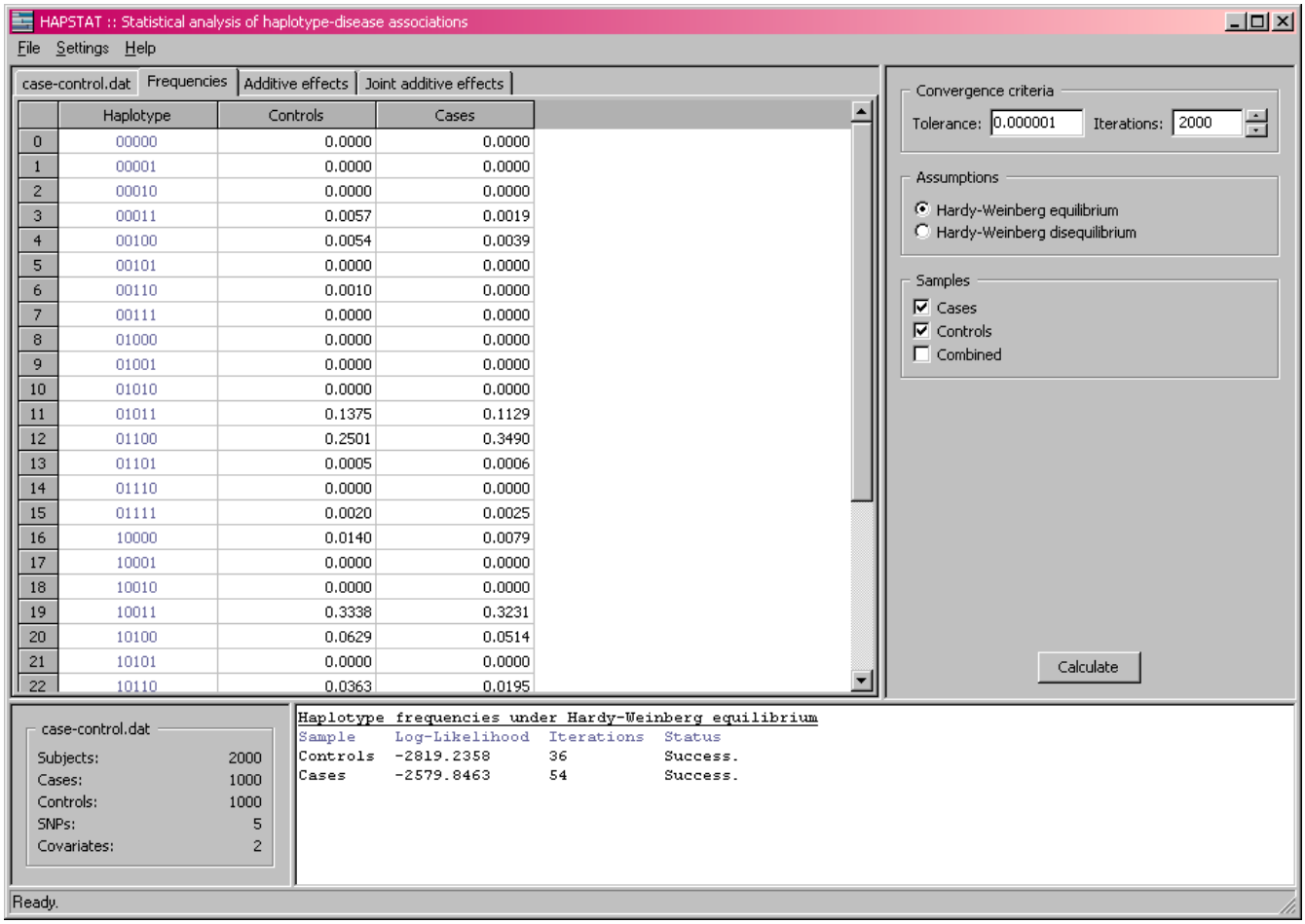


Figure 1.3: Estimating haplotype frequencies under Hardy-Weinberg equilibrium.

Convergence criteria

These settings apply to the EM algorithm used by HAPSTAT to estimate haplotype frequencies. The algorithm will terminate 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:

$$\max_k \mid \delta_{k,i} \mid \; < \; \varepsilon \,,$$

where ε denotes the specified value of *Tolerance*,

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

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

Toggling the radio buttons in this box will estimate haplotype frequencies assuming that the population is in Hardy-Weinberg equilibrium (default) or disequilibrium. For Hardy-Weinberg disequilibrium, HAPSTAT will return an estimate for the inbreeding coefficient (ρ).

Samples

For cross-sectional 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. 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

To sort the frequency estimates shown in the left panel, right click on the title of the column you wish to sort by and select *Sort ascending* or *Sort descending*. All columns will be 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 the frequency sample to filter by. 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*. In the dialog box, enter the desired number of significant digits. The default setting is 4 significant digits.

Saving

To save frequency estimates, select the menu option *File»Save*, 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* will save results of all open tabs. HAPSTAT result files are in text format and can be opened 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 ratio. 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.

Navigation

HAPSTAT will fit both separate and joint models. Under the separate model, each haplotype is compared in turn to all of the others. Under the joint model, all haplotypes are simultaneously compared to one reference haplotype. Estimate separate effects by selecting the tab in the left panel labeled *Additive effects*. The options available to the user display in the right panel, shown in Figure 1.4.

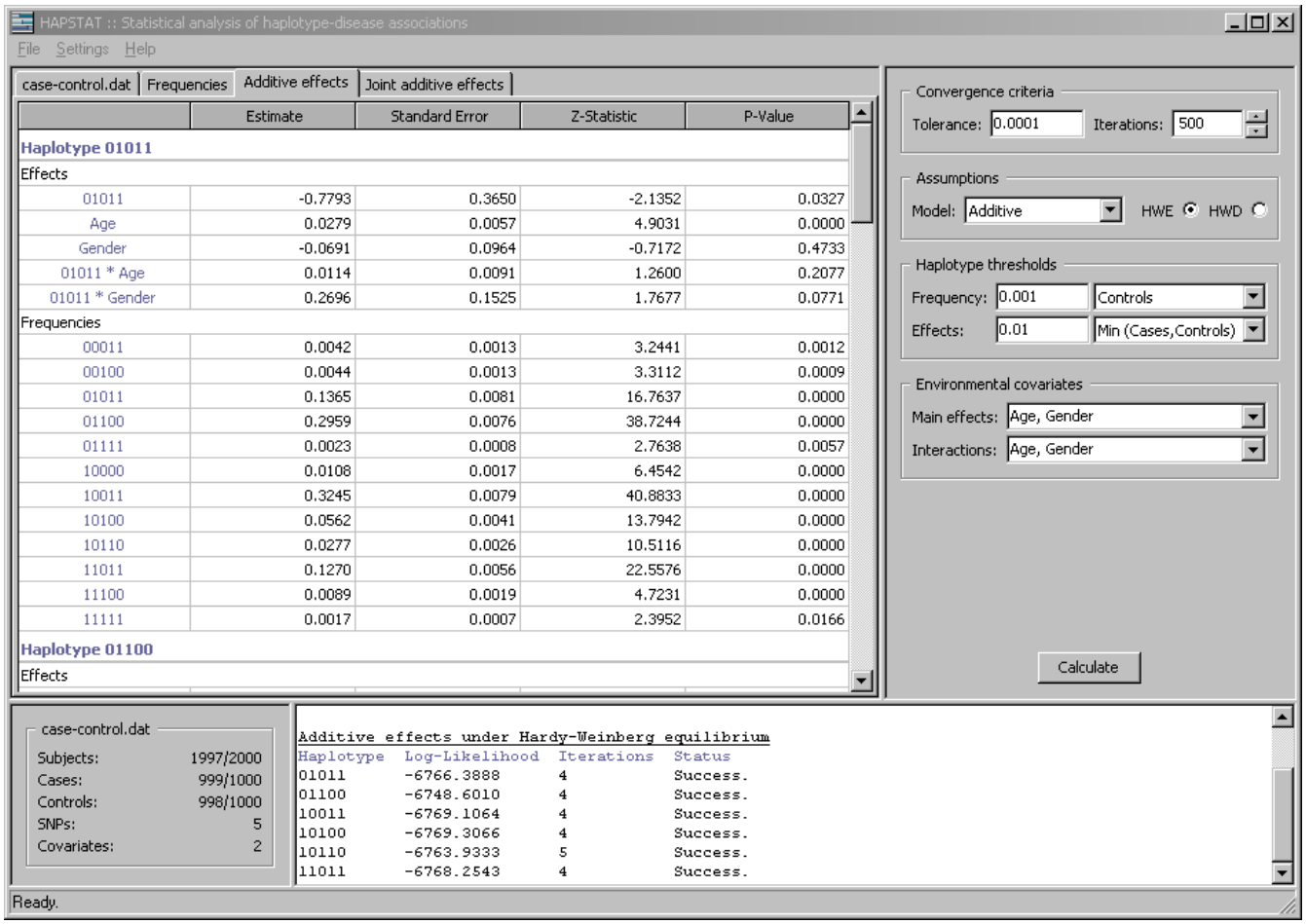


Figure 1.4: Estimating additive haplotype effects under separate models.

Select the tab in the left panel labeled *Joint additive effects* to estimate effects under the joint model. The options available to the user are shown in Figure 1.5.

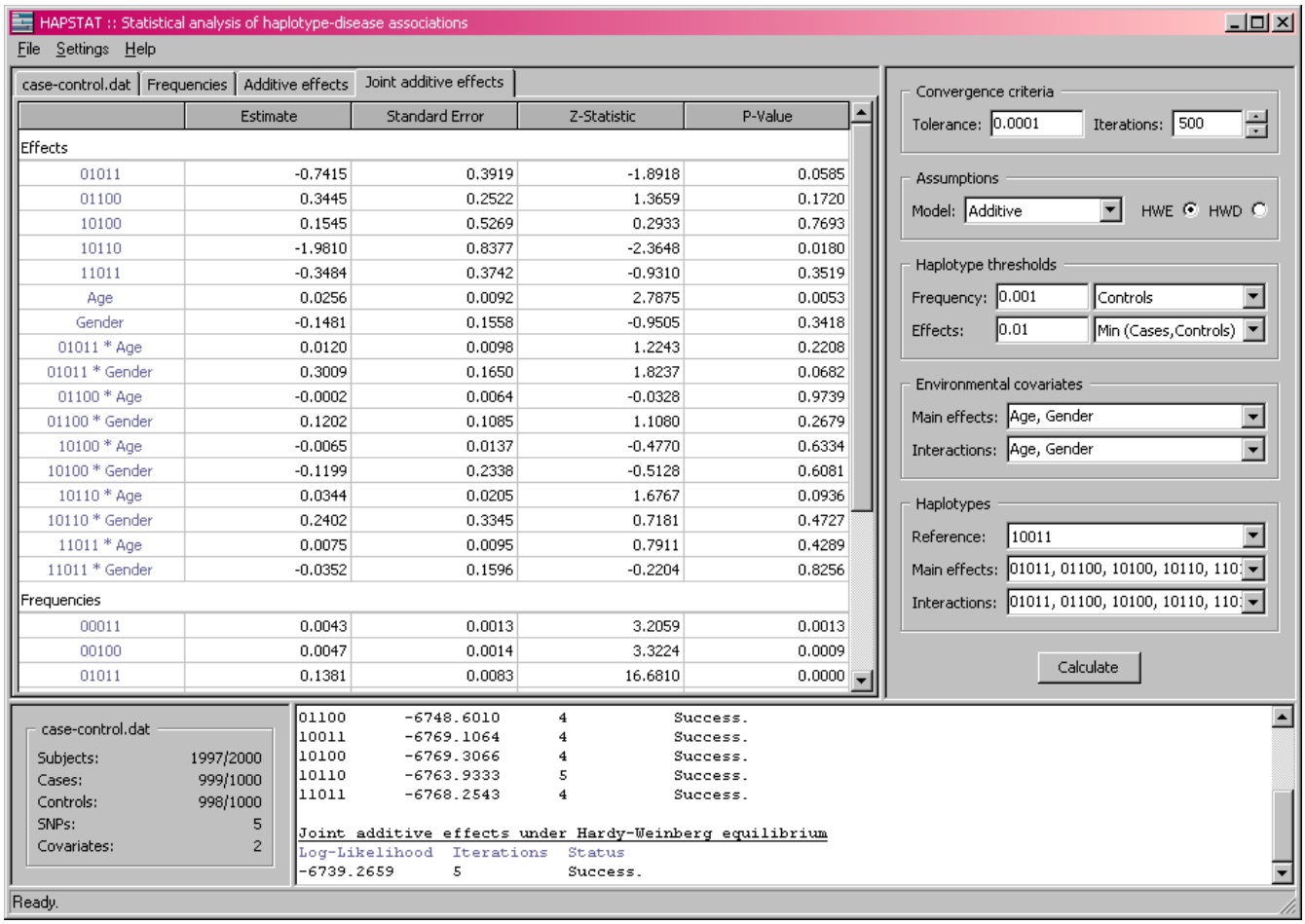


Figure 1.5: Estimating additive haplotype effects under the joint model.

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.

Convergence criteria

The same convergence criteria are used as the EM algorithm to estimate haplotype frequencies. See the previous section for details. 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 *Model* dropdown menu. Toggling the HWE and HWD radio buttons in this box will estimate haplotype effects under Hardy-Weinberg equilibrium (default) or disequilibrium. For Hardy-Weinberg disequilibrium, HAPSTAT will return an estimate for the inbreeding coefficient (ρ).

Haplotype thresholds

The haplotypes whose frequencies are no greater than the value specified by the *Frequency* threshold are removed from calculation. For case-control and cohort studies, frequencies are determined by the sample

chosen from the adjacent dropdown menu. 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. Under the separate model, haplotype effects are estimated for all haplotypes with frequencies exceeding the value specified by the *Effects* threshold. The default value is

$$\max (20/n , 0.01).$$

For case-control studies, this threshold is satisfied by both the case and control samples by default. For cohort studies, the default is the subcohort.

Under the joint model, the effects of all haplotypes with frequencies exceeding the value specified by the *Effects* threshold are compared to a reference group. The reference group consists of the reference haplotype and those haplotypes whose frequencies are below the *Effects* threshold. The threshold values should be increased if non-convergence is encountered. In Figure 1.6, for example, HAPSTAT failed to estimate recessive joint effects under Hardy-Weinberg disequilibrium using the default settings. By changing the *Effects* threshold to 0.02, you should obtain the results given in case-control.out.

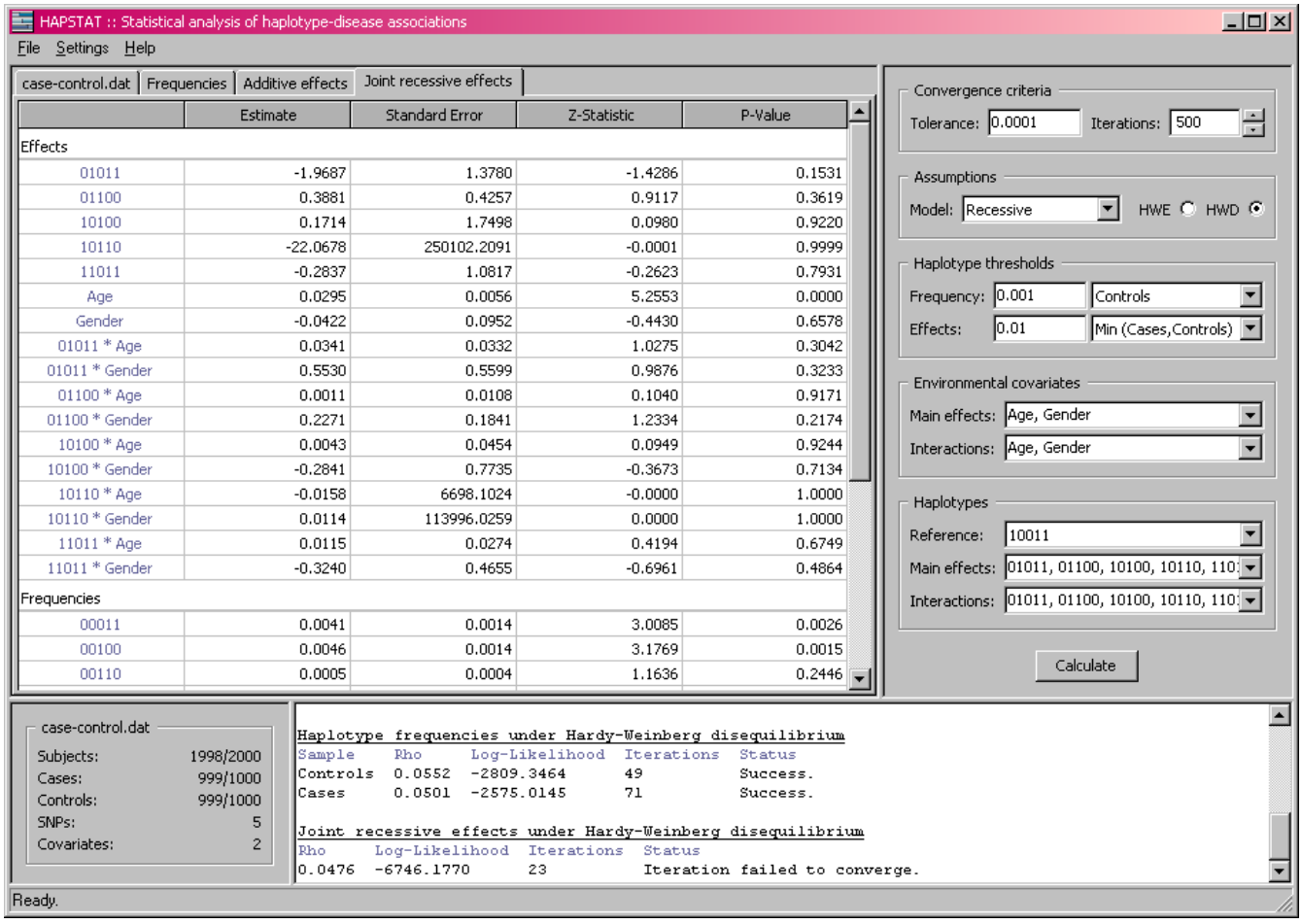


Figure 1.6: Estimating recessive haplotype effects under the joint model.

Environmental covariates

Calculate the haplotype effects and haplotype-environment interactions for (possibly empty) subsets of

environmental covariates using the *Main effects* and *Interactions* dropdown menus. By default, all covariates chosen from the input data are selected.

Haplotypes

Calculate the haplotype effects and haplotype-environment interactions using (possibly empty) subsets of the haplotypes determined from the *Effects* threshold using the *Reference*, *Main effects* and *Interactions* dropdown menus. By default, the reference haplotype is set to the most frequent, and all remaining haplotypes are selected for computation.

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*. In the dialog box, enter the desired number of significant digits. The default setting is 4 significant digits.

Saving

Select the menu option *File*»*Save* to save the effects estimates. This option will save the values shown in the current tab only. To save the results of all open tabs, including haplotype frequencies, use the *File*»*Save All* menu option. The results for the case-control data using the options shown in Figures 1.3 —1.5 are given in case-control.out.

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

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*. See Figure 2.1.

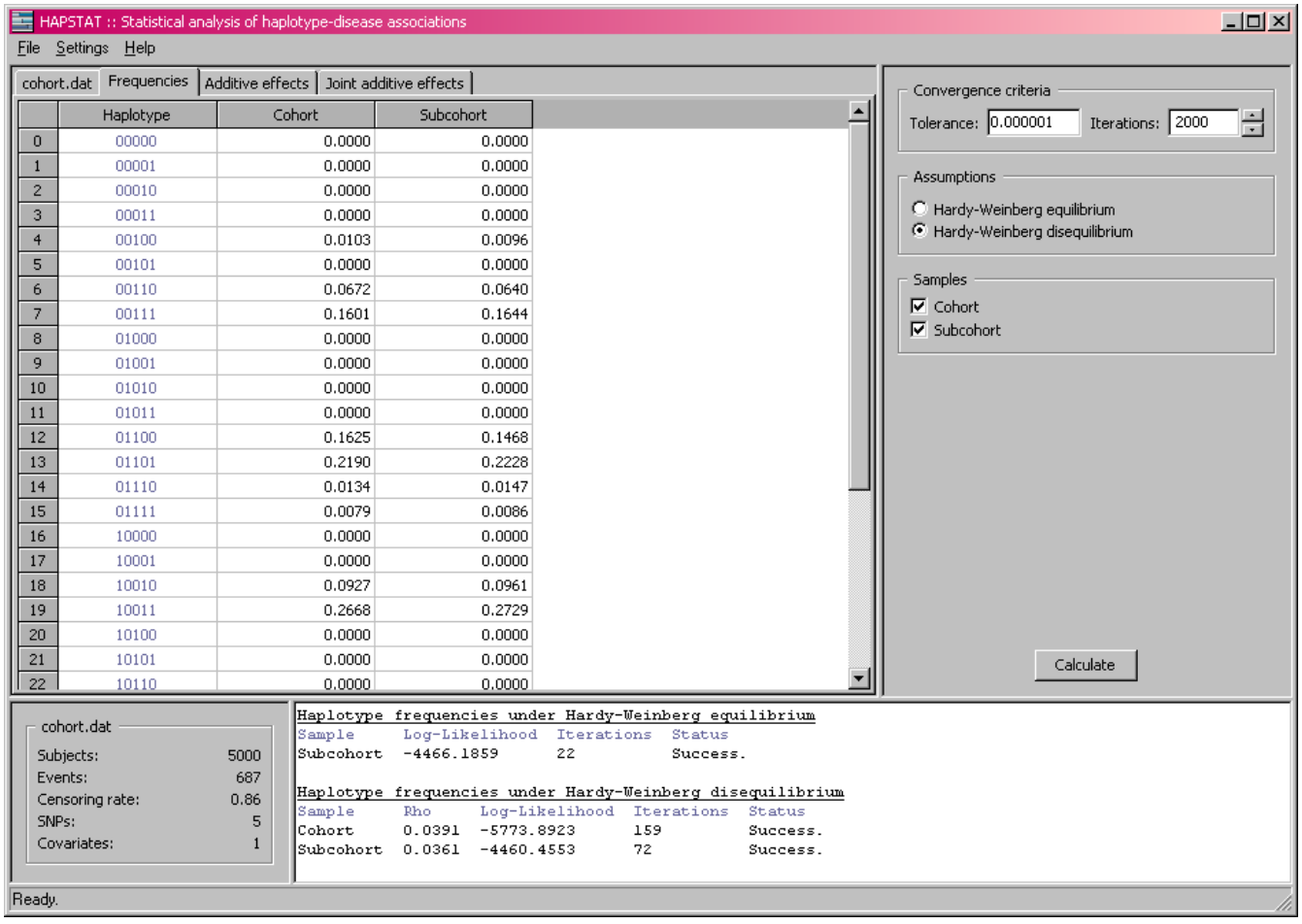


Figure 2.1: Estimating haplotype frequencies under Hardy-Weinberg disequilibrium.

Select the tab labeled *Additive effects*. To estimate dominant effects under Hardy-Weinberg disequilibrium, change the *Assumptions* settings by highlighting the *Dominant* model and selecting the *HWD* radio button. Click on *Calculate* to obtain the results in Figure 2.2.

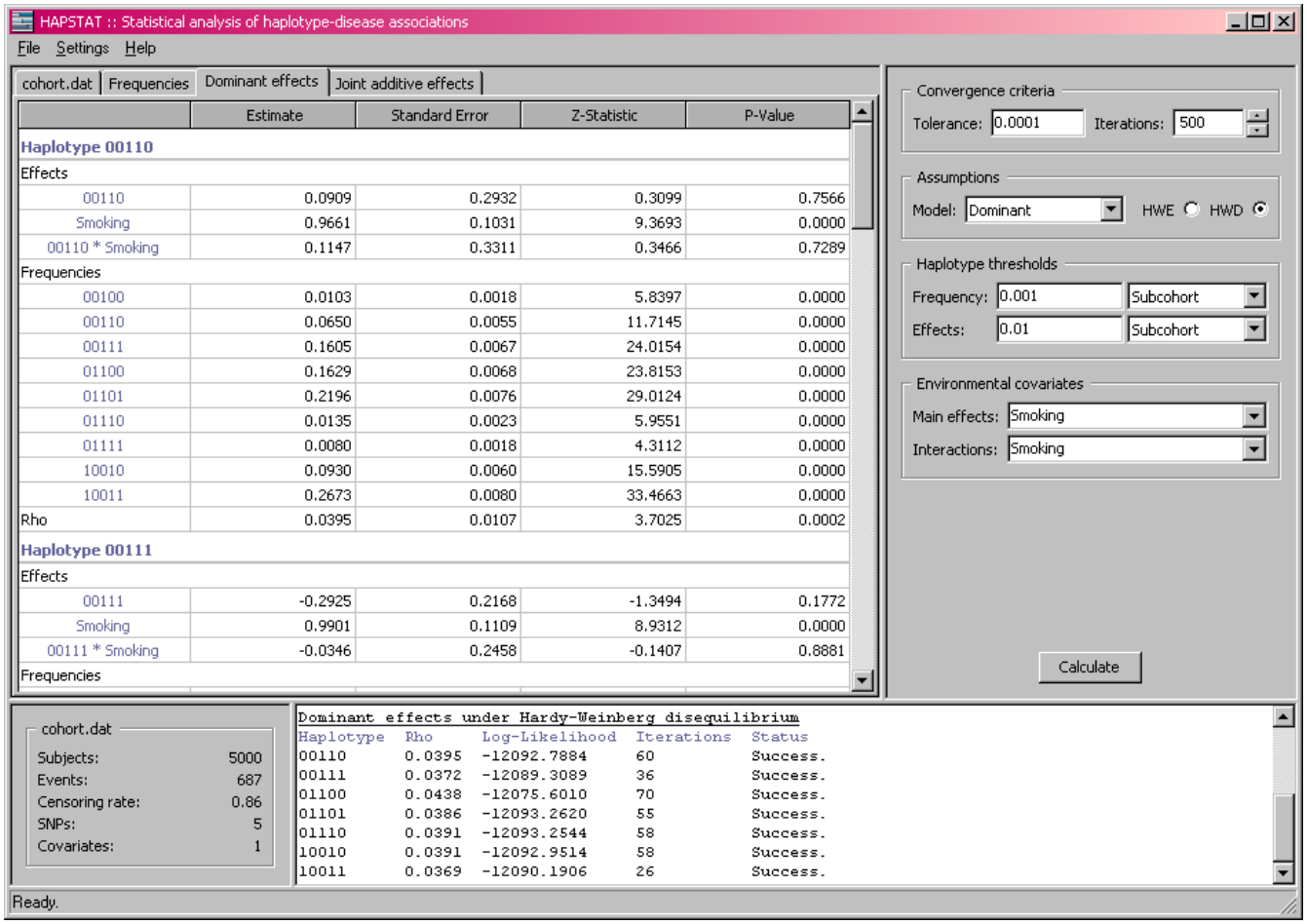


Figure 2.2: Estimating dominant haplotype effects under separate models.

To estimate joint dominant effects under Hardy-Weinberg disequilibrium, select the tab labeled *Joint additive effects* and change the *Assumptions* settings as in the previous example. The result after clicking on *Calculate* is shown in Figure 2.3.

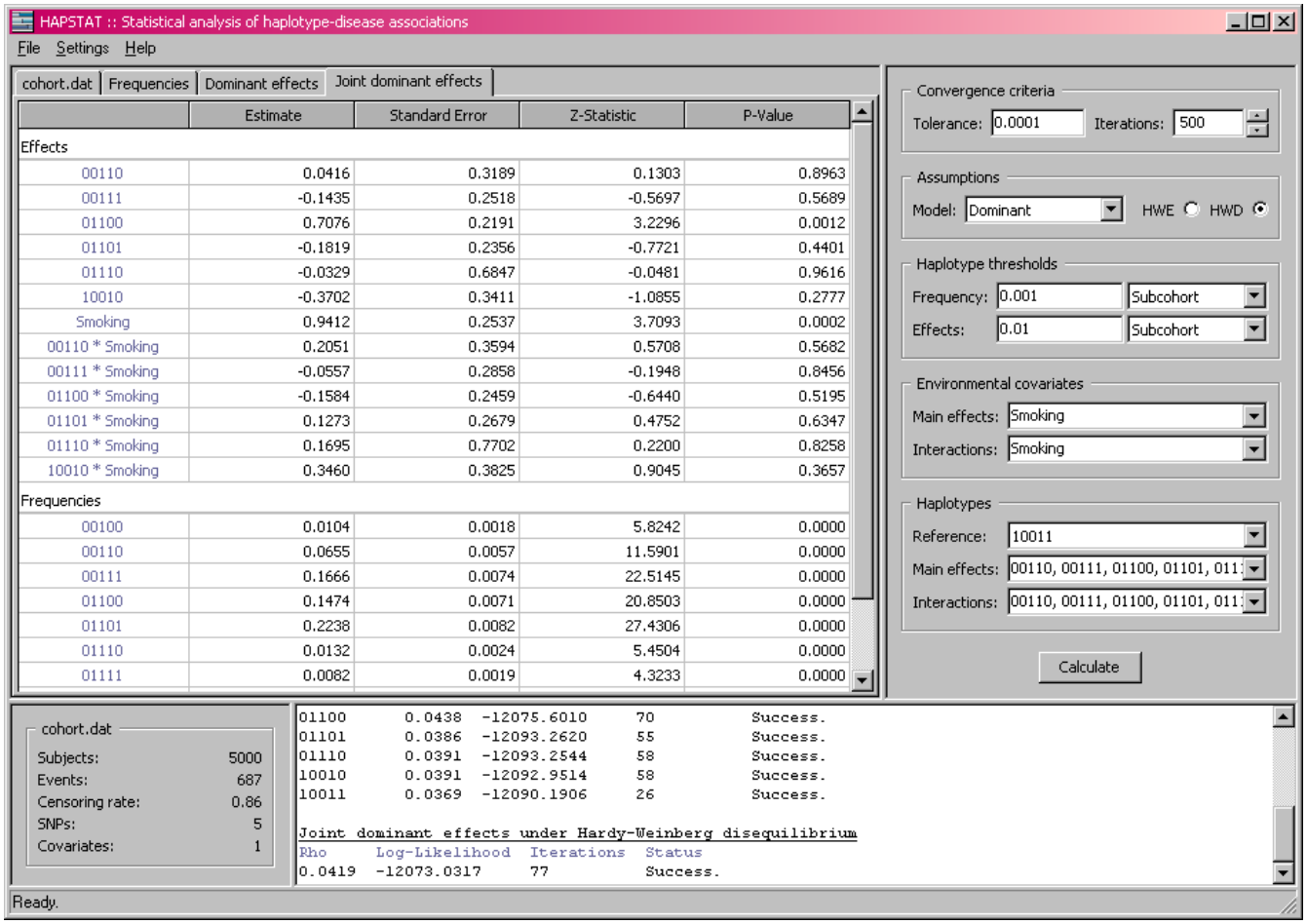


Figure 2.3: Estimating dominant haplotype effects under the joint model.

To estimate joint codominant effects under Hardy-Weinberg disequilibrium, select the *Codominant* model in the *Assumptions* settings. Note in Figure 2.4 that the iteration fails using the default settings.

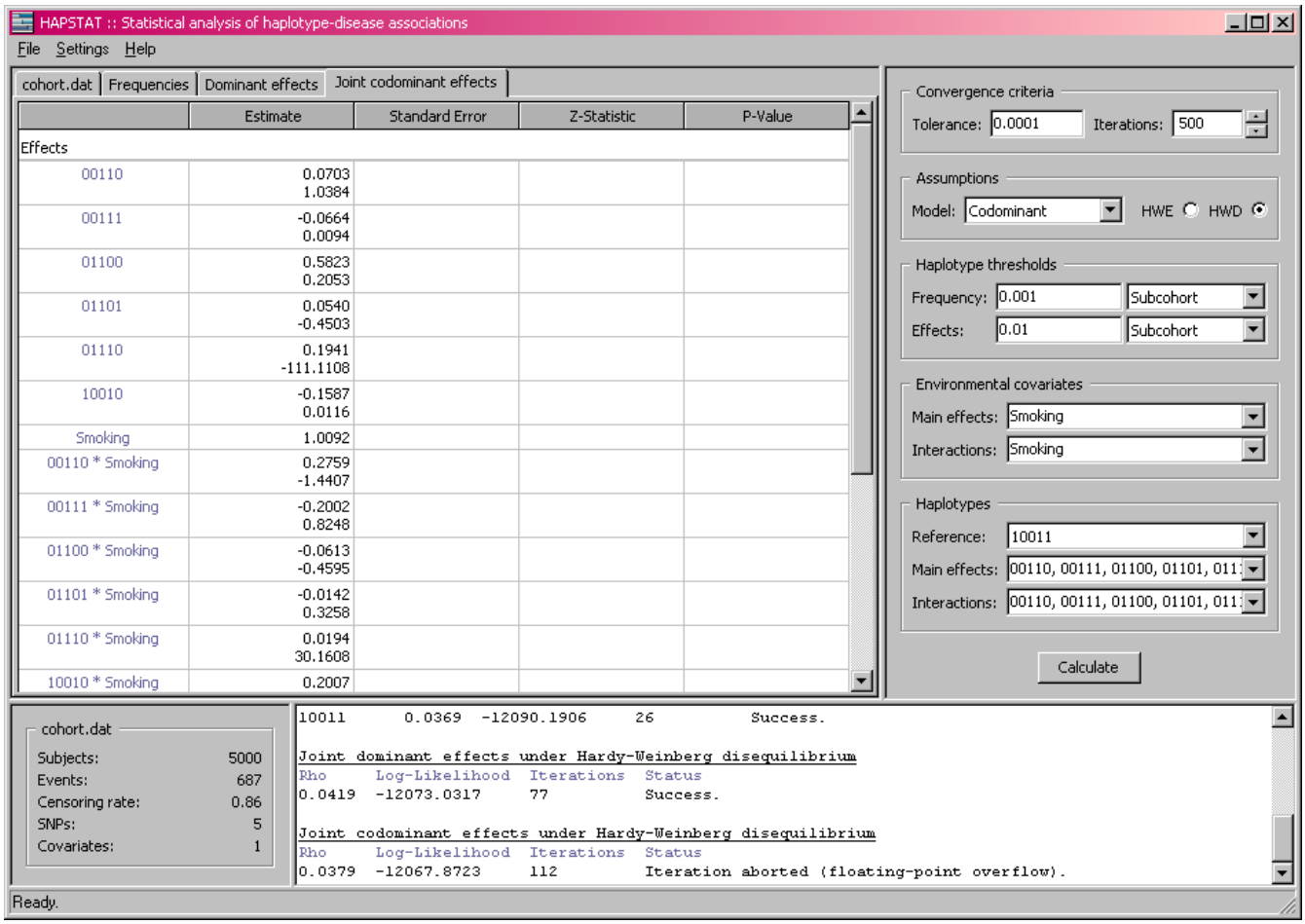


Figure 2.4: Estimating codominant haplotype effects under the joint model using default settings.

Selecting the haplotype 01110 as reference gives the results in Figure 2.5. Figure 2.6 shows the result after instead changing the *Effects* threshold to 0.02.

The results for the cohort data using the options shown in Figures 2.1—2.3, 2.5 and 2.6 are given in cohort.out.

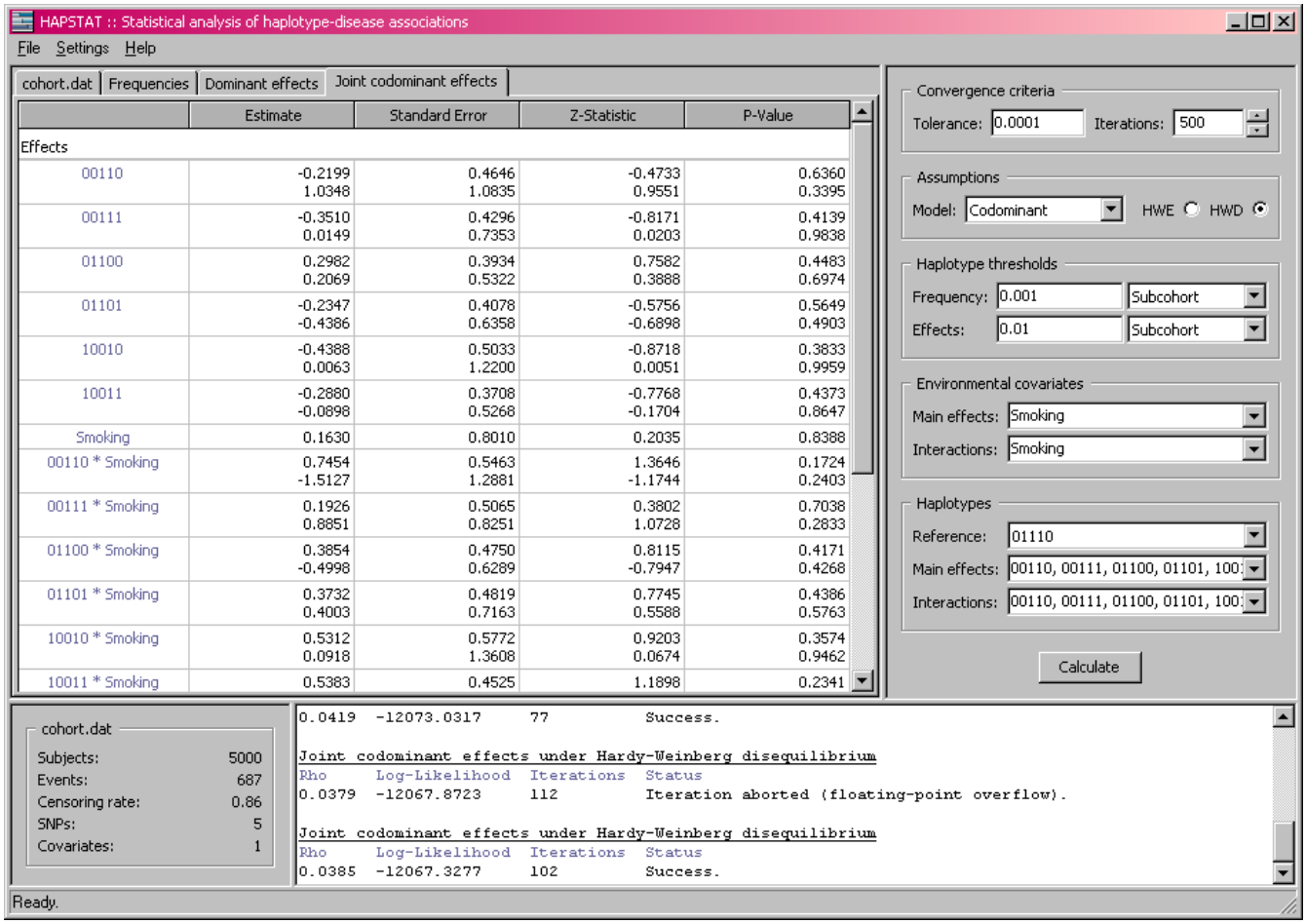


Figure 2.5: Re-estimating codominant haplotype effects under the joint model after changing the reference haplotype.

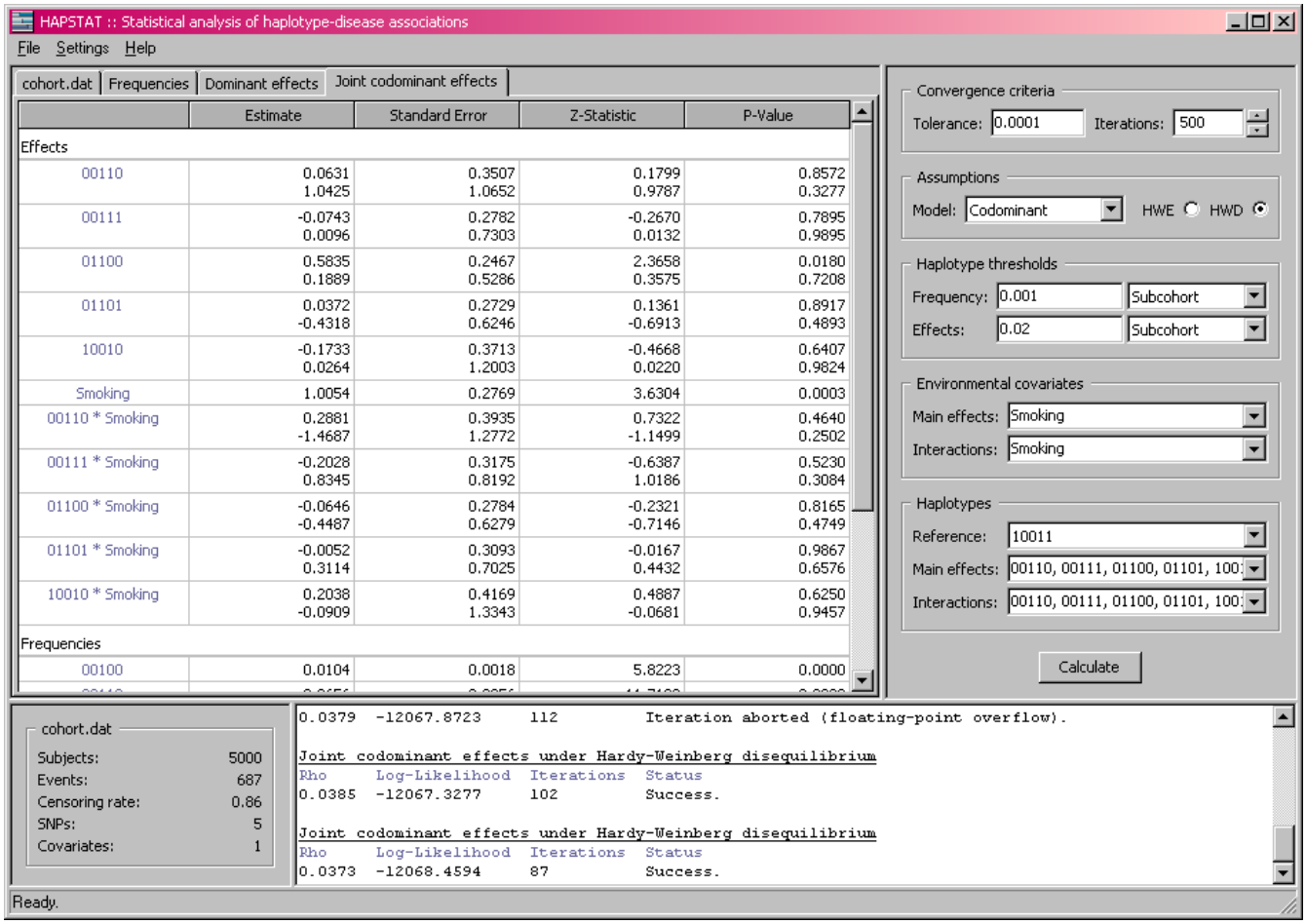


Figure 2.6: Re-estimating codominant haplotype effects under the joint model after changing the effects threshold.

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.

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

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

The column titled "Trait" contains disease-related trait data. The columns "Age", "Gender" and "Exposure" contain environmental covariate data and the columns SNP1-SNP6 represent the six SNP sites. Missing SNP values are denoted by '*'.

Select the tab labeled *Additive effects*. In the right panel, select *HWD*, change the *Effects* threshold to 0.05 and deselect haplotype×age interactions. Click on *Calculate* to obtain the display in Figure 3.1.

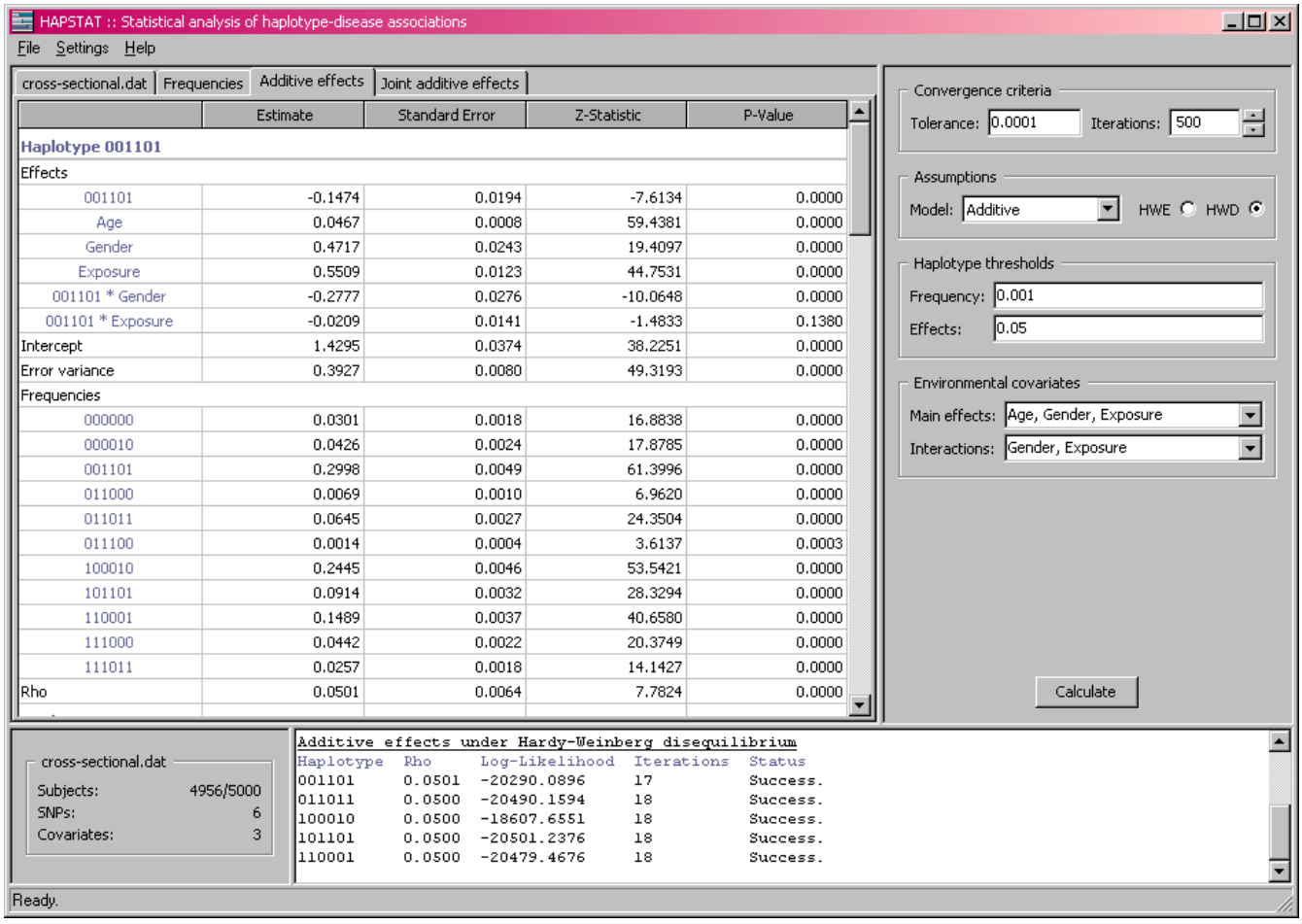


Figure 3.1: Estimating additive haplotype effects under separate models and Hardy-Weinberg disequilibrium.

Select the tab labeled *Joint additive effects* and change the settings as in the previous example; see Figure 3.2. The HAPSTAT display after changing the reference haplotype to 110001 is shown in Figure 3.3. Results are provided in the file cross-sectional.out.

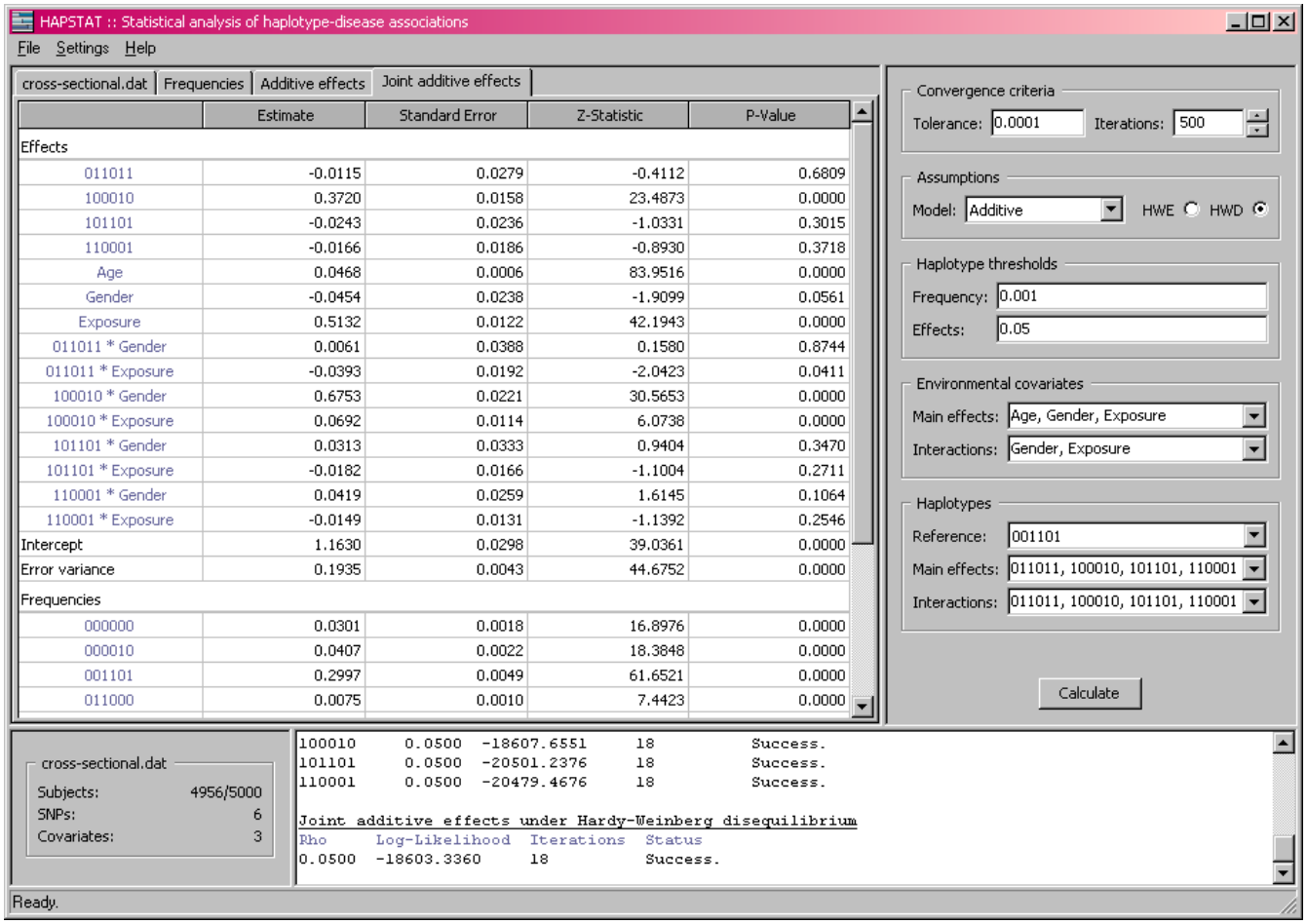


Figure 3.2: Estimating additive haplotype effects under the joint model using the most frequent haplotype as reference.

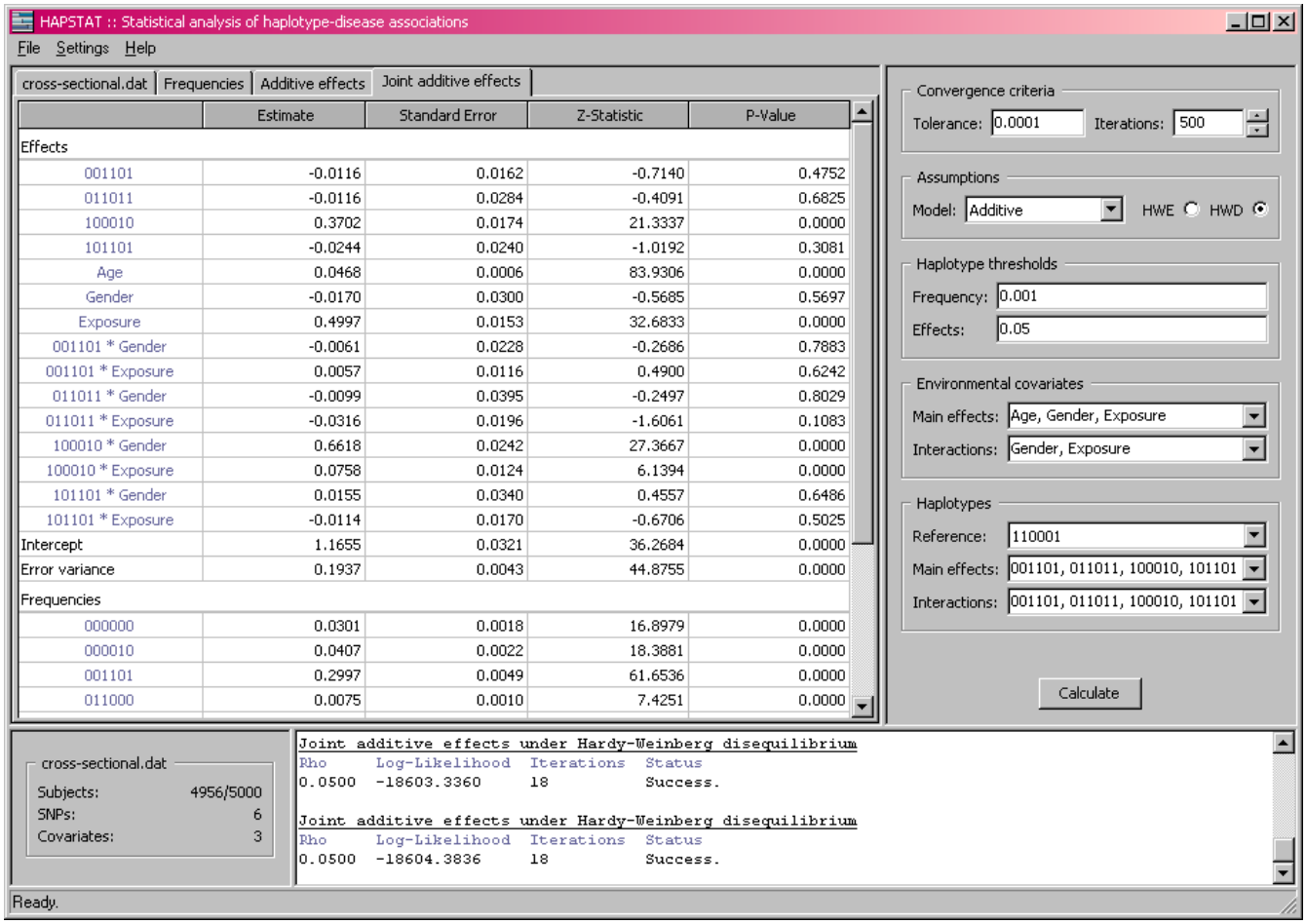


Figure 3.3: Estimating additive haplotype effects under the joint model using haplotype 110001 as reference.