Chaplin Case-control haplotype inference software

Emory University School of Medicine
Department of Human Genetics
615 Michael Street, Suite 301
Atlanta, GA 30322

Version 1.2.2, June 2006
http://www.genetics.emory.edu/labs/epstein/software/chaplin/index.html
Comments about this documentation should be sent to *statgen@genetics.emory.edu*.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Purpose

CHAPLIN is a software program for identifying specific haplotypes or haplotype features that are associated with disease using genotype data from a case-control study. The program utilizes statistical methodology initially developed in Epstein and Satten (2003) and explored further in Satten and Epstein (2004). For a thorough description of the statistical methodology, please review these papers (which are available for download here).

1.2 Background

Case-control study designs are popular for linkage-disequilibrium (LD) mapping of complex diseases. For such a design, one can test for LD of a marker with disease by applying traditional goodness-of-fit association tests that assess differences in marker allele frequencies between the case and control samples. However, a more powerful approach for detecting LD is to examine haplotypes, which are a series of tightly-linked marker alleles found on the same chromosome.

Haplotype analysis is routinely complicated by the fact that one typically only has access to genotype, rather than haplotype, data. Given marker genotype data, a subject's pair of haplotypes is ambiguous given a heterozygous genotype at more than one marker. Nevertheless, one can perform haplotype analysis from genotype data by implementing an expectation-maximization (EM) algorithm to infer haplotype information from sample genotype data. One can then test for LD by constructing an omnibus goodness-of-fit statistic that examines differences in EM-inferred haplotype frequencies between the case and control samples.

These omnibus haplotype-association tests do not provide inference on the effects of specific haplotypes or haplotype features on disease. Since such inference is valuable for identifying specific chromosomal segments that contain a disease variant, we developed a likelihood approach for this purpose that uses a variant of the EM algorithm, called the Expectation-Conditional-Maximization (ECM) algorithm, for analysis. This likelihood approach allows for multiplicative, dominant, and recessive modeling of specific haplotype features on disease risk. The approach relaxes the assumptions of Hardy-Weinberg equilibrium (HWE) of haplotype frequencies both in the control and case samples that are typically required for EM-based analyses. The approach also allows for missing genotype data at any or all of the markers considered in the analysis. .

CHAPLIN is a software program that implements this approach. Given user-defined haplotype effects, the program provides estimates of haplotype effects on disease and also tests whether such effects are significantly different from zero using likelihood-based statistics. Based on the results from a **CHAPLIN** analysis, an analyst can identify genetic variants or regions that are in LD with the disease and, hence, facilitate positional cloning efforts.

1.3 Assumptions

CHAPLIN makes the following assumptions:

1. All genetic markers are biallelic (e.g. SNPs), autosomal, and codominant.
2. Genotyping is performed without error
3. Control haplotype frequencies either are in HWE or have excessive homozygosity relative to HWE. Excessive heterozygosity of control haplotype frequencies may lead to improper inference (for details, see page 14 of the Satten and Epstein (2004) paper).

1.4 System requirements and installation

CHAPLIN is a Win32 application written in FORTRAN 90 using IMSL numerical routines and has been tested on both Windows XP and 2000 machines. **CHAPLIN** has not been tested on earlier versions of the Windows OS or on UNIX platforms but it is unlikely to perform correctly on these systems.

Included in the distribution is a program called **pChaplin** , which facilitates piping command line options to be main program. This allows users to run Chaplin as a part of a batch script, if needed. Further details of this can be found in Chapter 4.2 of this document.

The **CHAPLIN** website is http://www.genetics.emory.edu/labs/epstein/software/chaplin/index.html. The windows executable, documentation, and some sample files are contained in standard zip-compressed packages that can be downloaded from there. When one of these files is uncompressed, the documentation and sample data will be in written to the respective directories, `doc/` and `sample_data/`. In particular, the 4-SNP sample genotype data set, `sample.dat`, is used in the example analysis described in this documentation.

The program is a self-contained executable and does not require any complicated installation. The executable can be placed in and run from any location that is accessible to the user.

1.5 User Support

While we have tried to carefully document **CHAPLIN** and make it easy to use, problems may occur. Some problems can likely be resolved by reading through this documentation. In the event that problems do occur with the software, we would appreciate being notified so that these problems may be rectified. Such problems or questions either about running **CHAPLIN** or the statistical methodology involved should be directed to *statgen@genetics.emory.edu*.

1.6 How to cite this work

If you use **CHAPLIN** to analyze data for publication, we ask that you make the following citations:

- M.P. Epstein and G.A. Satten. Inference on haplotype effects in case-control studies using unphased genotype data. *Am. J. Hum. Genet.*, 73:1316-1329, 2003.
- Duncan R.D., Epstein M.P., Satten G.A. Case-Control Haplotype Inference (CHAPLIN). Version 1.2, September 2006.

2 Inputting Genotype Data

2.1 Data format

CHAPLIN requires the sample disease and genotype data to be in either a space- or tab-delimited text file. Each row of the file consists of a disease status and genotype data for a specific subject in a series of columns that are organized in the following fashion:

Column	Data value
1	Disease status (1=case, 0=control)
2	Genotype code at marker 1
3	Genotype code at marker 2
⋮	⋮
$N+1$	Genotype code at marker N

A data set with N markers will have $N+1$ columns of values. For allele information at a given marker, the set of genotype codes is summarized in Table 2.1.

Table 2.1: Summary of genotype codes for a given pair of alleles at a single locus. Missing data must be assigned a genotype code of -1. Columns 2 through $N+1$ in the genotype data file each must contain exactly one of these values.

Allele 1	Allele 2	**Genotype code**
0	0	**0**
0	1	**1**
1	0	**1**
1	1	**2**
missing data		**-1**

A portion of a 4-locus dataset is shown in Table 2.2. The top two rows in this set contain control data (first column value is zero) and the remaining three rows contain case data (first column value is one). The control subject in the first row is homozygous at each marker. The second control subject is heterozygous at the first marker and homozygous at the remaining three markers. The case subject in the fourth row has missing data at the fourth marker, indicated by the value -1.

Table 2.2: Part of a sample genotype data set showing two control subjects and three case subjects.

Case-control indicator	Marker 1	2	3	4
0	0	2	0	2
0	1	2	2	0
1	2	0	1	0
1	1	1	2	-1
1	2	0	2	0

2.2 Data loading

One can load an input data file into **CHAPLIN** in any of the following ways:

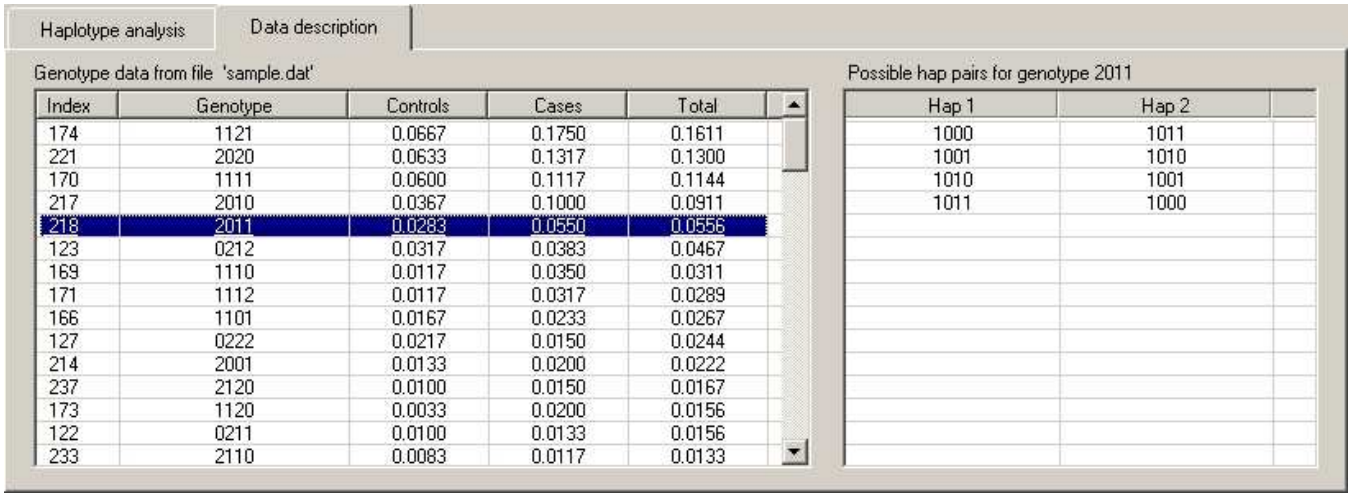
1. Clicking the 'Load data' button in the controls bar, shown in Figure 2.1;
2. On the 'File' menu, choosing the 'Load Genotype Data' submenu item;

The first two methods will initiate a standard 'File Open' dialog. From this dialog, select your data file for loading.

Figure 2.1: This controls bar contains many of the objects used to interact with **CHAPLIN** .

2.3 Data description

Upon successful data loading, **CHAPLIN** provides the user with a summary of the input genotype data. To examine this summary, click the 'Data Description' tab. An example summary is shown in Figure 2.2.



Index	Genotype	Controls	Cases	Total
174	1121	0.0667	0.1750	0.1611
221	2020	0.0633	0.1317	0.1300
170	1111	0.0600	0.1117	0.1144
217	2010	0.0367	0.1000	0.0911
218	2011	0.0283	0.0550	0.0556
123	0212	0.0317	0.0383	0.0467
169	1110	0.0117	0.0350	0.0311
171	1112	0.0117	0.0317	0.0289
166	1101	0.0167	0.0233	0.0267
127	0222	0.0217	0.0150	0.0244
214	2001	0.0133	0.0200	0.0222
237	2120	0.0100	0.0150	0.0167
173	1120	0.0033	0.0200	0.0156
122	0211	0.0100	0.0133	0.0156
233	2110	0.0083	0.0117	0.0133

Hap 1	Hap 2
1000	1011
1001	1010
1010	1001
1011	1000

Figure 2.2: The 'Data Description' tab provides the user with a summary of the input genotype data separated into cases and controls.

The first column displays a unique index for each genotype, which ranges in value from 0 to $4^N - 1$, where N denotes the number of markers in the sample. The second column lists the multi-marker genotype using the coding described in Table 2.1 on Page 3 with the exception that, for ease of reading, missing data are now represented by '*' rather than '-1'. Columns 3-5 shows the number of subjects with a particular genotype in the controls, cases, and pooled sample, respectively. The frequency of genotypes in each sample can be viewed by clicking the 'Frequency' radio button under the list view. In frequency mode, the number of displayed digits is specified using the '# digits' box that appears adjacent to the 'Frequency' button.

Clicking a column header causes the data to be sorted on the entries within that column. Subsequent clicks of the same column header will alternate the sorting between descending and ascending modes.

Clicking on a specific genotype index in the left list view results in a complete set of possible haplotype pairs consistent with that genotype appearing in the right list view. This includes both permutations of the '0/1' and '1/0' genotypes as well as all possibilities consistent with the missing data.

3 Haplotype Analysis

3.1 Sample haplotype frequencies

After data loading, **CHAPLIN** immediately estimates haplotype frequencies separately in the case and control samples. Examination of these estimated haplotype frequencies can help guide the choice of subsequent haplotype models. Page 1321 of Epstein and Satten (2003) provides an illustration of this idea.

To examine the estimated haplotype frequencies in the control and case samples, click the 'Haplotype Analysis' tab. An example table of haplotype frequencies is shown in Figure 3.1.

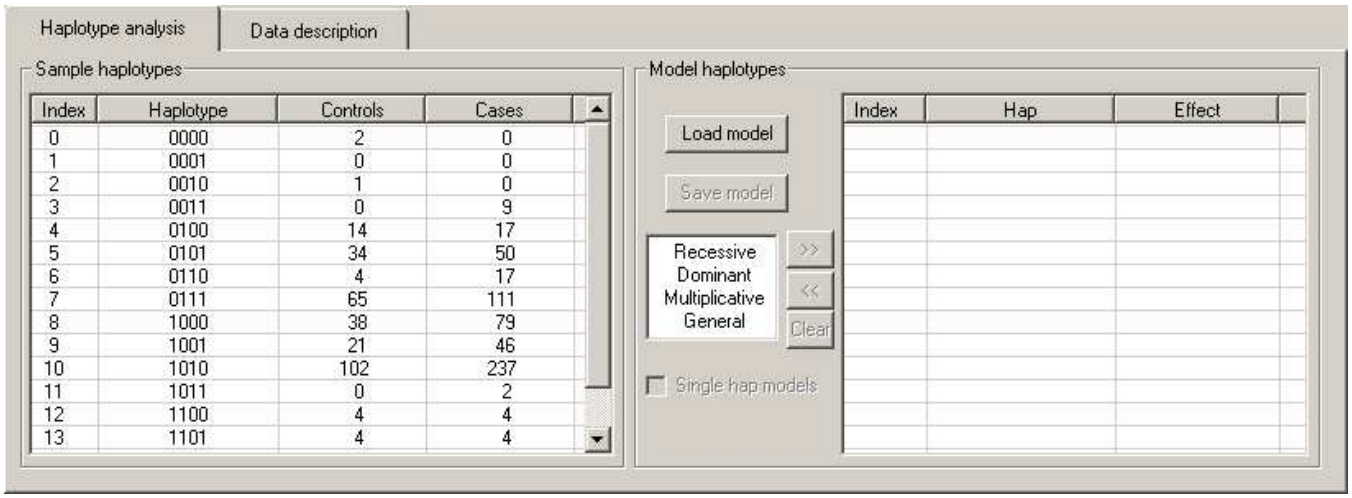


Figure 3.1: The 'Haplotype analysis' tab shows estimated haplotype frequencies and allows the user to interactively specify a model.

The first column displays a unique index for each haplotype, which ranges in value from 0 to $2^N - 1$, where N denotes the number of markers in the sample. The second column shows the specific N-marker haplotype. Columns 3 and 4 show the estimated frequency of the haplotype in the control and case sample, respectively. By default, **CHAPLIN** only displays haplotypes that have a non-zero frequency in at least one of the two samples. To show all haplotypes, click the 'Show All' box under the list view. To control the number of displayed digits in the haplotype frequency, use the '# digits' box.

3.2 Haplotype models

Using the initial haplotype frequencies as a guide, one can then construct models that examine the effects of various haplotypes on disease status. To model haplotype effects on disease, **CHAPLIN** requires the specification of the odds of disease given a haplotype pair (h, h'). This disease odds can be parameterized as $\theta_{hh'} = \exp(X_{hh'}\beta)$, where β is a list of disease-risk parameters (to be estimated) and $X_{hh'}$ is a list of numerical design codes that relate the haplotype pair (h, h') to β. Modeling the effects of specific haplotypes on the disease odds is accomplished by appropriate coding of $X_{hh'}$ for all haplotype pairs (h, h'). This can be achieved interactively through **CHAPLIN** or by directly inputting a prewritten text file into the program. These methods are discussed in the next two sections.

3.2.1 Model input: interactive mode

Specification of $X_{hh'}$ in interactive mode can be accomplished through the graphic user interface (GUI) by adding and/or removing haplotypes to the model list. Furthermore, once a model is developed, it can be saved for reuse. These functions are now described.

Adding Haplotypes to a Model

1. Click on the first haplotype that you wish to model in the left region of the 'Haplotype Analysis' tab.
2. Click on the desired type of haplotype effect by choosing the appropriate one from the 'Effect' box found in the central region of the 'Haplotype Analysis' tab. Possible effects consist of:
   - **Recessive effect**: subjects with one copy of the haplotype have the same risk as those subjects with no copies.
   - **Dominant effect**: subjects with one copy of the haplotype have the same risk as those subjects with two copies.
   - **Multiplicative effect**: subjects with one copy of the haplotype are at an intermediate risk (on log scale), with respect to those subjects with zero or two copies.
   - **General effect**: subjects with one copy of the haplotype have a general change in risk compared to those with zero or two copies.
3. Once an effect is chosen, press the 'Add' button below the 'Effect' box. The selected haplotype with the desired effect should now appear in the 'Model haplotypes' listing. If more than one haplotype is selected, all selected haplotypes will be added to the model with the same effect type.
4. Repeat steps 1-3 for other haplotypes until you have obtained the desired model

By performing steps 1-4, one specifies the appropriate $X_{hh'}$ that **CHAPLIN** uses for analysis.

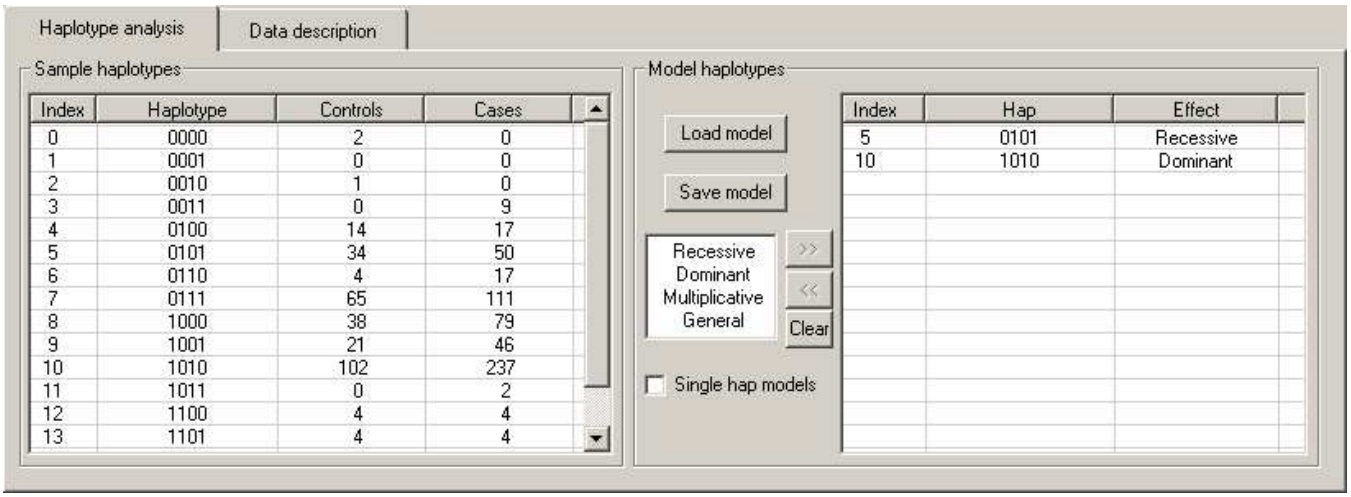


Figure 3.2: The 'Model haplotypes' list is populated with a recessive haplotype 5 and a dominant haplotype 10.

Removing Haplotypes From a Model

If you wish to remove one or more haplotypes from the overall model in the 'Model haplotypes' listing:

1. Select the haplotype(s) to remove within the model list
2. Click the 'Remove' button below the 'Effect' box.

Clearing a Haplotype Model

To remove all haplotypes from the 'Model haplotypes' listing, click the 'Clear' button below the 'Effect' box.

Saving a Haplotype Model

An existing model can be saved to a file and retrieved for later use (see Section 3.2.2.). A model can be saved in either of the following ways:

- Clicking on the 'Save' button below the 'Model haplotypes' listing
- On the 'Model' menu, selecting the 'Save model to file' option

Either action will initiate a standard 'Save As' dialog. Use this dialog to specify a filename for the model and subsequently save it to disk.

3.2.2 Model input: file mode

A haplotype model can be input into **CHAPLIN** by directly specifying $X_{hh'}$ within a space-delimited text file. To model a specific haplotype h^* on $\theta_{hh'}$, code $X_{hh'}$ according to the type of effect that h^* has on the disease odds. Examples of different design codes for different types of effects of h^* are shown below:

Recessive effect:
$$X_{hh'} = \begin{cases} 1 & h = h' = h^*, \\ 0 & \text{otherwise;} \end{cases}$$

Dominant effect:
$$X_{hh'} = \begin{cases} 1 & h = h^* \quad \text{and/or} \quad h' = h^*, \\ 0 & \text{otherwise;} \end{cases}$$

Multiplicative effect:
$$X_{hh'} = \begin{cases} 2 & h = h' = h^*, \\ 1 & h = h^* \quad \text{or} \quad h' = h^*, \\ 0 & \text{otherwise;} \end{cases}$$

For a general effect model of h^*, $X_{hh'}$ consists of two elements $X_{hh'} = (X_{hh',1}, X_{hh',2})$ that model the separate effects of the first and second copies of h^* on $\theta_{hh'}$. In this case, the coding becomes:

General effect:
$$X_{hh'} = (X_{hh',1}, X_{hh',2}) = \begin{cases} (1,1) & h = h' = h^*, \\ (1,0) & h = h^* \quad \text{or} \quad h' = h^*, \\ (0,0) & \text{otherwise;} \end{cases}$$

In addition to a general effect model, $X_{hh'}$ will also consist of more than a single element when modeling the effects of multiple haplotypes on the odds of disease. For example, when modeling both a recessive effect of haplotype h^* and a general effect of haplotype h^{**} on $\theta_{hh'}$, $X_{hh'}$ will have three elements; one for the recessive effect of h^* and two for general effect of h^{**}.

The text file must conform to a specific format. The first line of the file is an integer corresponding to the number of elements in $X_{hh'}$. Subsequent lines are organized into a series of space-delimited columns that are organized in the following fashion:

Column	Data value
1	Index of first haplotype h
2	Index of second haplotype h'
3	Design code for first element of $X_{hh'}$
4	Design code for second element of $X_{hh'}$
$\vdots$	$\vdots$
$P+2$	Design code for P^{th} element of $X_{hh'}$

The design code corresponding to any haplotype pair, (h,h'), that has no effect on disease is $X_{hh'} = 0$; these entries do not require explicit assignment within the file, as they will be assigned a value of zero by default. Furthermore, only a single specification for any haplotype pair should be provided, as **CHAPLIN** will automatically symmetrize the design code matrix. For example, if you specify the coding for $h = 1$ and $h' = 2$, you do not need to specify the coding for $h = 2$ and $h' = 1$.

As an example, suppose we have a dataset consisting of 4-SNP haplotypes, with the 16 possible haplotypes indexed by 0,1,...,15. We wish to construct a 2-parameter model that first fits the recessive effect of haplotype 5 and then fits the dominant effect of haplotype 10. The corresponding model file is shown in Figure 3.3.

```
 2
 0  10  0.0000000000  1.0000000000
 1  10  0.0000000000  1.0000000000
 2  10  0.0000000000  1.0000000000
 3  10  0.0000000000  1.0000000000
 4  10  0.0000000000  1.0000000000
 5   5  1.0000000000  0.0000000000
 5  10  0.0000000000  1.0000000000
 6  10  0.0000000000  1.0000000000
 7  10  0.0000000000  1.0000000000
 8  10  0.0000000000  1.0000000000
 9  10  0.0000000000  1.0000000000
10  10  0.0000000000  1.0000000000
10  11  0.0000000000  1.0000000000
10  12  0.0000000000  1.0000000000
10  13  0.0000000000  1.0000000000
10  14  0.0000000000  1.0000000000
10  15  0.0000000000  1.0000000000
```

Figure 3.3: File data for a model with a recessive effect haplotype 5 and a dominant effect haplotype 10. This file was created from the with the 'Save' button in the **CHAPLIN** GUI.

The first line of this file specifies the number of parameters. Here, there are two parameters; one corresponding to the recessive effect of haplotype 5 and one corresponding to the dominant effect of haplotype 10. Subsequent lines specify the haplotype pair indices along with the corresponding design code components as described in the text above. The model file shown in Figure 3.3 was created and outputted using the **CHAPLIN** interactive interface. Model files that are directly output from **CHAPLIN** in this way (see Section 3.2.1) will automatically conform to the required format; therefore, using the interactive interface is the most user-friendly and reliable method of creating a model file.

3.2.3 Joint versus single haplotype analyses.

Once a haplotype model is specified, one of two different types of analyses can be run. The first analysis option is a joint analysis that models all haplotype effects simultaneously on disease. This is the default analysis option in **CHAPLIN** .

The second analysis option is to perform a series of single haplotype analyses on each of the haplotype effects specified in the 'Model haplotypes' listing. Therefore, if the 'Model haplotypes' listing consists of N haplotype effects, then **CHAPLIN** will run N different analyses-one for each effect specified in the listing. Such an analysis could be useful as an introductory step for identifying haplotypes of interest to include in more complicated models. To run single haplotype analyses, click the 'Single hap models' button under the 'Model haplotypes' listing.

3.2.4 Initiating an analysis

Once a model is inputted and the analysis type is chosen, **CHAPLIN** can run the analysis by pushing the 'Run' button located in the upper-right corner of the main window. Results from the analysis will appear in the Output window (for output description, see Section 3.3).

3.2.5 Program options

Under the 'Edit' menu, selection of the submenu item 'Options' will display a dialog window with user adjustable parameters, as shown in Figure 3.4.

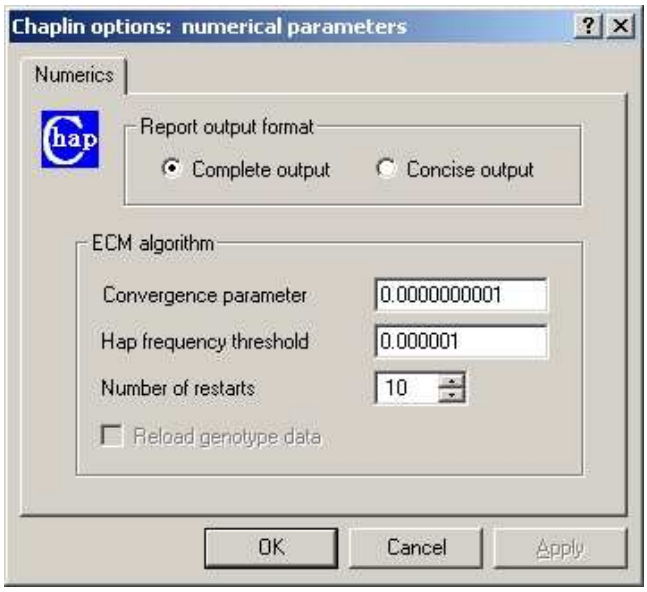


Figure 3.4: Popup dialog containing options available to the user.

Report output The radio buttons within the group titled 'Report output format' allow selection between concise and verbose report output (see Section 3.3). A concise report contains only information about the haplotype-specific effects and the global tests. A complete report contains this same information along with a breakdown of case-control counts, and a list of observed haplotype frequencies with each respective score vector component.

ECM algorithm parameters This group contains numerical parameters specific to the ECM algorithm routines.

- **Convergence parameter,** ε: Used to help declare convergence of the ECM algorithm. If the square root of the sum of squares of the parameter values at successive iterations is less than ε, **CHAPLIN** assumes the ECM has converged. The default value of the parameter is $\varepsilon = 10^{-10}$. Increasing this value will speed up computation time but may lead to convergence of parameter values to satellite points and not the true maxima.
- **Hap frequency threshold**: Any haplotype with estimated frequency below this threshold will be effectively assigned a frequency value of zero. Increasing this value will speed up the calculation time, but may result in missing rare haplotypes that have an effect on disease.
- **Number of restarts**: Parameter estimates for the ECM algorithm may converge to satellite points rather than true maxima, which can yield unreliable results. To protect against convergence to satellite points, one should restart the ECM algorithm multiple times at random initial parameter values. **CHAPLIN** performs this procedure assuming 10 restarts by default. Decreasing the number of restarts decreases the computation time but may increase the possibility of convergence to satellite points.

If genotype data has been loaded prior to making changes to this section, the EM algorithm for the null model analysis can be re-run with the new settings upon closing this dialog by checking the 'Reload genotype data' box.

3.2.6 Identifiability issues

CHAPLIN will attempt to run any model specified in the 'Model haplotypes' listing specified by the user. However, we recommend that one carefully considers the chosen model before analysis. An overly complicated model may lead to non-identifiable estimates of parameters, leading to unreliable results. To ensure the chosen model yields identifiable parameter estimates, one can use the conditions described in Appendix A of Epstein and Satten (2003).

CHAPLIN will also yield non-identifiable estimates of a particular haplotype effect if the modeled haplotype has a frequency of zero in the control sample. In this situation, one can still model the haplotype's effect on disease, but it requires swapping the disease outcome value of cases and controls in the dataset by hand and subsequently modeling the effect of the haplotype on being a control (for a more thorough description of this idea, please read page 1322 of Epstein and Satten (2003).

3.3 Output

Once an analysis is run, results will appear in the 'Output' window

3.3.1 Joint haplotype analysis

CHAPLIN first provides results for the effects of individual haplotype effects in the joint haplotype model. The output is organized in the following way:

- Column 1: haplotype index;
- Column 2: haplotype;
- Column 3: haplotype effect type;
- Column 4: β, the estimated haplotype effect;
- Column 5: standard error of β;
- Column 6: Wald statistic for testing whether $\beta = 0$ and the associated p-value;
- Column 7: R^2_h, the haplotype uncertainty measure of Stram et al. (2003).

Below these results, an estimate of the fixation index in the control sample is presented. Non-zero values of the fixation index provide evidence that haplotype frequencies in the control sample are not in Hardy-Weinberg equilibrium.

The next items presented are the robust score and likelihood-ratio (LR) statistics for testing the global null hypothesis that all model haplotypes have no effect on disease. For a properly-specified model, the LR statistic will be a more powerful test than the robust score statistic. However, the robust score statistics is less sensitive to model misspecification than the LR statistic and may be preferable if one is not confident in the model choice.

Reported next is the Akaike Information Criterion (AIC) of Akaike (1985), which can be used for model selection. Models with smaller AIC values are preferable over models with larger values.

Finally, a summary of the data is provided. First, the number of cases and controls in the sample are shown. Next, **CHAPLIN** presents a list of estimated haplotype frequencies under the global null model (all haplotypes independent of disease) as well as each haplotype's score vector component (all score vector values should approximately equal zero; non-zero values indicated that the haplotype frequencies may not have converged properly).

3.3.2 Single haplotype analysis

For a single haplotype analysis, **CHAPLIN** lists the following output for each haplotype:

- Column 1: haplotype index ;
- Column 2: haplotype;
- Column 3: haplotype effect type;
- Column 4: β, the estimated haplotype effect;
- Column 5: Wald statistic for testing whether $\beta = 0$ and the associated p-value;
- Column 6: AIC, Akaike Information Criterion;
- Column 7: R^2_h, the haplotype uncertainty measure of Stram et al. (2003).

As mentioned earlier, the analysis of each haplotype is not adjusted for the effects of the other haplotypes in the model.

3.3.3 Saving results

The output produced by **CHAPLIN** can be saved to a file by selecting the "Save output to file" submenu item under the "File" menu. This will cause a standard "Save as" dialog to appear whereby a directory and filename may be selected for the output data.

3.3.4 Rare haplotypes

Epstein and Satten (2003) showed that asymptotic inference tended to be unreliable for estimating and testing the effects of rare haplotypes on disease. As all p-values in **CHAPLIN** are based on asymptotic theory, we caution the user to be careful in interpreting results in these situations. **CHAPLIN** will generate a warning message detailing this caution when one models rare haplotypes (frequency < 0.002).

In these situations, we recommend the use of permutation or bootstrap methods for proper analysis. Such methods are not currently available in **CHAPLIN** but will be made available in a future version of the program.

4 Running Chaplin

4.1 Graphic user interface (GUI)

The principle means for a user to interact with Chaplin is through the provided GUI. For convenience, controls for loading and viewing genotype data, specifying haplotype models, initiating analyses, and saving output are provided. These various controls are described throughout this manual.

4.2 Command line interface

Provided in the downloaded package is a wrapper program called **pChaplin** that assists users seeking to run **CHAPLIN** analyses on the command line or in batch scripts. Using **pChaplin** , one can run **CHAPLIN** on a DOS command-line interface. **pChaplin** does not provide an interactive shell, but rather inputs in SNP and disease data from the command line and outputs data to a specified results file using user-specified options. Note that **pChaplin** will not be able to properly facilitate a **CHAPLIN** analysis through a Cygwin terminal.

A summary of these options can be displayed in the terminal by typing the command **pChaplin −−help**, as shown in Figure 4.1.

```
C:\Chaplin>pChaplin.exe --help
Usage: pChaplin [OPTION] [FILE] [-]
-v, --version         show Chaplin version and exit
-h, --help            show this help
-b, --batch           close Chaplin after completing analysis
-g, --genfile FILE    file containing genotype data
-m, --modelfile FILE  file containing haplotype model
-o, --outfile FILE    file for output
--hap INTEGER         single-haplotype index
--model INTEGER       1=Rec, 2=Dom, 3=Mul, 4=Gen

Case-control haplotype inference package:  Chaplin 1.2.2
http://www.genetics.emory.edu/labs/epstein/software/chaplin/index.html
Report bugs to <statgen@genetics.emory.edu>
```

Figure 4.1: Summary of available command line options displayed using the command line **−−help** option.

4.2.1 Command line options

−−batch
This command runs Chaplin in batch mode, thereby suppressing the GUI interface. If the **−−batch** option is not specified, Chaplin will display the GUI interface.

−−genfile FILE
This option specifies the filename of the input genotype data, as described in Chapter 2.1.

−−modelfile FILE
This option specifies the filename for the haplotype model, which is described in Chapter 3.2.2. Note that models involving multiple haplotypes must be specified in this way.

−−outfile FILE
This option specifies the name of the output file containing results of the analyses. If no name is specified, the default name of the output file will be 'output.dat'. If one specifies the same output file in multiple batch analyses, that file will be overwritten.

−−hap INTEGER
For single-haplotype models only, this option specifies the index of the haplotype of interest (these indexes are listed in the GUI framework). This option is only used when one specifies the **−−model** option.

−−model INTEGER
For single-haplotype models only, this option specifies the model of the particular haplotype of interest. The coding of the model is shown in the following table:

Integer	Model
1	Recessive
2	Dominant
3	Multiplicative
4	General

This option is only used when the **−−hap** option is also specified.

Examples

In the following examples, suppose the file *genotypes.dat* contains genotype data and the file *model.des* contains the haplotype model.

To run Chaplin in batch mode using the genotype file and the model file, while outputting results to a file called *chaplin-analysis.out*, issue the following command:

```
pChaplin --batch --genfile genotypes.dat --modelfile model.des --outfile chaplin-analysis.out
```

To run Chaplin in batch mode assuming a single-haplotype model that assumes a dominant effect for haplotype 5, we can use the command:

```
pChaplin --batch --genfile genotypes.dat --hap 5 --model 2 --outfile chaplin-analysis.out
```

4.2.2 Errors while in batch mode

In the event of an error in batch mode (e.g., data files not properly formatted, filenames misspecified), Chaplin will write an error summary to the output file and subsequently exit.

5 Example Analysis

5.1 Many single-haplotype analyses

5.1.1 Input genotype data

Load sample data

Load the sample genotype data set using either the submenu or the ”Load data” button. The null model frequencies broken down by cases and controls will be displayed in the left listview under the ”Haplotype analysis” tab.

Sort haplotype frequencies by controls

Sort this data by control-group frequencies by clicking the ”Control Freq” column header.

5.1.2 Model selection

Add haplotypes to model input

Next, select all haplotypes having non-zero frequencies in both the cases and controls; then, select ”Multiplicative” from the Effects listview.

Push the enabled ”Add” button to add these haplotypes to the model input listview. In this preliminary analysis, all of these haplotypes will be modeled independently of each other to ascertain which, if any, haplotypes have significant risk associations. Select the ”Single hap models” check box.

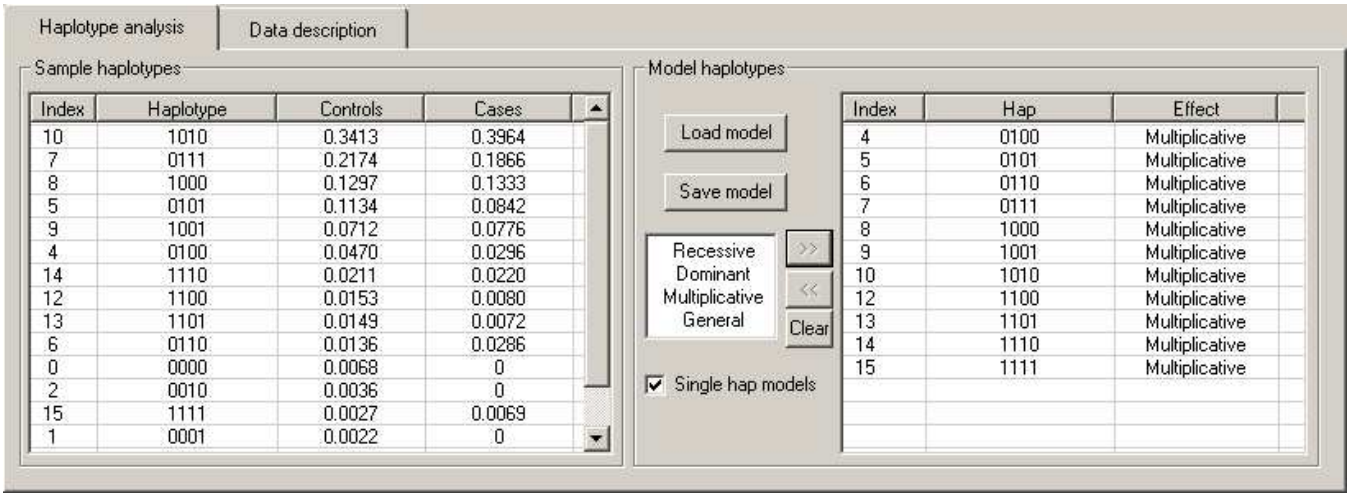


Figure 5.1: Ready to run preliminary analysis of every individual non-zero control frequency haplotypes assuming a multiplicative effect for each. The list of haplotypes is sorted on control frequencies in descending order.

Run the program

Push the ”Run” button to begin the analysis. Output from this sample run is shown in Figure 5.2. From this preliminary analysis, haplotypes 5 and 10 exhibit the greatest significance based on their Wald statistics and low AIC values. Based on these, we perform further analysis on these two haplotypes.

```
Individual haplotype model results:
-------------------------------------------------------------------------------
Index      Hap      Effect   Beta        Wald (p-value)        AIC       Rh^2
------  ----------  ------  -------  --------------------  ----------  -------
     4   0 1 0 0     Mul    -0.5393     1.8786 (0.060305)     5338.31   0.7985
     5   0 1 0 1     Mul    -0.4727     2.4948 (0.012604)     5335.67   0.7509
     6   0 1 1 0     Mul     0.6101     1.2913 (0.196595)     5339.87   0.7225
     7   0 1 1 1     Mul    -0.1679     1.2654 (0.205733)     5340.19   0.8689
     8   1 0 0 0     Mul    -0.0330     0.1977 (0.843260)     5341.73   0.7763
     9   1 0 0 1     Mul     0.1053     0.5168 (0.605296)     5341.50   0.8982
    10   1 0 1 0     Mul     0.2970     2.6898 (0.007149)     5334.43   0.9002
    12   1 1 0 0     Mul    -0.8820     1.3543 (0.175639)     5339.67   0.6229
    13   1 1 0 1     Mul    -0.7696     1.2693 (0.204328)     5340.11   0.6669
    14   1 1 1 0     Mul    -0.0052     0.0134 (0.989340)     5341.77   0.7784
    15   1 1 1 1     Mul     0.7971     0.7290 (0.465985)     5341.11   0.6994
-------------------------------------------------------------------------------
```

Figure 5.2: Output from running single haplotype analysis on all haplotypes with non-zero frequencies in both the case and control groups.

5.2 More detail on significant haplotypes

New model input

We next assess the most likely effects that haplotypes 5 and 10 have on disease. To do this, we fit each haplotype separately under different effect models and select the most appropriate model as that with the smallest AIC value.

Clear the current model input using the ”Clear” button. Add haplotype 5 with a recessive effect to the model inputs list view. With only one haplotype populating the model inputs list view, the ”Single hap models” checkbox is not enabled. Rerun the program. The concise set of output (see Section 3.2.5) from this run is shown in Figure 5.4.

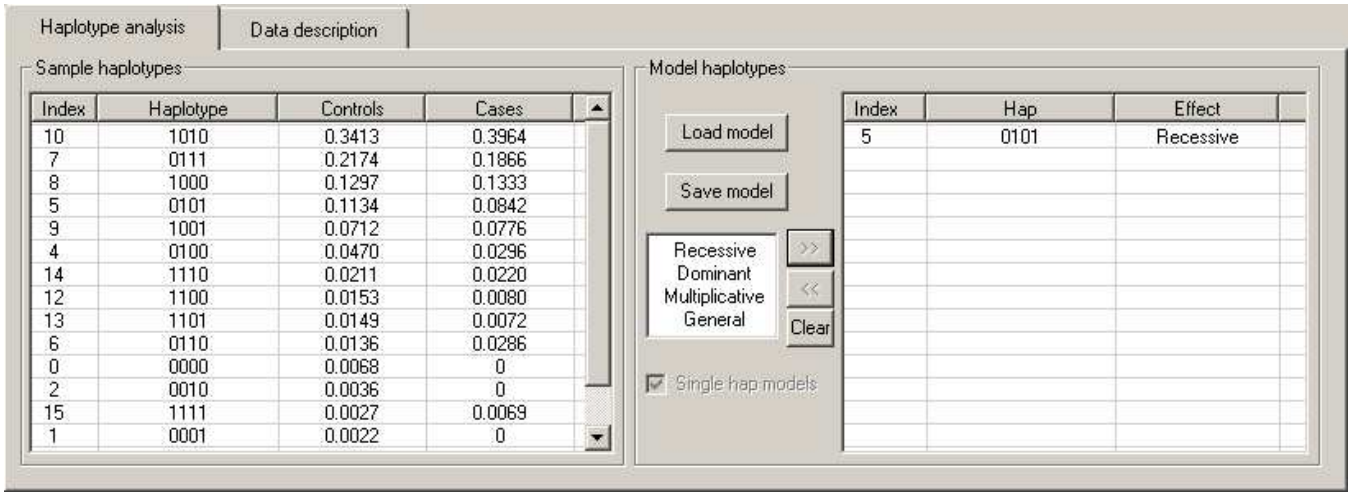


Figure 5.3: Ready to run analysis only with a recessive haplotype 5.

```
Summary of haplotype-specific effects:
-------------------------------------------------------------------------
Index      Hap      Effect    Beta       SE       Wald (p-value)      Rh^2
------  ----------  ------  ---------  -------  ------------------  ------
     5   0 1 0 1     Rec     -0.67439   0.6212  -1.0856 (0.277660)   0.9949
-------------------------------------------------------------------------

Global tests of haplotype effects:
---------------------------------------------------
fixation index:                            0.0000000
Akaike information criterion (AIC):          5340.35
---------------------------------------------------
Robust score (p-value)       LR (p-value)
---------------------  ---------------------
   1.9222 (0.165610)       1.4248 (0.232608)
```

Figure 5.4: Concise output generated by running the program with haplotype 5 under a recessive effect.

Repeat this process for Dominant, Multiplicative, and General effects at haplotype 5. Based on the AIC value, it appears that haplotype 5 has a multiplicative effect on disease. A summary of these analyses is provided in Table 5.1.

Table 5.1: Results from analyzing haplotype 5 for each possible effect type.

Effect	AIC
Recessive	5340.35
Dominant	5337.24
Multiplicative	5335.67
General	5337.42

We next repeat the same procedure as above using haplotype 10. Results are summarized Figure 5.2. Based on the AIC value, it appears that haplotype 10 has a dominant effect on disease.

Table 5.2: Results from analyzing haplotype 10 for each possible effect type.

Effect	AIC
Recessive	5341.51
Dominant	5327.46
Multiplicative	5334.43
General	5329.46

5.3 Joint model analysis

In the initial analysis, the set of likelihood ratio statistics suggest that both haplotypes 5 and 10 may be significant. The individual analyses on these haplotypes for each of the four effects (summarized in Tables 5.1 and 5.2) suggests analyzing a joint model consists of haplotype 5 with a multiplicative effect and haplotype 10 with a dominant effect.

The results of this joint model are shown in Figure 5.5.

```
Summary of haplotype-specific effects:
------------------------------------------------------------------------
Index      Hap      Effect    Beta       SE       Wald (p-value)      Rh^2
------  ----------  ------  ---------  -------  ------------------  ------
     5   0 1 0 1     Mul     -0.21805   0.2037  -1.0702 (0.284533)   0.7509
    10   1 0 1 0     Dom      0.45479   0.1509   3.0147 (0.002572)   0.8135
------------------------------------------------------------------------

Global tests of haplotype effects:
--------------------------------------------------
fixation index:                          0.0029762
Akaike information criterion (AIC):        5328.32
--------------------------------------------------
Robust score (p-value)       LR (p-value)
----------------------  ----------------------
  14.6366 (0.000663)      15.4513 (0.000441)
```

Figure 5.5: Output from running a joint analysis on haplotypes 5 multiplicative and 10 dominant.

Note that in this joint model, haplotype 5 under a multiplicative effect no longer appears to be significant when adjusting for the effects of haplotype 10 under a dominant effect. However, haplotype 10 still is significant in this joint model. This suggests that the disease variant is located on this haplotype and suggests a genomic region to further investigate using positional cloning. Note also that the AIC measure of this joint model (5328.33) is higher than a model that includes only the dominant haplotype 10 (5327.46).

Glossary

This is a summary list of abbreviations and acronyms used in this document.

AIC	Akaike information criterion
CI	confidence interval
ECM	expectation-conditional-maximization algorithm
EM	expectation-maximization algorithm
GUI	graphic user interface
HWE	Hardy-Weinberg equilibrium
IMSL	International Mathematical and Statistical Libraries
LD	linkage disequilibrium
LR	likelihood ratio
SNP	single nucleotide polymorphism

Index

Bibliography

Akaike, H. (1985). Prediction and entropy. In A. Atkinson and S. Fienberg (Eds.), *A celebration of statistics*, pp. 1–24. Springer, New York.

Epstein, M. and G. Satten (2003). Inference on haplotype effects in case-control studies using unphased genotype data. *Am. J. Hum. Genet. 73*, 1316–1329.

Satten, G. and M. Epstein (2004). Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Am. J. Hum. Genet. 27*(3), 192–201.

Stram, D., C. Haiman, J. Hirschhorn, D. Altshuler, L. Kolonel, B. Henderson, and M. Pike (2003). Choosing haplotype-tagging snps based on unphased genotype data from a preliminary sample of unrelated subjects with an example from the multiethnic cohort study. *Hum. Herid. 55*, 27–36.