

SimHap: A comprehensive modelling framework and a multiple-imputation approach to haplotypic analysis of unrelated individuals

GUI Release v1.0.2: User Manual
January 2009

If you find this software useful, please refer to:

Carter KW, McCaskie PA, Palmer LJ (2008). SimHap GUI: An intuitive graphical user interface for genetic association analysis. *BMC Bioinformatics*. 2008 Dec 25;9(1):557.

This program is free for non-commercial use only. It is distributed in the hope that it will be useful, but without any warranty.

© McCaskie PA, Carter KW, Palmer LJ (2004)

Contents

1. Introduction	1
2. Installing SimHap	2
2.1. Installing Java 1.5+ on your computer	2
2.2. Installing R 2.4.0+ on your computer	2
2.3. Install SimHap on your computer	2
3. Input Files and Data Preparation	3
3.1. Formatting SNP Data	3
3.2. Missing Data	3
4. Example Data Sets	4
4.1. Data Set 1: Binary and Quantitative Outcomes	4
4.2. Data Set 2: Longitudinal Outcomes	5
4.3. Data Set 3: Right-censored Outcomes	5
5. Running SimHap	6
5.1. Haplotype Analysis	8
5.1.1. Generating haplotypes	8
5.1.2. Defining the model	9
5.1.3. Specifying model parameters	12
5.1.4. SimHap output	13
5.2. Single SNP Analysis	14
5.3. Epidemiological Analysis	15
6. SimHap Licence and Referencing the Software	15
References	16

1. Introduction

When dealing with phase ambiguous genotype data, as is often the case with population data, an individual's haplotype pair (or diplotype) may not be known with certainty. For individuals homozygous at each loci of interest or homozygous at all but one locus, diplotypes may be determined with certainty. For individuals heterozygous at more than one locus and without the genotypic information of close relatives inference is often used to determine possible haplotypes. When inferring diplotypes for individuals with ambiguous phase (e.g. from phase unknown genotype data), uncertainty is inherent. It is common practice to infer possible diplotypes for an individual and calculate the likelihood of each diplotype. Often, the most likely diplotype is treated as known and used in association analyses. Where one diplotype is clearly most probable, this method may be adequate but it may become decreasingly reliable when multiple diplotypes are possible and the likelihoods of these possibilities are close. **SimHap** is a program that we have developed to impute haplotype frequencies at the individual level using biallelic SNP genotype data. **SimHap** also tests for haplotype associations with outcomes of interest while using simulation to incorporate the uncertainty around inferred haplotypes into the modelling procedure.

SimHap allows simple epidemiological, single SNP and haplotype association analyses of quantitative, binary, longitudinal and right-censored outcomes under a range of genetic models. **SimHap** can accommodate large data sets, and can model genetic and environmental effects, including complex haplotype:environment interactions. **SimHap** features cross-platform functionality via Java, and a sophisticated graphical user interface (GUI), so you need not have a comprehensive knowledge of statistical modelling or command line operation to perform complex analyses. This approach uses current estimation-maximisation based methods for the estimation of haplotypes from unphased genotype data¹ and incorporates simulation techniques to model haplotypic associations in population-based samples.

SimHap will also perform association analyses on more simple epidemiological data, with or without the inclusion of SNPs or haplotypes. The current implementation of **SimHap** has been written to utilise the statistical computing package R² when resolving haplotypes; all possible haplotype configurations are resolved for each individual within the program itself, and the posterior probability of each configuration is calculated. This information is then passed into a generalised-linear modelling, linear mixed effects or Cox proportional hazards framework where (using simulation to deal with the uncertainty around the imputed haplotypes) association tests are performed.

Example datasets with quantitative, binary, longitudinal and right-censored outcomes can be downloaded from <http://www.genepi.org.au/simhap>ⁱ.

ⁱ **Note:** any reference to example data (e.g. variable names or values) or user input throughout this manual is displayed in fixed width font e.g. Aa or `simhap.bat`.

2. Installing SimHap

SimHap v1.0.0 can be obtained by following the download link from the following website: <http://www.genepi.org.au/simhap>. In order to successfully run **SimHap**, you must first have Java 1.5+ and R 2.4.0+ installed.

2.1. Installing Java 1.5+ (5.0+) on your computer

You must have Java (Runtime or Standard Edition) 1.5 or later installed to use **SimHap**. Check if Java is already installed on your machine by starting a command prompt (or shell) and typing:

```
java -version
```

If Java is not found, Java is either not installed or not set up correctly on your machine. The appropriate version of Java Standard Edition or Runtime Environment can be downloaded from <http://java.sun.com/javase/downloads/index.jsp>.

Please check with your IT support staff if you are unsure how to install Java.

2.2. Installing R 2.4.0+ on your computer

You must also have R 2.4.0 or later installed on your computer in order to run **SimHap**. If you do not have a correct version of R, you can download a version for your operating system from <http://cran.r-project.org/>. Please check with your IT support staff if you are unsure how to install R.

2.3. Install SimHap on your computer

To install **SimHap** on your computer, please download one of the following **SimHap** R package files (whichever is appropriate for your operating system):

SimHap_1.0.0.tar.gz	- SimHap R package for Linux and Mac users
SimHap_1.0.0.zip	- SimHap R package for Windows users

and the following java installer:

simhap1.0.0-install.jar	- Java-based installer for SimHap v1.0.0 for all users
-------------------------	---

from the **SimHap** download link at <http://www.genepi.org.au/simhap/>.

If you are using Windows, install the **SimHap** R library by launching R and choosing the 'Install package(s) from local zip files...' option under the 'Packages' tab. If you are using Linux or Mac, install the SimHap library by running the command:

```
R CMD INSTALL SimHap_1.0.0.tar.gz
```

in the directory where you saved the package. Note, you may need to be administrator/root to do this.

You can now run the 'simhap1.0.0-install.jar' installer by either double-clicking on the download, or typing the following in the command prompt/shell where you saved the installer:

```
java -jar simhap1.0.0-install.jar
```

You will be presented with a typical "Windows" style installer that will ask you where to install **SimHap** and will set up shortcuts for you.

Within the directory where you installed **SimHap**, you will find a file called 'simhap1.0.0-examples.zip'. Unzip (eg using WinZip) the example data file to a directory where you store your files – eg. the same directory where you installed **SimHap**.

3. Input Files and Data Preparation

SimHap requires that your genotypic and phenotypic data are in separate, comma separated text files (files with a .csv extension). If your data is in an excel spreadsheet, you can save this file as a comma separated file using 'csv' as a file type when saving in excel. The first column of both genotype and phenotype data files should contain an identification code for the individuals within the data set. These identifiers can be any character string. The first row of each file should contain column or field names.

3.1. Formatting SNP Data

The biallelic SNP genotypes in the genotype data file can be alphabetical or numeric eg. Aa, GG, 12. You may have a separator character between the two alleles of a SNP genotype eg. A/a or G_G but you do not need one. **SimHap** will read either format. Each allele of the genotype is not restricted to one character in length. For example, you may have a genotype v1v2, where v1 is the first allele, and v2 is the second. The only limitation on the naming of alleles is that the names of the two alleles of a genotype must be the same length. In the given example, the first allele v1 is two characters in length, and so is the second allele v2.

3.2. Missing Data

The missing data symbol is defined by the user and can be denoted by any character or character string. **SimHap** automatically detects a white space or an empty cell as missing data, however you can use any character or character string (e.g. *, ?, NA). We encourage that you use a blank cell to denote missing data, however if your missing data is denoted by something other than a white space or an empty cell, we encourage that you use

something unambiguous that will not be confused with variable values, such as an asterisk. The missing data character or character string must be the same in both your genotype and phenotype files.

4. Example Data Sets

Three example data sets have been provided to help you learn to use **SimHap**: three genotype files together with their associated phenotype files. These files exist in the 'simhap1.0.0-examples' directory referred to in Section 2.3. You can view these files with any text editor.

4.1. Data Set 1: Binary and Quantitative Outcomes

pheno.csv

This phenotype data file contains 16 columns of data on 180 individuals. The first column contains a unique ID for each individual and the remaining columns contain biological variables important to cardiovascular disease. Below is a description of the variables within the phenotype file. Note: the missing data character for this data set is a white space.

<u>Variable</u>	<u>Description</u>
ID	patient identifiers.
SEX	1=male, 0=female.
AGE	age in years.
SBP	systolic blood pressure (mmHg).
DBP	diastolic blood pressure (mmHg).
BMI	body-mass index.
WHR	waist-hip ratio.
HDL	plasma high density lipoprotein (mmol/L).
LDL	plasma low density lipoprotein (mmol/L).
DIABETES	a binary indicator of history of type 2 diabetes.
FH_IHD	a binary indicator of family history of ischaemic heart disease.
PLAQUE	a binary indicator of the presence of 1 or more carotid plaques.
SMOKE	a binary indicator of smoking history (0=never smoked, 1=ever smoked).
PY	pack-years of smoking.
DISEASE	a binary indicator of ischaemic heart disease.
STRAT	a matching variable indicating the pairs of matched cases and controls.

geno.csv

This genotype data file contains genotype data for 4 biallelic SNPs in a particular gene. Note that there is no separator character between the alleles for these SNPs and that the missing data character for this data set is a white space.

These files can be used to practice analysis with quantitative (e.g. HDL) or binary (e.g. PLAQUE) outcomes.

4.2. Data Set 2: Longitudinal Outcomes

longpheno.csv

This phenotype data file contains 12 columns of longitudinal data related to asthma for 99 individuals. Multiple observations on one individual (e.g. multiple hospital visits) are represented by multiple rows in the data file, with the same value in the ID column. Below is a description of the variables in the phenotype file. Note: the missing data character for this data set is a white space.

<u>Variable</u>	<u>Description</u>
id	Individual identification number
year	Year in which individual visited the hospital clinic
age_time1	Age at first clinic visit (years)
age_time1c	Age at first clinic visit centred around zero (years) (i.e. $\text{age_time1} - \text{mean age_time1}$)
sex	0 = male, 1 = female
age	Age at time of measurement (years)
agec	Age at time of measurement (years) centred around zero (i.e. $\text{age} - \text{mean age}$)
height	Height (metres)
weight	Weight (kilograms)
bmi	Body-Mass Index
fev1f	Forced expiratory volume in 1 second (mLs): measure of lung function

longgeno.csv

This genotype data file contains genotype data on three biallelic SNPs from a particular gene. The genotype file should consist of only one row per individual, containing their genotypic information. Note that there is no separator character between the alleles for these SNPs and that the missing data character for this data set is a white space.

These files can be used to practice analysis with longitudinal quantitative outcome types. The most sensible outcome here is `fev1f`.

4.3. Data Set 3: Right-censored Outcomes

survpheno.csv

This phenotype data file contains data on the recurrence times to infection, at the point of insertion of the catheter, for 38 kidney patients using portable dialysis equipment. Catheters may be removed for reasons other than infection, in which case the observation is censored. Each patient has exactly 2 observations. Below is a description of the variables in the phenotype file. Note: missing data in this file is characterised by an empty cell.

<u>Variable</u>	<u>Description</u>
id	Individual identification number
time	Time to infection (days)
status	Indicator of censoring (0=not censored, 1=censored)
age	Age at observation time (years)
sex	1 = male, 2 = female
disease	Categorical indicator of disease type

survgeno.csv

This genotype data file contains genotype data for kidney patients on three SNPs in a particular gene. The genotype file should consist of only one row per individual, containing their genotypic information. Note that there is no separator character between the alleles for these SNPs and that the missing data character for this data set is an empty cell.

These files can be used to practice analysis with right-censored outcome types.

5. Running SimHap

To begin **SimHap**:

If you are using Microsoft Windows, once you have installed **SimHap** you can run the program by clicking on the shortcut created - either on the desktop or in your Start Menu. Alternatively, and for Linux users, to run **SimHap** please type the following into the command prompt (shell) - from the directory where you installed **SimHap**.

`simhap.bat` - for Windows users

`sh simhap.sh` or `./simhap` - for Linux and Mac users

You may be prompted to choose where you have installed R. If **SimHap** does not automatically detect where R is installed, you will be required to navigate to the correct location in the window that appears. If you are using Windows, browse to the location of your R install and select the executable file called `R.exe`. If you are using Linux or Mac, browse to the location of your R install and select the executable file called `R`. Assuming everything is installed correctly, the initial **SimHap** window (as seen in Figure 1.) should appear.



Figure 1. SimHap start screen

Click **START SIMHAP** to load your genotype and phenotype data ready for analysis. The screen shown in Figure 2. will appear and you will be able to load your data.

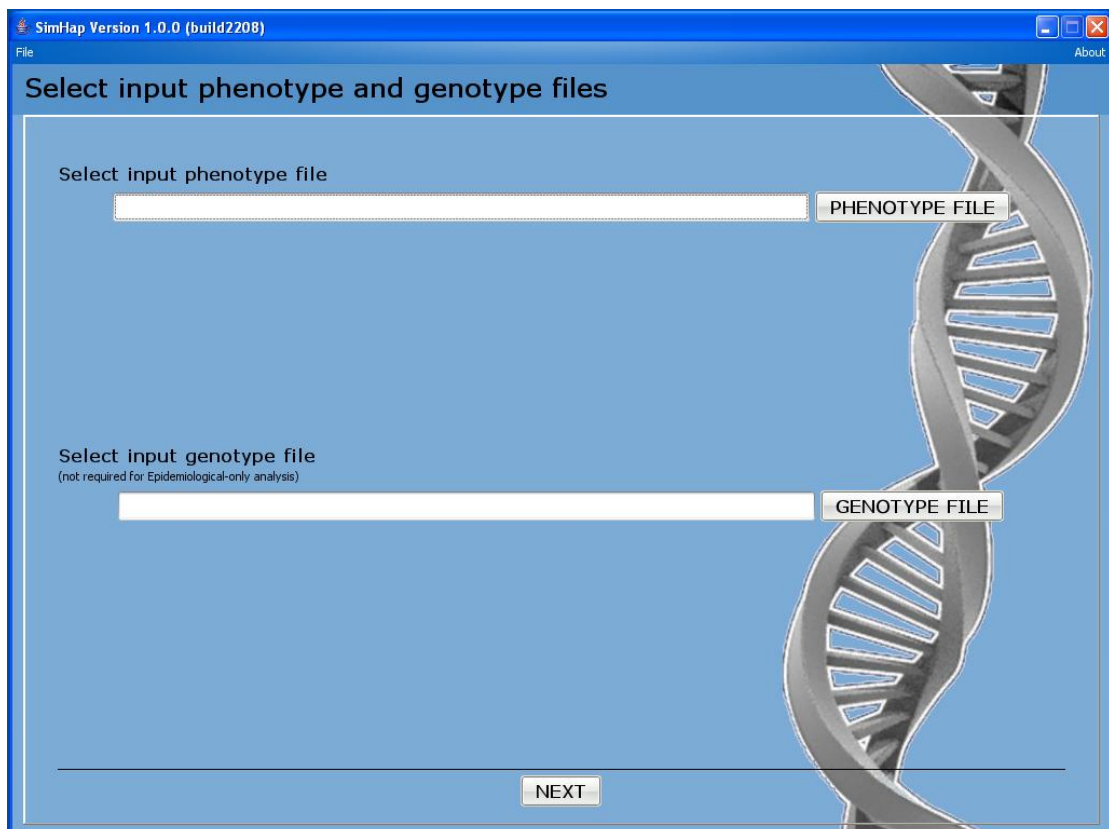
The image shows the SimHap data input screen. The window title bar reads "SimHap Version 1.0.0 (build2208)". The main heading is "Select input phenotype and genotype files". Below this heading, there are two sections. The first section is labeled "Select input phenotype file" and contains a text input field followed by a button labeled "PHENOTYPE FILE". The second section is labeled "Select input genotype file" with a sub-note "(not required for Epidemiological-only analysis)" and contains a text input field followed by a button labeled "GENOTYPE FILE". At the bottom center, there is a button labeled "NEXT". A 3D DNA double helix illustration is visible on the right side of the screen.

Figure 2. Data input

Click on the **PHENOTYPE FILE** button to search for your stored phenotype file, otherwise type in the path for the location of this file in the space provided. If you intend to perform epidemiological analysis only (without the inclusion of genetic data), a genotype file is not required at this stage. If you intend to perform haplotype or single SNP analysis, you must load both a phenotype and a genotype file in the format described in Section 3 of this manual. Once you have loaded your data files, you will be asked to specify a missing data and allele separator characters, if applicable. If missing data is characterised by a blank cell or a space in your data file, simply leave this field blank. Note that missing data should be characterised in the same way for both genotype and phenotype files. Once you have set the missing data and allele separator fields, you will be presented with a screen asking you whether you would like to perform **HAPLOTYPES**, **SINGLE SNP** or **EPIDEMIOLOGICAL** analysis.

5.1. Haplotype Analysis

5.1.1. Generating haplotypes

If you choose to perform haplotype analysis, you will be presented with a screen (Figure 3.) showing the available SNPs. To select a SNP to be included in the haplotyping process, simply click to highlight it and click **>>** to move it to the **Selected SNPs** box. To add multiple SNPs simultaneously, hold down Ctrl while highlighting them. Note the order that you select the SNPs is the order that they will be used to generate haplotypes. At this stage, if you are performing a case-control study, you can select to generate haplotype frequencies independently in cases and controls. If you select this option, you be asked to specify which variable in your phenotypic data indicates case status. **SimHap** will now generate estimates of the haplotype frequencies and print them to the screen. This may take up to several minutes depending on the size of your data, the number of SNPs and the speed of your computer. The EM-algorithm implemented in **SimHap** works most efficiently with up to 5 SNP haplotypes. The more SNPs you add, the longer the haplotype generation step will take, and we do not recommend that you attempt to generate more than 8 SNP haplotypes. Doing so can result in the algorithm failing to reach convergence.

If you have chosen to infer haplotypes separately in cases and controls, you will be presented with two sets of frequencies, otherwise you will be presented with the combined estimated haplotype frequencies (Figure 4.).

Because we often have little power to detect an association between a trait and rare haplotypes (say less than 1%), such haplotypes are often grouped together. **SimHap** allows you to choose a threshold frequency, and haplotypes with a frequency below this threshold will be grouped together. The default threshold is 5%, however this can be changed either typing a new value into the frequency box, or by using the arrows. When including individual haplotypes in a model, haplotypes with a frequency below this threshold can be included in your analyses as a group, or can be left out altogether. More information on including covariates in a model is described later in this section.

SimHap Version 1.0.0 (build2208)

Select SNP order to create haplotypes

Available SNPs

S_snp_1
S_snp_2
S_snp_3
S_snp_4

>>

<<

Selected SNPs (min 2)

If this is a case-control dataset, would you like to infer haplotypes in cases and controls separately?

Which variable indicates case status?

Yes ▾

id ▾

PREVIOUS NEXT

Figure 3. Haplotype SNP selection

SimHap Version 1.0.0 (build2208)

Set minimum haplotype frequency

Initial Haplotypes and Frequencies

	Haplotype	Frequency	Std.Error
1	h.M1G	0.4840	0.0263
2	h.N1A	0.1525	0.0189
3	h.N2A	0.1104	0.0165
4	h.M2G	0.1066	0.0163
5	h.N1G	0.0766	0.0140
6	h.N2G	0.0606	0.0126
7	h.M2A	0.0058	0.0040
8	h.M1A	0.0036	0.0032

SNP order used to create haplotypes:

S_snp_1
S_snp_2
S_snp_3

Please select minimum haplotype frequency 5 ▾

PREVIOUS VIEW INDIVIDUALS SAVE TO FILE NEXT

Figure 4. Haplotype frequencies

5.1.2. Defining the model

Once the frequency threshold has been set, you will be asked what type of outcome you wish to analyse: quantitative, binary, longitudinal or right-censored, and whether you would like to model all haplotypes together, or select individual haplotypes for inclusion into the model. On this same screen you may choose to examine the distribution of your variables by clicking on the **NORM. PLOTS** button. Selecting this option will allow you to view histograms of your variables and to perform a Shapiro Wilks test of normality. You can perform natural log or \log_{10} transformations and re-run these tests (Figure 5.).

Choosing to model all haplotypes together will cause all haplotypes with a frequency greater than the specified threshold will appear in your model and their effects will be calculated relative to a baseline haplotype. Choosing to include haplotypes individually will allow the user to include only those haplotypes of interest. The effect of each individual haplotype will be calculated relative to not having that haplotype. If you chose to include all haplotypes, you will be presented with a screen similar to the Model selection screen shown in Figure 6.

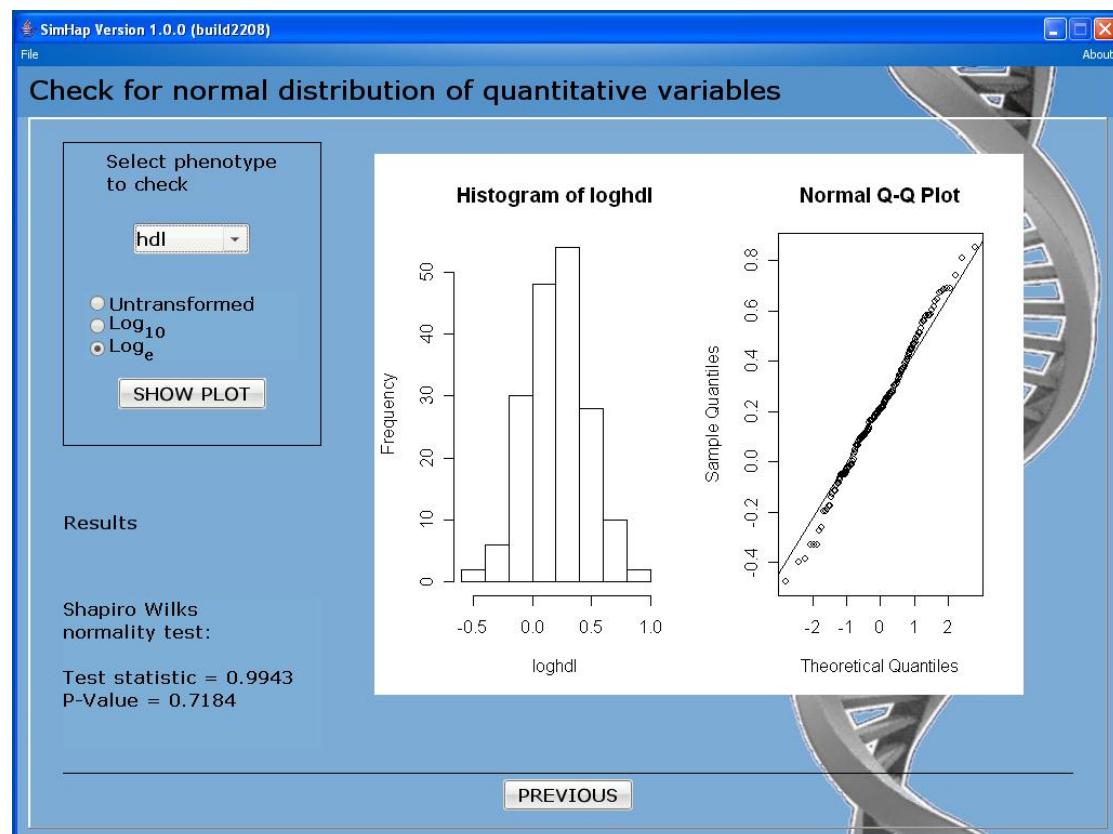


Figure 5. Check for normality

The effects of haplotypes on the outcome of interest will be relative to the effect of this haplotype. **SimHap** will default to the most common haplotype as a baseline, but you can change this if you wish. If you have chosen to include haplotypes individually, each haplotype will appear in the **Available Covariates** menu. They will appear with "h." in front of the haplotype name (eg h.n2a). A baseline haplotype is not required if haplotypes are included individually as

the reference category in this case is simply not having that haplotypeⁱⁱ. If you do not wish to model the effects of the rare haplotypes, simply do not move the variable called “h.rare” over to the **Selected Covariates** box on the right of the screen.

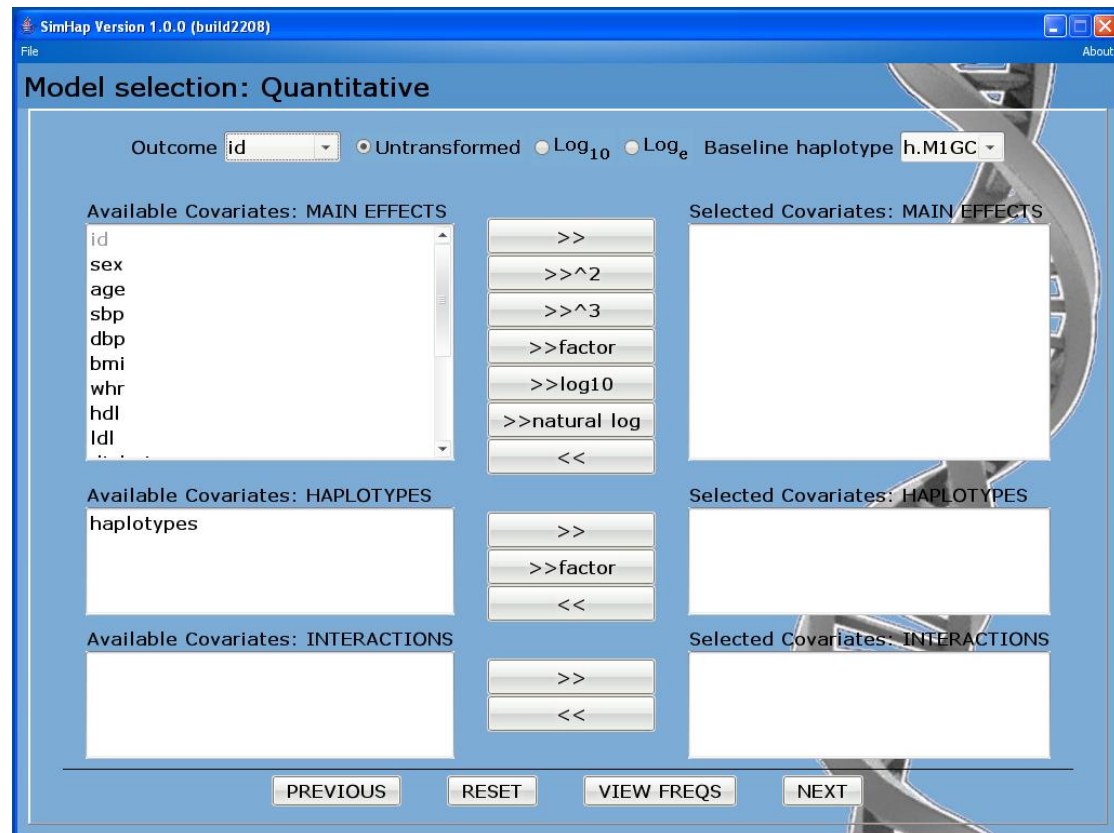


Figure 6. Model selection

Select the outcome of interest using the drop box at the top of the screen. You can choose to leave this variable untransformed or apply a natural log (\log_e) or \log_{10} transformation. To adjust for covariates in your model, highlight the covariate on the left of the screen and move it across to the **Selected Covariates** box on the right using **>>**. Variables can be added to the model as factors by using the **>>factor** buttonⁱⁱⁱ. Note that you can add quadratic and cubic terms by using the **>>^2** and **>>^3** buttons respectively. Covariates can also be transformed using natural log or \log_{10} before inclusion in the model using the **>>natural log** and **>>log10** buttons respectively. You can also include interaction terms to your model. To fit a two-way interaction, use Ctrl to select the two variables and then move them across to the **Selected Interactions** box with **>>** to

ⁱⁱ If you choose to model all haplotypes together in the model, haplotypes are included as linear terms. If you choose to include haplotypes individually, you can include them as a factor, in which case a coefficient (or odds ratio) and p-value will be derived for each level of the factor. In the case of an additive effect, the levels of a haplotype are 0, 1 and 2 representing the number of copies of that haplotype. For a dominant effect, the levels of a haplotype are 0 and 1, where 0 represent no copies of the haplotype and 1 represents at least one copy. In the case of a recessive effect, the levels of a haplotype are 0 and 1, where 0 represents less than two copies of the haplotype and 2 represents exactly two copies of the haplotype.

ⁱⁱⁱ A covariate modelled as a factor will be treated as a categorical variable. If the covariate is numeric, the baseline will be the smallest value by which the effect of each other level of the factor is compared. If the covariate is non-numeric, the baseline will be the first category when ordered alphabetically.

add them to the model. More complex interactions (e.g. 3-way) can be added in a similar manner. Note that SimHap requires that the main effects of variables included in an interaction term to be entered into the model, therefore the possible variables for use in interactions will be limited to those you have already chosen as selected **MAIN EFFECTS** and **HAPLOTYPES**.

This screen will be the same for quantitative, binary and longitudinal outcomes. The only change for right-censored outcomes is that instead of selecting an outcome variable, you select a censoring variable (i.e. an indicator of censoring).

5.1.3. Specifying model parameters

Once you have defined your model, a new screen will appear where you can specify some model parameters. You can model the haplotype effects as additive, dominant or recessive. Select the type of effect you wish to model using the drop box. If you are modelling a right-censored outcome, you must also specify which variable in your phenotype file represents time. If you are using longitudinal data, you must also select a grouping variable (such as a subject ID) and a time variable (which indicates a change in time between observations). You can also specify a correlation structure and a value for your data set on this screen. Three correlation structures are possible:

corAR1	An autocorrelation structure of order 1. Value is the value of the lag 1 autocorrelation which must be between -1 and 1. Defaults to 0.2.
corCAR1	An autocorrelation structure of order 1, with a continuous time covariate. Value is the correlation between two observations one unit of time apart which must be between 0 and 1. Defaults to 0.2.
corCompSymm	A compound symmetry structure corresponding to uniform correlation. Value is the correlation between any two correlated observations. Defaults to 0.2.

SimHap will choose **corCAR1** with a **value** of 0.2 if a correlation structure is not defined by the user.

On this screen you can also specify the number of simulations (imputed data sets) you wish to use in the modelling process, and (optionally) choose one or two subsets of your data to use in your analysis. For example, to perform an analysis in males over 50 (with **SEX** coded as 0 for males and 1 for females) you can choose "**SEX = 0**" as the first subset and "**AGE > 50**" as the second subset. The following symbols can be used in the sub-setting option:

=	equal to
!=	not equal to
>	greater than
>=	greater than or equal to
<	less than
<=	less than or equal to

While the modelling procedure takes place, a progress bar will appear telling you approximately how long the simulations will take to run. For large data sets and many simulations, this can take some time.

5.1.4. SimHap output

Once the modelling procedure is complete, results (along with a description of your model) will be displayed to the screen. Because **SimHap** uses simulation, parameter estimates are generated for each simulation you run. The summary results that are given are the mean value of each parameter estimate taken over all simulations and are adjusted for all other parameters in the model. Note: p-values less than 0.05 appear in red. To view a more detailed set of results, click the **VIEW DETAILS** button. You can then save these results to a text file by clicking **SAVE RESULTS**. Under the **VIEW DETAILS** option, you can also print the results directly from the screen.

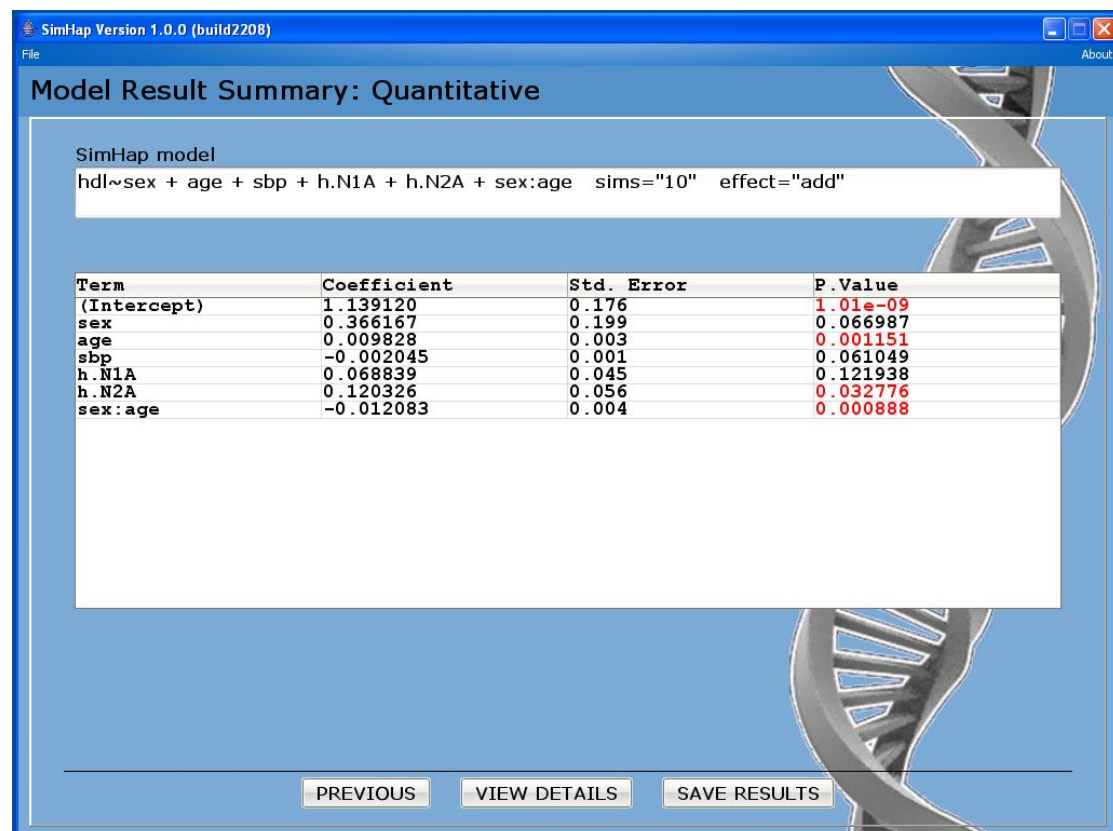


Figure 7. Results

5.2. Single SNP Analysis

As well as performing haplotype association analysis, **SimHap** can also be used to model the effects of single SNPs on an outcome of interest. After you have loaded your genotype and phenotype files, select **SINGLE SNPS** as your type of analysis when prompted. You will be prompted to specify a missing data character and an allele separator character as with haplotypes. Once you have completed this, a screen like the one shown in Figure 8. will appear. You are required to specify which allele for each SNP is the major allele. This determines which genotype will be used as a baseline, by which to compare the effects of the other genotypes. Selecting the major allele will define the wildtype as the genotype composed of two copies of this allele. This wildtype will then be used as a baseline, by which the effects of the other genotypes will be compared. Allele frequencies displayed on this screen are calculated from your genotypic data.

SimHap Version 1.0.0 (build2208)

File About

Set major (wildtype) allele for each SNP

SNP	Major Allele	Allele 1 Frequency	Allele 2 Frequency
S_snp_1	M	M=216(60.3%)	N=142(39.7%)
S_snp_2	1	1=258(72.1%)	2=100(27.9%)
S_snp_3	G	G=262(72.8%)	A=98(27.2%)
S_snp_4	C	C=197(55.0%)	A=161(45.0%)

PREVIOUS NEXT

Figure 8. SNP analysis

You will then be required to specify your model in a similar manner to that required for haplotype analysis. Variables (including SNPs) can be included as factors in order to derive a coefficient (or odds ratio) and p-value for each genotype. The SNP effects can also be fitted as additive, dominant or recessive^{iv}. A SNP will appear on the list of covariates as “S_” followed by the

^{iv} When SNPs are included as factors, if you choose to model the SNPs as additive, the wildtype genotype (which you defined by selecting the “major” allele) will be treated as a baseline and be coded by a 0, the heterozygote will be coded as 1 and the other homozygote will be coded as 2. If a SNP effect is modelled as dominant, the wildtype will remain 0 and the other two genotypes will be coded as 1 (indicating that an effect can be seen with 1 or 2 copies of the rare allele). If a SNP effect is modelled as recessive, both the wildtype and the heterozygote will be coded as 0 and the rare homozygote will be coded as 1 (indicating that 2 copies of the rare allele are required to see an effect). SNPs must be

SNP name, followed by an indicator of additive, dominant or recessive; e.g. **S_SNP1_add**. Move the SNP of interest over to the **Selected Covariates** menu using the >> button.

Data can be subset in the same way as in haplotype analysis, and results are output to the screen in a similar manner. Because no simulation procedure is implemented in single SNP analysis, the results should be generated very quickly. Detailed results can be viewed by clicking on the **VIEW DETAILS** button and will be saved by utilising the **SAVE RESULTS** option. Under the **VIEW DETAILS** option, you can also print the results directly from the screen.

5.3. Epidemiological Analysis

SimHap allows the analysis of epidemiological only data, without the inclusion of genetic covariates. Epidemiological analysis can be performed by either loading only a phenotype file into **SimHap** or by loading both a phenotype and genotype file and clicking on the **EPIDEMIOLOGICAL** button when prompted for the type of analysis you wish to perform. You will be required to specify your model in a similar manner to that required for haplotype and single SNP analysis. Variables can be included as factors in order to derive a coefficient (or odds ratio) and p-value for each level of the factor and various transformations such as quadratic and cubic terms, as well as log transformations that are available in haplotype and single SNP analysis options are also available for epidemiological analysis.

6. SimHap Licence and Referencing the Software

SimHap is provided free for non-commercial use. You may copy, distribute, display, and perform your work for non-commercial purposes only, provided you acknowledge the authors of **SimHap**. Commercial use is prohibited without express consent of the copyright holders.

SimHap is provided "as is" without warranty of any kind, either expressed or implied. We accept no responsibility for damages, including loss of data, to you or third parties.

If you find the software useful, please refer to:

Carter KW, McCaskie PA, Palmer LJ (2008). SimHap GUI: An intuitive graphical user interface for genetic association analysis. *BMC Bioinformatics*. 2008 Dec 25;9(1):557

Author contact:

Pamela A. McCaskie

Centre for Genetic Epidemiology

Western Australian Institute for Medical Research

Ground Floor, B Block

Hospital Avenue, Nedlands Western Australia 6009

included as factors to derive coefficients (or odds ratios) and p-values for each genotype or combination of genotypes.

AUSTRALIA

Email: pmccask@cyllene.uwa.edu.au

Phone: +61-8-9346 1612

Fax: +61-8-9346 1818

SimHap is an ongoing project. We have tested the software as comprehensively as is possible with the data available to us. If you encounter any problems, questions or queries regarding **SimHap**, please direct them to Pamela McCaskie at pmccask@cyllene.uwa.edu.au.

References

1. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 1995; 12: 921-7.
2. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* 1996; 5(3): 299-314.