User Guide

Overview of SimHOEPI

SimHOEPI is a simulator which is user-friendly and fits various simulation purposes. It requires an existing real SNP dataset as a starting point and generates simulated data by resampling genotype fragments sequentially from different samples and concatenating them which can largely preserve the minor allele frequencies (MAFs) in the dataset. Then it calculates an epistasis model represented by a penetrance table and randomly assigns labels to samples based on this model. SimHOEPI allows the user to calculate penetrance tables for a specified epistasis model with the restriction of prevalence or heritability when baseline penetrance is user-specified. The advantage of SimHOEPI is that it not only supports the calculation of second-order epistasis model, but also supports the calculation of epistasis model with high-order epistasis model. In addition, a graphical user interface (GUI) is provided for the convenience of users to calculate new epistasis models.

Availability

SimHOEPI and the detailed manual are freely available online at https://github.com/CDMBlab/SimHOEPI.

Requirements

MATLAB (tested against version R2021a, it is likely to work on many others).

Parameter settings

In the component of setting parameters, several parameters including general parameters, epistasis model parameters, and output parameters need to be specified using the graphical user interface.

General parameters

- Case

  It is the case number (i.e., the number of samples affecting a complex disease) of the simulation data. The default value is 1000 (Integer).

- Control

  It is the control number (i.e., the number of samples affecting no complex disease) of the simulation data. The default value is 1000 (Integer).

- SNP

  It is the number of SNPs in the simulation data. The default value is 100 (Integer).

Epistasis model parameters

- MAF

  It is the MAFs of the SNPs in the model. The value should be in the range of [0, 0.5] (double).

- Heritability

  It is the heritability of the model (h^2). The value is less than 1 (double).

- Prevalence

  It is the population of cases ($P(D)$). The value is less than 1 (double).

- The baseline effect

  Baseline penetrance (α) is the part of the penetrance function that makes up the model with relative penetrance. The value is less than 1 and as small as possible, such as 0.05 (double).

Output parameters

- FileName of SNP data

  It is the file name of the SNP data. The default setting is SNP (string).

- SNP data (.txt)

  It shows whether the output file format of SNP data is txt. The default setting is 0 (Boolean).

- SNP data (.mat)

  It shows whether the output file format of SNP data is mat. The default setting is 0 (Boolean).

- FileName of the penetrance table

  It is the file name of the penetrance table. The default setting is table(string).

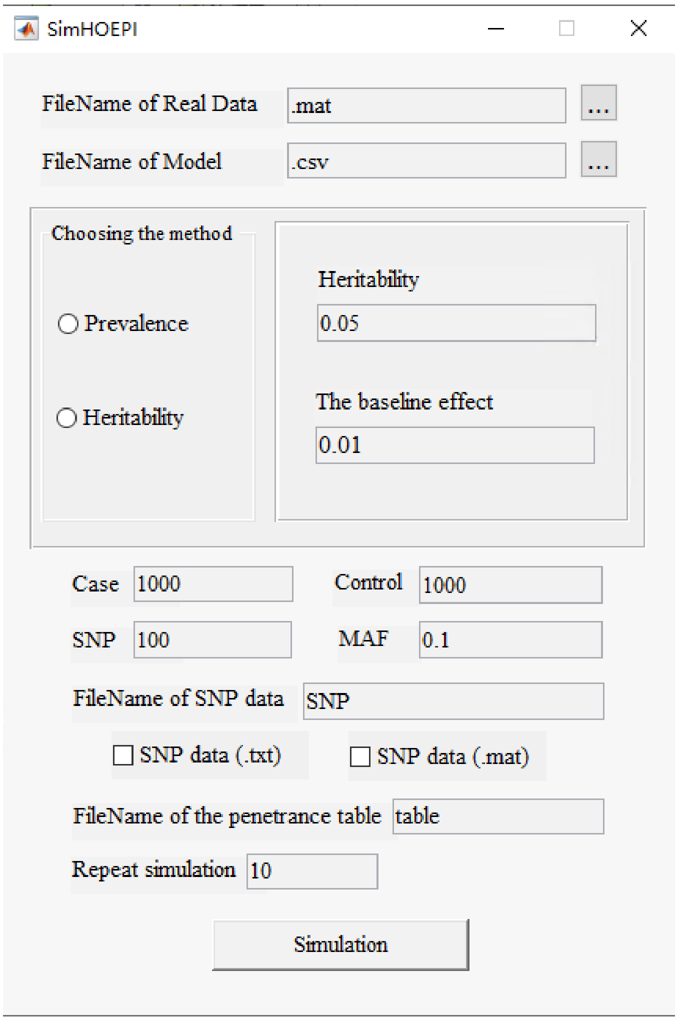
- Repeat simulation

  It is the specified number for batch processing. The default value is 10 (integer).

Usage

Starting the program

After downloading these files, there will be a file named SimHOEPI.m in the SimHOEPI folder. This file is the main function of the program and the user need to run this file in MATLAB to get the graphical user interface of this simulator. The original Interface of SimHOEPI is shown in Figure 1.



SimHOEPI

FileName of Real Data: .mat

FileName of Model: .csv

Choosing the method

○ Prevalence

○ Heritability

Heritability: 0.05

The baseline effect: 0.01

Case: 1000

Control: 1000

SNP: 100

MAF: 0.1

FileName of SNP data: SNP

□ SNP data (.txt)

□ SNP data (.mat)

FileName of the penetrance table: table

Repeat simulation: 10

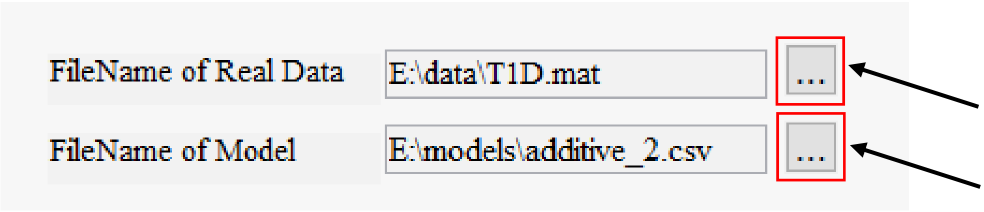
Figure 1 - The original Interface of SimHOEPI

Input

SimHPEPI requires two files, the real data file and the model file.

- The real data file in this program is from the database of Type 1 Diabetes (T1D). This data is represented as a matrix, with each row representing a sample and each column representing a SNP.
- The model file is written to a file using CSV (Comma-Separated Values) format, where each row is a different genotype, and two columns are the genotype and its associated penetrance expression function. There are two variables in the expression function named x and y which correspond to baseline penetrance α and relative penetrance f.

The user needs to click the two buttons in the red box in the Figure 2 respectively to input the real data file and the model file. The real data file and the model file are also provided in the corresponding folders. The users can choose the type and order of the model according to their needs. Figure 2 is an example of input real data and a second-order additive model.



FileName of Real Data: E:\data\T1D.mat

FileName of Model: E:\models\additive_2.csv

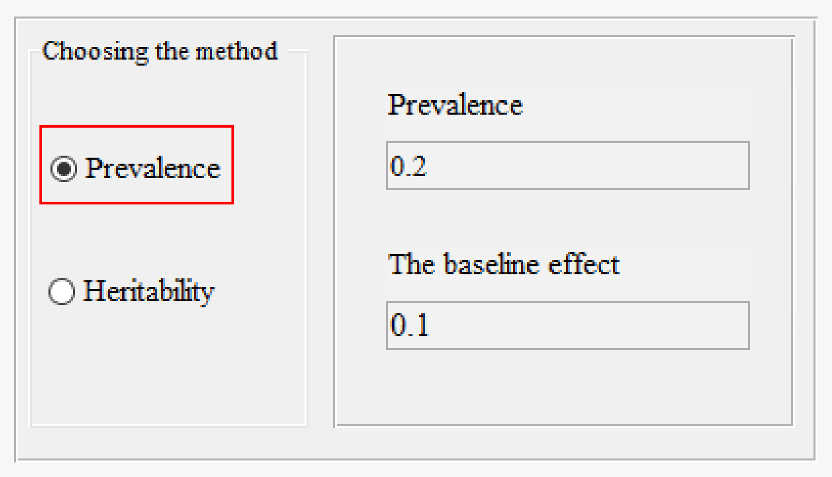
Figure 2 - Interface for inputs

Choosing the method

SimHOEPI provides two ways to calculate epistasis models: one is based on the heritability h^2 and the other is based on the prevalence $P(D)$. Users can choose one of the methods to calculate the SNP epistasis model according to their needs.

- An example of the one method based on the population prevalence $P(D)$ and the baseline effect α is shown in Figure 3. And the user should set parameters, including the baseline effect α, the population prevalence $P(D)$.

- An example of the other method based on heritability h^2 and the baseline effect α is shown in Figure 4. And the user should also set these two corresponding parameters.



Choosing the method

● Prevalence

○ Heritability

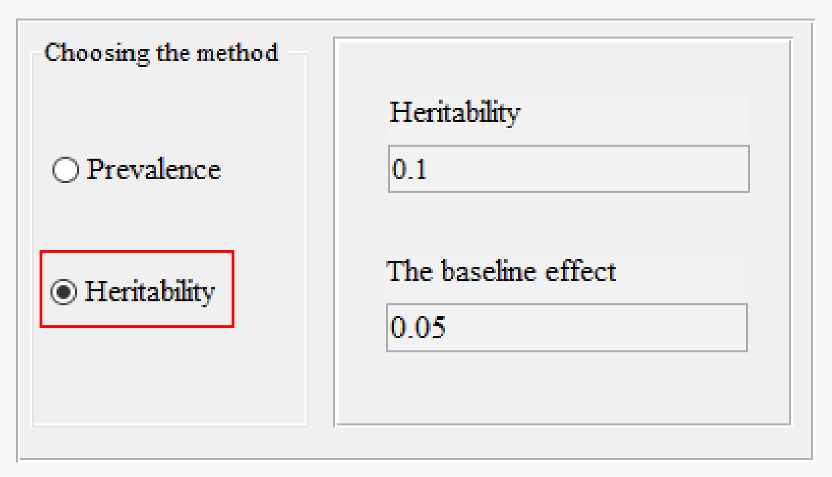
Prevalence

0.2

The baseline effect

0.1

Figure 3 - Interface for choosing the method of $P(D)$



Choosing the method

○ Prevalence

● Heritability

Heritability

0.1

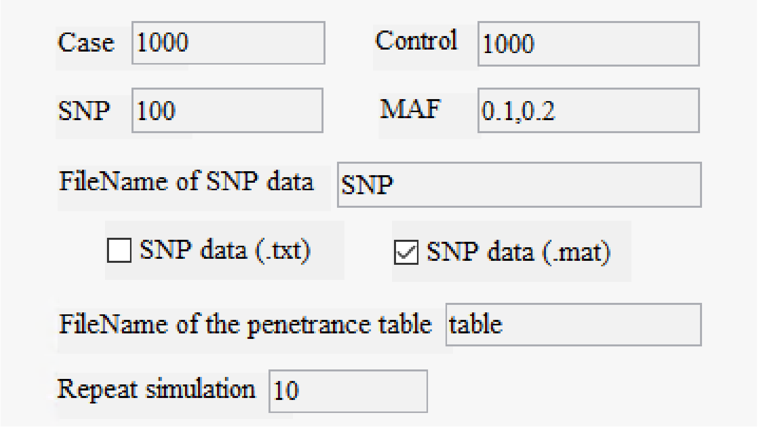
The baseline effect

0.05

Figure 4 - Interface for choosing the method of h^2

Simulation of SNP data

As shown in Figure 5, four general parameters, including Cases, Controls, SNPs, MAFs, should be set by users freely. The number of MAFs should be consistent with the order of the model in the model file, and the numbers of MAFs should be separated by ",". And five output parameters, such as FileName of SNP data, FileName of the penetrance table, SNP data (.txt), SNP data (.mat), and the number of Repeat simulation can also be freely set by the user. All the values in the calculated penetrance table are transformed by a function to ensure that the penetrance values are between 0 and 1.



Case 1000 Control 1000

SNP 100 MAF 0.1,0.2

FileName of SNP data SNP

□ SNP data (.txt) ☑ SNP data (.mat)

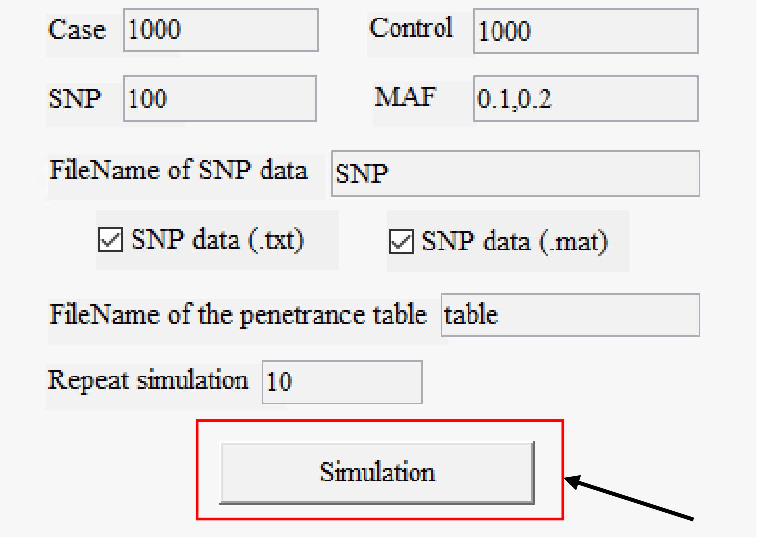
FileName of the penetrance table table

Repeat simulation 10

Figure 5 – An example of an interface for general parameters and output parameters

Click Simulation button

As shown in Figure 6, when the user clicks the button after setting all the parameters, the software can calculate the model and generate the simulation data according to the parameters input by the user.



Case 1000 Control 1000

SNP 100 MAF 0.1,0.2

FileName of SNP data SNP

☑ SNP data (.txt) ☑ SNP data (.mat)

FileName of the penetrance table table

Repeat simulation 10

Figure 6 – An example of clicking Simulation button