HWEBayesForensic User Manual

FUKUTA Mamiko*

Department of Forensic Medicine, Nagoya City University
Graduate School of Medical Sciences

January 17, 2023

*m.fukuta@med.nagoya-cu.ac.jp

Contents

1 Introduction

HWEBayesForensic is a script for the Bayesian approach evaluating the deviation from Hardy-Weinberg equilibrium (HWE) in forensic population data of multi-allelic makers. This script is written in R language and freely available in GitHub repository
(`https://github.com/MamikoFukuta/HWEBayesForensic`).
The evaluation can be carried out in two ways: parameter estimation and model comparison. In parameter estimation, the degree of deviation from the HWE is aggregated into a single parameter (f), and its value is estimated. The model comparison uses a Bayes factor to indicate whether the data fit better in a model with or without f.

2 Getting started

2.1 Preparing the operating environment

This script needs “R” and “RStan” software, which can be installed from the official websites.

- R website: `https://www.r-project.org/`
- RStan website: `https://mc-stan.org/users/interfaces/rstan`

The installation of RStan is a bit complicated, so please read the instructions on the official website carefully and check the operation using the example in ”RStan Getting Started” (`https://github.com/stan-dev/rstan/wiki/RStan-Getting-Started`) .

RStan also encourages using “RStudio” software, which is coding environment for R. This script, however, can also be run in R. RStudio can be installed from the official websites.

- RStudio: `https://posit.co/downloads/`

2.2 Downloading the script

Access the GitHub repository (`https://github.com/MamikoFukuta/HWEBayesForensic`), click the “Code” button, then “Download Zip”.

2.3 Input data format

The input data must be created in a CSV file in the following format.

Locus 1,	Locus 2,	$\cdots$	Locus m
S_1Al_1,	S_1Al_1,	...	S_1Al_1
S_1Al_2,	S_1Al_2,	...	S_1Al_2
S_2Al_1,	S_2Al_1,	...	S_2Al_1
S_2Al_2,	S_2Al_2,	...	S_2Al_2
$\vdots$	$\vdots$		$\vdots$
S_nAl_1,	S_nAl_1,	...	S_nAl_1
S_nAl_2,	S_nAl_2,	...	S_nAl_2

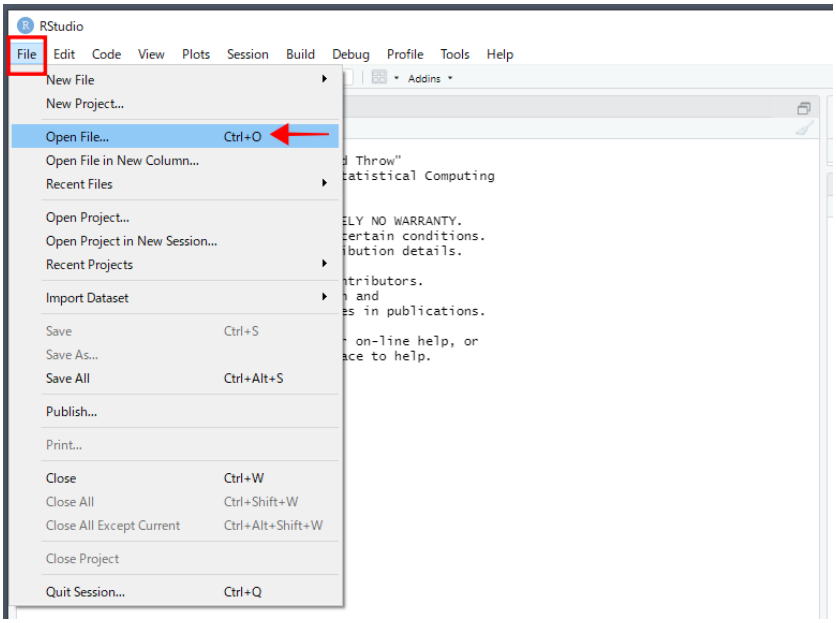
The first line contains the name of the loci, and the second and subsequent lines include the genotypic data, two lines per person. In the figure above, the data of m loci are shown for n samples (S_1, S_2, ..., S_n), where the genotype is Al_1/Al_2. If there are missing data, entering "NA" will result in the individual's genotype being ignored at that locus. The number of individuals needs to be the same for all loci in a single CSV file. The alleles can be numeric or alphabetic, and intermediate alleles such as "12.3" can also be entered. An example of an input file is shown below.

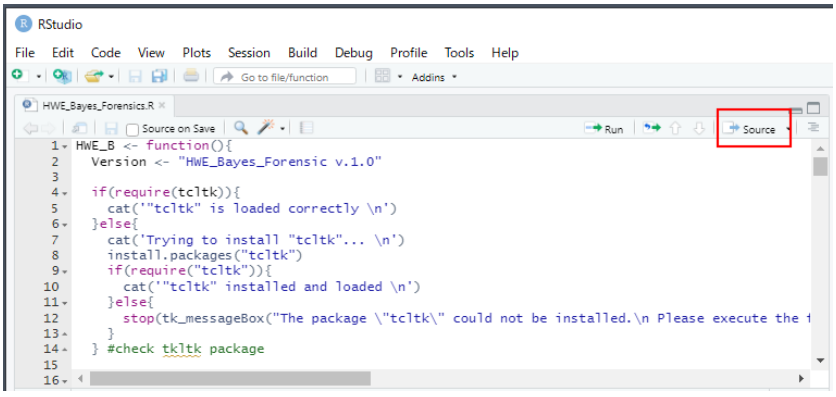
```
DXS8377,   DXS10147,   DXS7423
46,        7,          14
47,        8,          15
43,        6,          15
44,        8,          15
52,        8,          14
54,        8,          15
50,        6,          15
51,        8,          15
50,        6,          15
```

2.4 Runnig the script

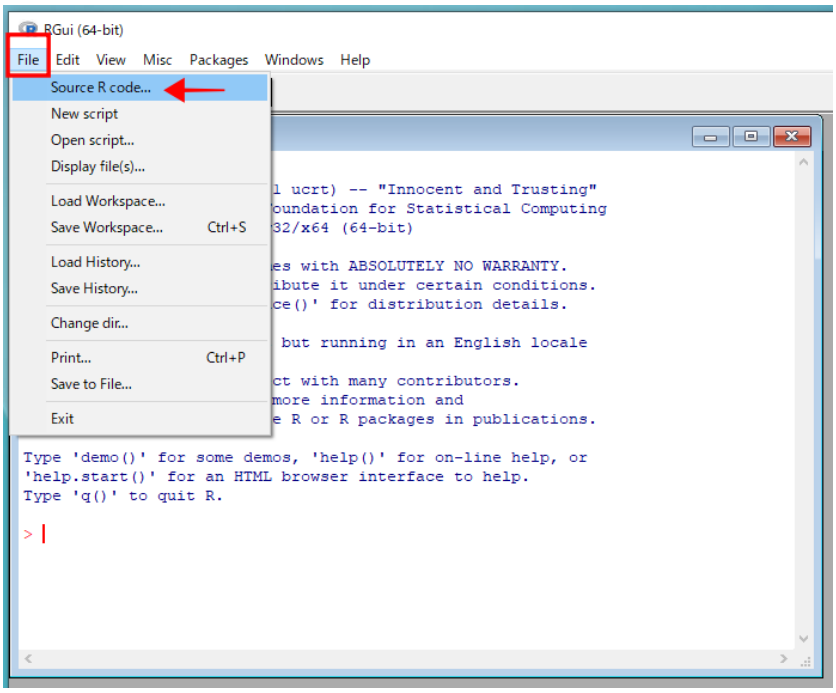
After downloading the script, open the RStudio (or R).

- In Rstudio, open the script then click on the source button shown in the following figure.





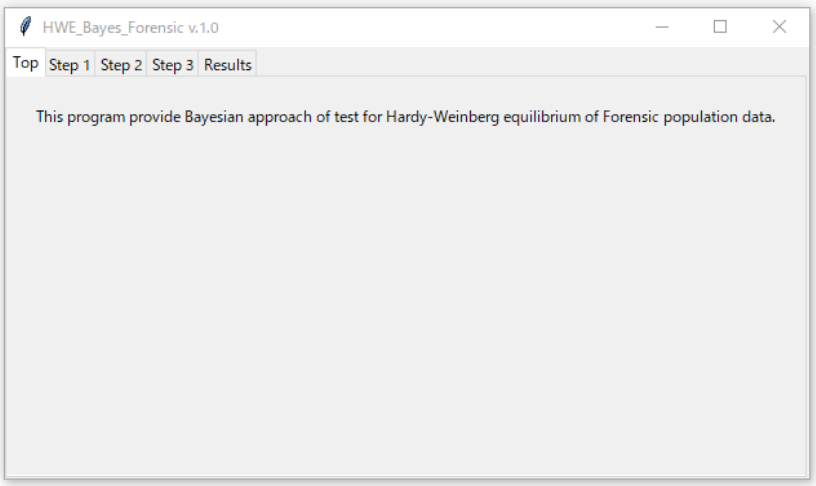
- In R, click “Source R code” in the “File” tab, then select the script.



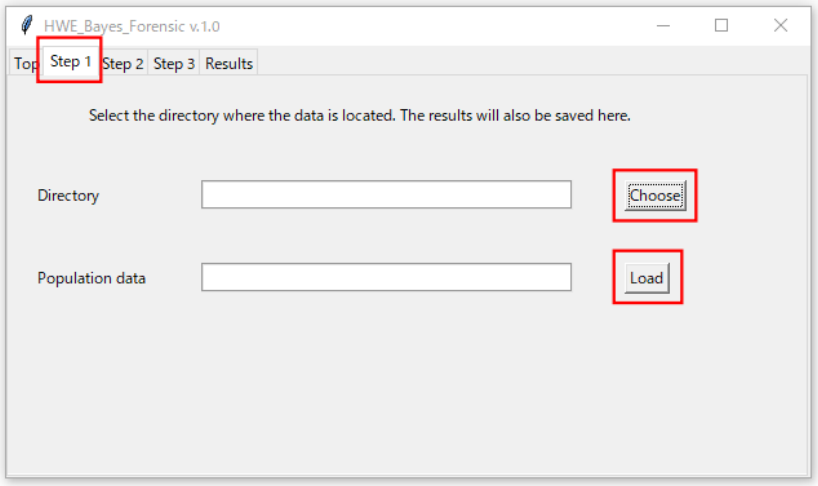
Then the necessary R packages ("tcltk", "tcltk2", "Bridgesampling", "ggplot2") will be automatically installed. However, as mentioned earlier, the "RStan" package needs to be installed beforehand to ensure a proper running environment.
It takes time to complete the complete the package installation and loading and compile the Stan model. You can check the progress on the console screen like in the below figure, and any errors will be displayed.

```
> HWE_B()
Loading required package: tcltk
"tcltk" is loaded correctly
Loading required package: tcltk2
"tcltk2" is loaded correctly
Loading required package: bridgesampling
"bridgesampling" is loaded correctly
Loading required package: rstan
Loading required package: StanHeaders
Loading required package: ggplot2
rstan (Version 2.21.2, GitRev: 2e1f913d3ca3)
For execution on a local, multicore CPU with excess RAM we recommend calling
options(mc.cores = parallel::detectCores()).
To avoid recompilation of unchanged Stan programs, we recommend calling
rstan_options(auto_write = TRUE)
Do not specify '-march=native' in 'LOCAL_CPPFLAGS' or a Makevars file
"rstan" is loaded correctly
Compiling F_stanmodel. Please wait for a few minutes.
Compiling nF_stanmodel. Please wait for a few minutes.
```

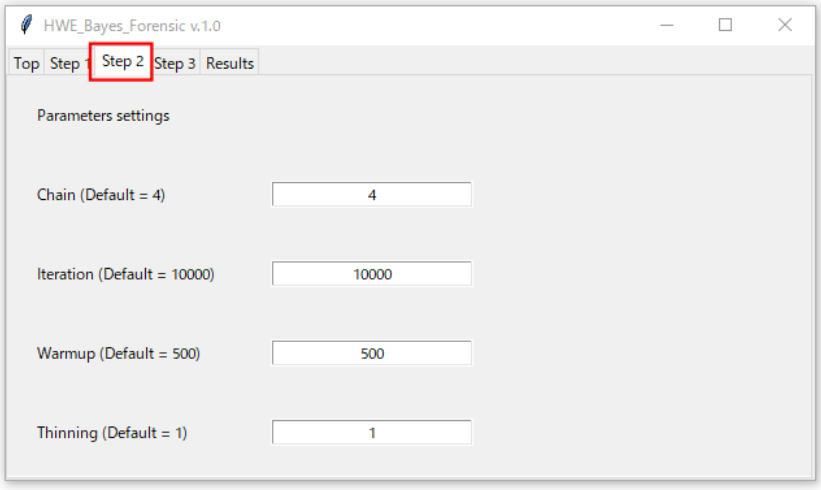
After all the preparation is done, the GUI window shown in the figure will open.



Click on the “Step 1” tab to select the directory where the data file is saved and the CSV file of the population data.



Next, click on the “Step 2” tab and set up the conditions for the sampling of Stan.



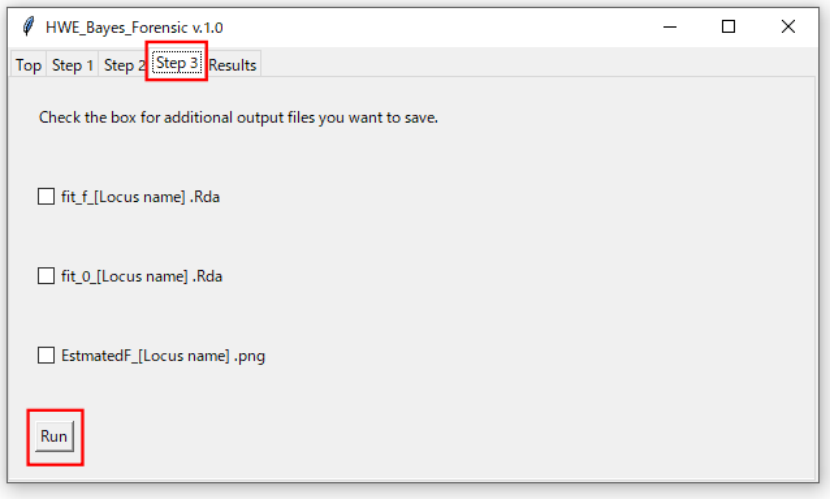
The number of samplings is described as the following:

$$Sampling\ number = \frac{Iteration - Warmup}{Thinning} \times Chain$$

Iteration is the total number of samples, and the first part indicated by Warmup is not used for estimation because it depends on the initial random

value. Thinning indicates how many results should be used every other time, and increasing the number reduces the effect of autocorrelation. This condition of sampling is repeated as the number of Chains, to make sure that the results agree even if the initial values change. If no convergence occurs, it is recommended to change the values of these parameters and try again. Note that reducing the number of sampling is not recommended for the correct estimation of the Bayes factor with bridge sampling.

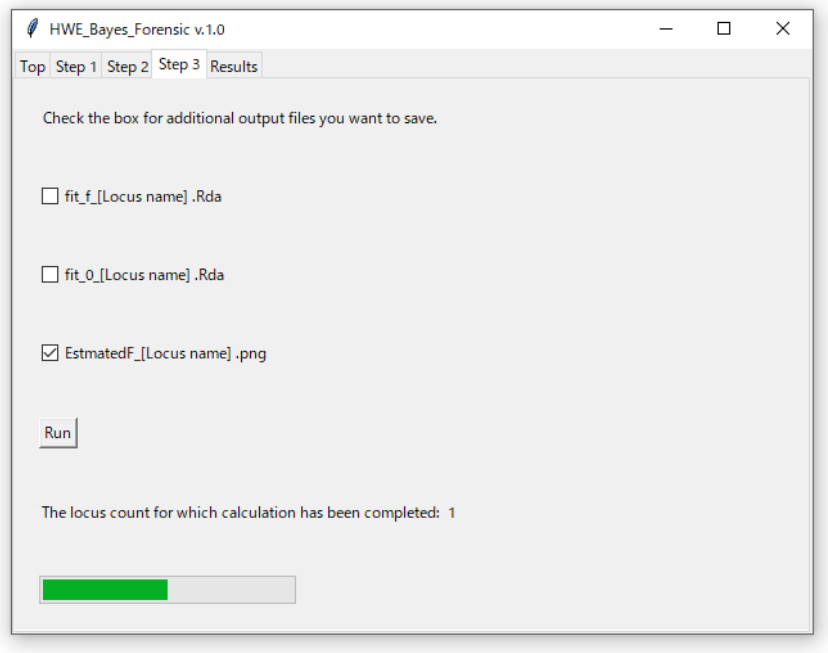
In the "Step 3" tab, select the file you want to save in the options.



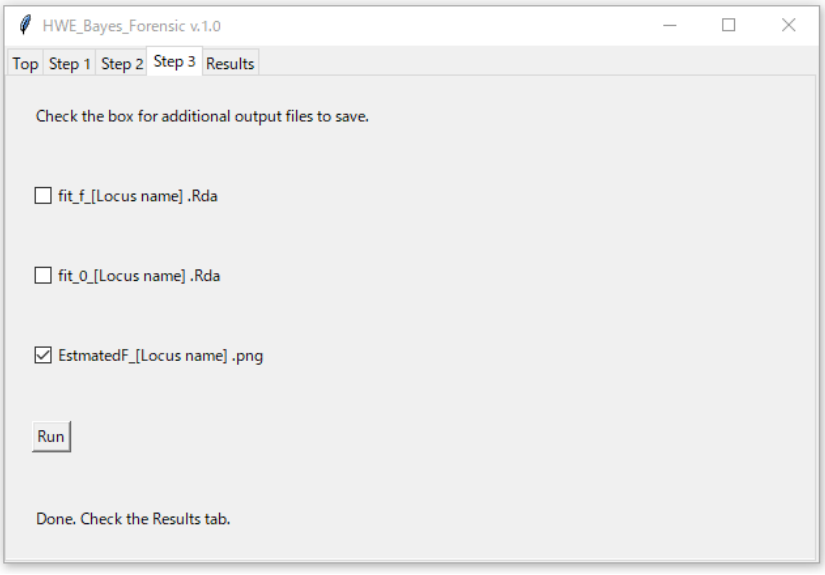
The fit file is very large because it contains all the sampling data. fit_f is the sampling result using the single-f-model (M_1), and fit_0 is one using the model assumed to be in HWE (M_0). These data applications will be explained in detail in later sections. It is recommended to save the posterior distribution image (PNG file) of f to check for convergence. The directory to save the file is the one you selected in Step 1.

After selecting the file to save, click the Run button to execute the sampling. Please make sure that sampling cannot be stopped, and take some time, depending on the performance of your PC.

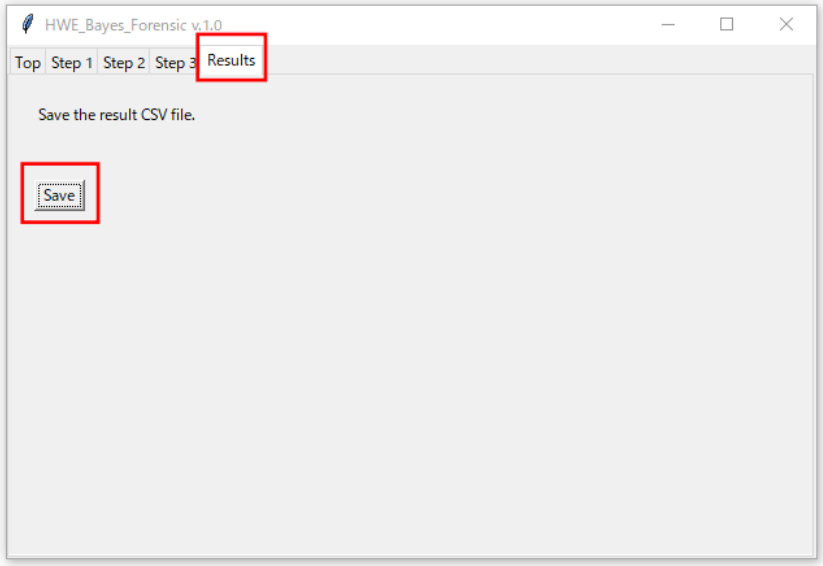
The progress is indicated by a bar.



When the sampling is finished, the progress bar will disappear.



Click on the “Results” tab to save the result CSV file.



If you have another file to analyze, click on the “Step 1” tab, change the file, and do the same.

2.5 Result file

The results file contains an estimated range of f, the judgement of $\hat{R}$, "Ratio", F_{IS}, and Bayes factors.

- The estimated range of f gives the value of minimum, 2.5 percentile, median, mode, 97.5 percentile, and maximum value.
- $\hat{R}$ is marked "TRUE" if it is less than 1.1 for all parameters, i.e., it is considered convergent.
- "Ratio" is the ratio of all heterozygotes to all homozygotes in a homogeneous allele frequencies population that is equally polymorphic with the locus. This value is 2 for HWE and < 2 if homozygous frequencies increase.
- F_{IS} is a value defined by Weir and calculated from the observed and expected heterozygosity.
- BF_{01} is a Bayes factor with M_0 as the numerator. If it is greater than 1, it means the data fit better to M_0 than M_1, while there is no consensus yet on how large to determine the data are in HWE.

  $$BF_{01} = \frac{p(data|M_0)}{p(data|M_1)}$$

  BF_{10} is the Bayes factor when the numerator and denominator are reversed to BF_{01}.

3 Advanced analysis

The MCMC sample optionally saved in Step 3 can be used to examine each parameter's distribution, correlation, and so forth. There are three parameters as follows:

- the deviation from HWE: f
- allele frequency: pa
- genotype frequency: pg

Please refer to the help page on Stan's functions for more information on utilizing fit data.