ISFG Educational Workshop, Copenhagen 2012

LRmix tutorial

Hinda Haned
Netherlands Forensic Institute, The Hague, The Netherlands

Contents

1 What is LRmix?

Forensim is an R-package dedicated to facilitate the statistical interpretation of forensic DNA evidence. It also provides simulation tools made to mimic data from casework. A detailed description of forensim is given in the package tutorial, available from: http://forensim.r-forge.r-project.org/. The present tutorial aims at describing one particular module of Forenim, LRmix, which facilitates the calculation of likelihood ratios of LTDNA samples with drop-out, drop-in, any number of contributors and replicates. It is programmed after the model proposed by Curran et al. (2005) and Gill et al. (2007). LRmix is programmed in the R language and offers a user-friendly graphical interface (based on Tcl/Tk) that facilitates the interaction with the program. In order to use LRmix, you first need to install the R software, and then the Forensim package. LRmix and Forensim are available for free, under the GNU General Public licence version ≥ 2. The following section details the installation process. The Hammer case, published in Gill et al. (2007) and available from Forensim website, is used for illustration purposes.

A note on notation A few typographical conventions are used in this tutorial: different colours are used for the R commands and for the R results. A verbatim font is used for `R commands`.

2 Installation

Before we start, make sure you have installed R properly. The R software is available from the Comprehensive R Archive Network (CRAN). Hereafter we explain how the software can be installed.

2.1 Install the R software

- Go to http://www.cran.r-project.org/
- Dependent on which operating system you use, click on one of the links:
  - Download R for Linux
  - Download R for MacOSX
  - Download R for Windows
- For Windows, simply follow the link Install R for the first time
- Click the link "Download R 2.15.0 for Windows", run the file and the installation program will start.
- Click on R-2.15.0.exe to install the set-up file
- After installation, a blue colored icon appears on your desktop, click on the icon to launch an R session (Figure 1).

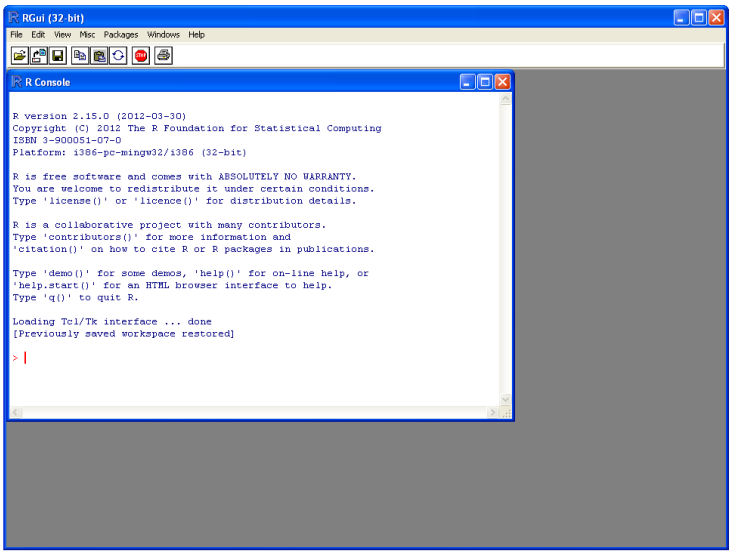

Figure 1: An R session (Windows)

A note for Mac users The LRmix module has a user graphical interface that relies on the Tcl/Tk language. The Tcl/Tk distribution is provided separately for Maxc OS system, and you will need to download it. Go to: http://cran.r-project.org/bin/macosx/tools/ and download the tcltk-8.5.5-x11.dmg file and install it on your system.

Once is downloaded on your system, you have to download Forensim and its dependencies.

2.2 Install the Forensim package

If your computer is connected to the Internet, follow Option 1, otherwise, follow OPtion 2.

2.2.1 Option1: install the packages directly from the R environment

Follow these steps:

1. Open R
2. type the following command in the R console:

   ```
   > install.packages('forensim')
   ```

   This automatically opens a list of 'CRAN mirrors' from which you can install Forensim. You can choose a mirror that it is close to you, thus if you are in France, you can choose Lyon, or if you are in the Netherlands, you can choose Amsterdam (Figure 2).

To make the Forensim package fully functional in R you need some additional packages. Repeat the previous step for all the following packages:

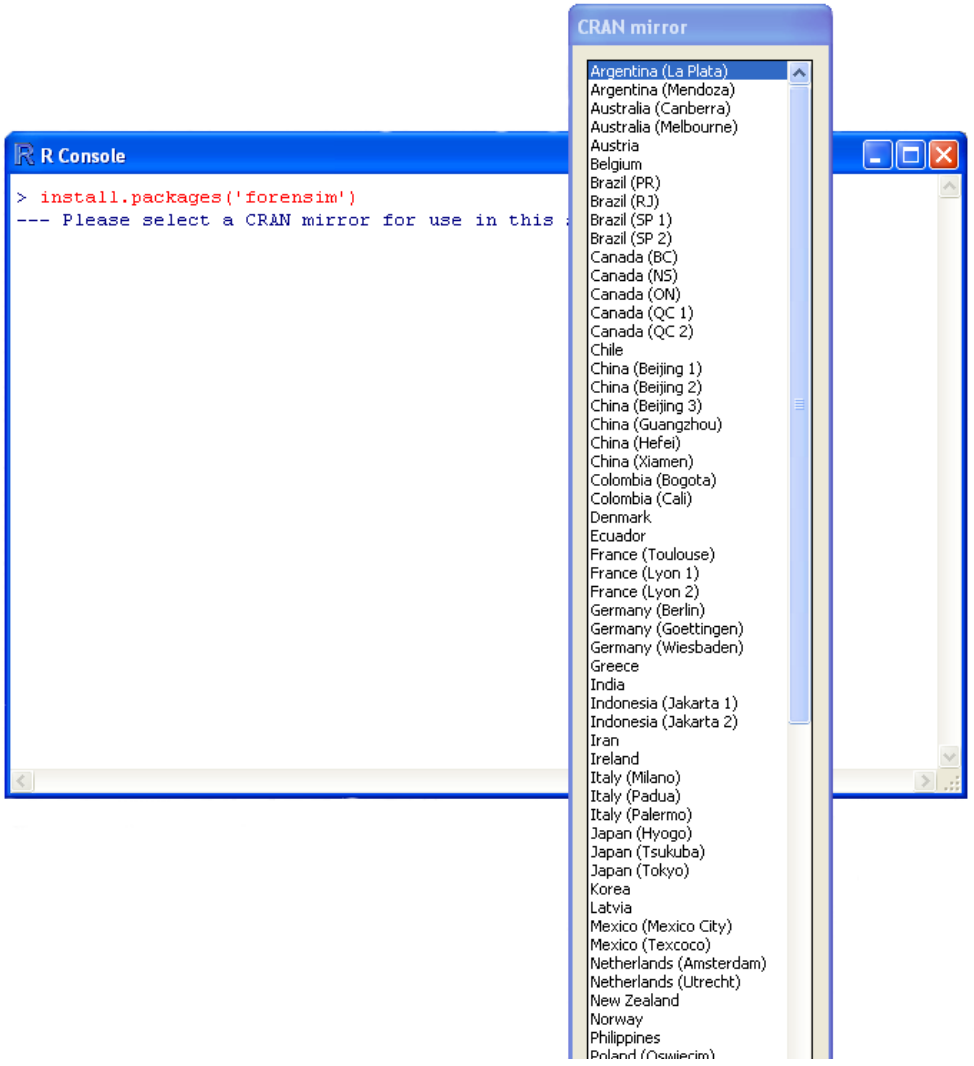

Figure 2: Install packages from the CRAN repository.

1. tcltk2
2. tkrplot.

The forensim package is now ready to be used!

2.2.2 Option2: manual installation

It is possible to install Forensim and its dependencies manually, this is useful if you do not have a connextion to Internet, so you can first load the relevant files and then install them into your computer. Forensim and its dependencies can be found on the CRAN website http://www.cran.r-project.org. In the left menu, under 'Software', click the link 'Packages', then click on 'Table of available packages, sorted by name'. Search for the Forensim package. Click the link with the appropriate file. If you use windows it is the one next to Windows binary, for the Forensim package, it is the forensim_3.1.zip file. Save the file into your working folder. **Do not unzip the file, as this is the required format for R packages**. To make the Forensim package fully functional in R you need some additional packages. Repeat the previous step for all the following packages:

1. tcltk2
2. tkrplot.

These two packages are needed to enable the correct display of the LRmix module.

All downloaded packages now need to be activated in R. Follow these steps:

- Open R
- Install packages using the R function `install.packages`:

```
> install.packages('forensim_3.1.zip', repos=NULL)
```

Do this for every downloaded package. Change the information within the quotation marks according to each package. The forensim package is now ready to be used!

Tip for windows users Download all the zip files in the same folder, then click on the Packages tab: install packages from zip files. It is possible to select all the packages at once, and install them at the same time (Figure 3).

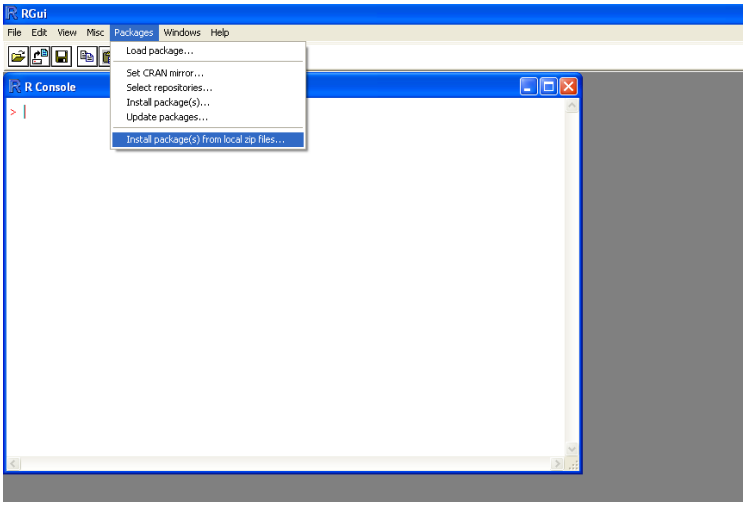

Figure 3: Package installation under Windows.

3 The LRmix module

Forensim implements a number of statistical methods that can be used in the statistical interpretation of evidentiary DNA samples. These methods are documented in the manual of the Forensim package as well as in Haned (2011).

The LRmix module implements a model for the qualitative evaluation of DNA samples. It is a direct implementation of the model described in Curran et al. (2005). The LRmix module allows the calculation of likelihood ratios for different replicates, with any number of contributors, and in case dropout and drop-ins occur. Population substructure is also accounted for using the classical θ correction (Balding and Nichols, 1994).

3.1 Getting started

The first step is to launch R. To do so, simply click on the blue R icon. This should open an R session as shown in Figure 1. The LRmix module is programmed into the R language, and its graphical user interface is programmed in Tcl/Tk.

Load the package forensim to your current R session using the function `library`:

```
> library(forensim)
```

Note! Every time R is closed and opened again a new session starts and the forensim package needs to be loaded again, using the command `library(forensim)`. This command loads the library into your R session, which will enable you to use all the functions available in Forensim. The LRmix module is launched by the LRmixTK command[1]:

```
> LRmixTK()
```

[1]TK stands for Tcl/Tk, the programming language used for the graphical user interface

This launches a window that is the main interface to the LRmix module (Figure 4).

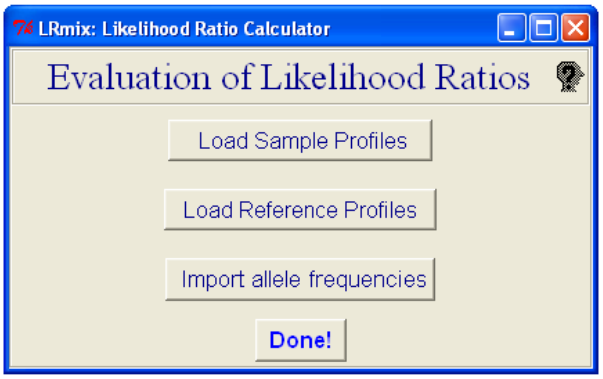

Figure 4: LRmix main graphical user interface.

To be able to use the module you have to make sure that your R session is open, but you can minimize the R windows, and continue using the LRmix interface independently. The module has three buttons that correspond to three steps: first, load the sample profiles, second, load the reference profiles, and third, import the allele frequencies.

3.2 Load sample Profiles

Pressing this button launches a window that allows you to select the files that contain the profiles of the evidence (Figures 5 and 6).

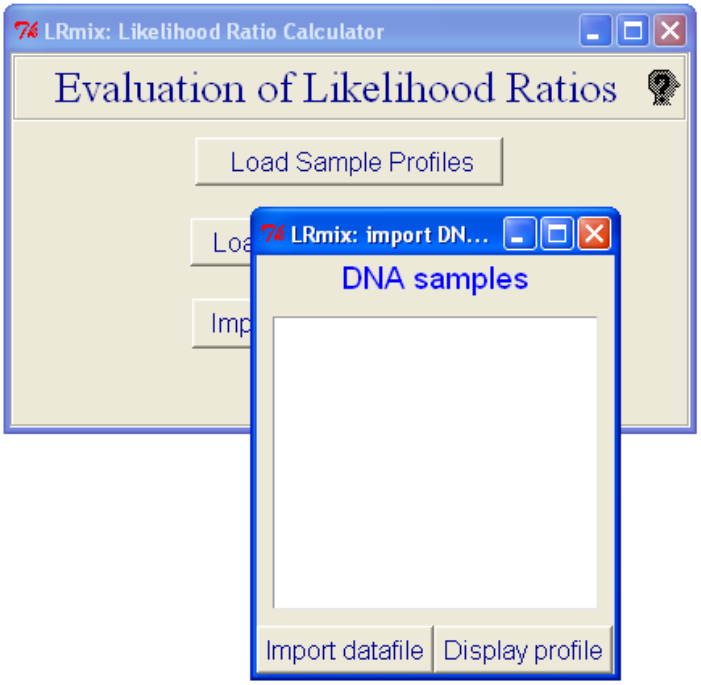

Figure 5: LRmix file upload window for the evidence profile. IN this example, the sampleHammer.csv file has been uploaded into LRmix.

The input files can either be text or CSV files. They are typically obtained by exporting your data using genotyping software as text file table. Table 1 gives an example of such file. The names of the replicates must be indicated using the SampleName column. The Marker column indicates the names of the markers. In this example, the user chose to use the data for the first five alleles. In practice, any number of alleles can be provided to the software. Empty or NA columns will be ignored by LRmix.

SampleName	Marker	Allele1	Allele2	Allele3	Allele4
R1	D3S1358	14	16		
R1	VWA	15	16	19.00	
R1	D16S539	11	13	14.00	
R1	D2S1338	20	23	24.00	25.00
...	...	...	...	...	...
R2	D3S1358	14	16		
R2	VWA	15	16	17.00	19.00
R2	D16S539	11	13	14.00	
R2	D2S1338	20	24	25.00	

Table 1: Required format for the input file for the evidence profile(s), extract of the Hammer case profiles. Note tha there are two replicates R1 and R2.

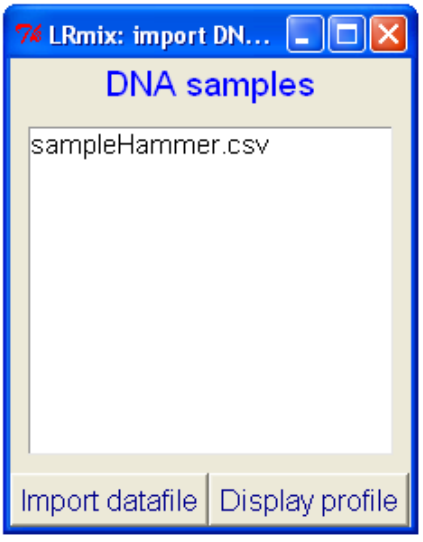

Figure 6: LRmix file upload window for the evidence profile.

Once the file is chosen, the program allows you to see the profiles, and to eventually select the loci as well as the replicates to be analysed (Figure 7). Note that for the purpose of the course, only four replicates can be analysed simultaneously.

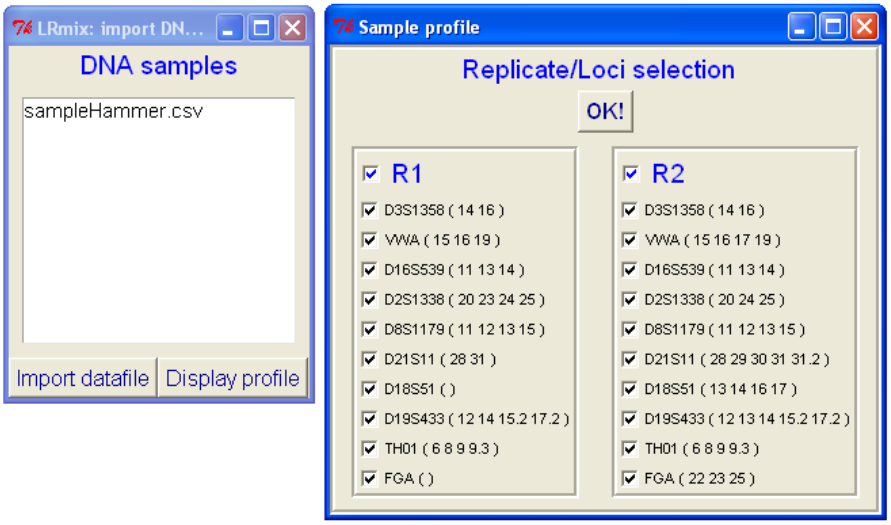

Figure 7: DNA profiles from the Hammer case.

The alleles for each replicate are given between brackets for each locus. By default, all loci are included in the calculations, but you can unselect the loci that you want to exclude from the analysis. Note that if there are no alleles at a given locus (see for example at locus FGA in the Hammer case, replicate 1) LRmix displays empty brackets. Once your choice is made, press OK!, this will close the window. At this stage, the program has recorded your preferences.

3.3 Load reference profiles

The next step now is to load the reference profiles, namely the suspect and the victim. Press OK when you finish uploading your files (Figure 8).

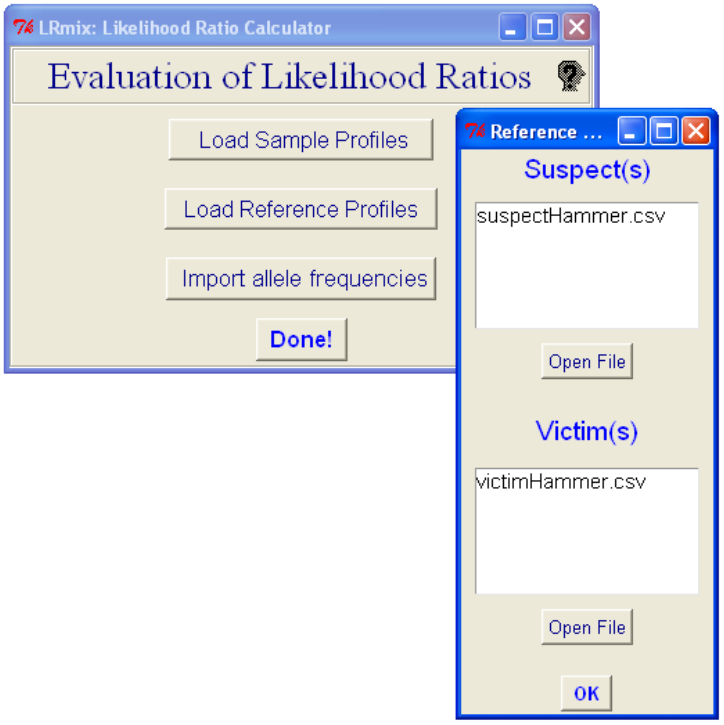

Figure 8: Uploading the reference DNA profiles from the Hammer case.

The selected files should be in the same format as the files used for the sample file (see Table 2). Any number of suspects and victims can be uploaded into the program.

SampleName	Marker	Allele1	Allele2
suspect	D3S1358	14	16
suspect	VWA	15	19
suspect	D16S539	11	14
suspect	D2S1338	24	25
suspect	D8S1179	12	13
suspect	D21S11	28	31
suspect	D18S51	14	17
suspect	D19S433	15.2	17.2
suspect	TH01	9	9.3
suspect	FGA	22	23

Table 2: Required format for the input file for the reference profile(s).

Note that if there two or more suspects, you need to upload a file which contains the profiles of all these suspects. This implies that you want to analyse all the suspects at the same time, if you want to analyse them separately, you need to do separate analyses with different suspect files. It is always compulsory to provide a suspect, but you don't have to provide a victim file. If you have more than one

victim, you need to provide the relevant profiles in a single file, where different individuals have different IDs, see the example in Table 3.

SampleName	Marker	Allele1	Allele2
victim1	D3S1358	16	16
victim1	VWA	15	16
victim1	D16S539	13	13
victim1	D2S1338	20	20
victim1	D8S1179	11	15
victim1	D21S11	29	30
victim1	D18S51	17	17
victim1	D19S433	12	14
victim1	TH01	6	8
victim1	FGA	22	25
victim2	D3S1358	15	17
victim2	VWA	16	19
victim2	D16S539	12	13
victim2	D2S1338	18	25
victim2	D8S1179	11	13
victim2	D21S11	29	30
victim2	D18S51	15	17
victim2	D19S433	14	14
victim2	TH01	6	7
victim2	FGA	20	22

Table 3: Required format for the input file for the reference profile(s). The table gives the profiles of two profiled victims, victim 1 and victim2.

3.4 How to import your own allele frequencies

Since version 3.1 of Forensim, LRmix module allow the user to import allele frequencies. The format is generally found in forensic journals, and is described below. The files can either be given in CSV format (comma separated values), or in text format (with tab separated values). Table 4 gives an example of such file:

Allele	CSF1PO	FGA	TH01	TPOX	VWA	D3S1358
5			0.002	0.002		
6			0.232	0.002		
7			0.190			
8	0.005		0.084	0.535		
8.1						
...	...	...	...	...	...	...
16.2						
17					0.281	0.215
17.2						
18		0.026			0.200	0.152
18.2						
19		0.053			0.104	0.012
19.2						
20		0.127			0.005	0.002
21		0.185			0.002	
21.2		0.005				
22		0.219				
22.2		0.012				
...	...	...	...	...	...	...

Table 4: Required format for the allele frequencies file. Extract from the Identifiler (Applied Biosystems) allele frequencies (Butler et al., 2003).

The first colum 'Allele' gives the allele lengths, the other columns correspond to the loci. The allele frequencies are given in row for each allelic form. Once the file containing the allele frequencies is selected, press the OK! button,

3.5 Analysis

The analysis button launches a window where you have to specify the model parameters.

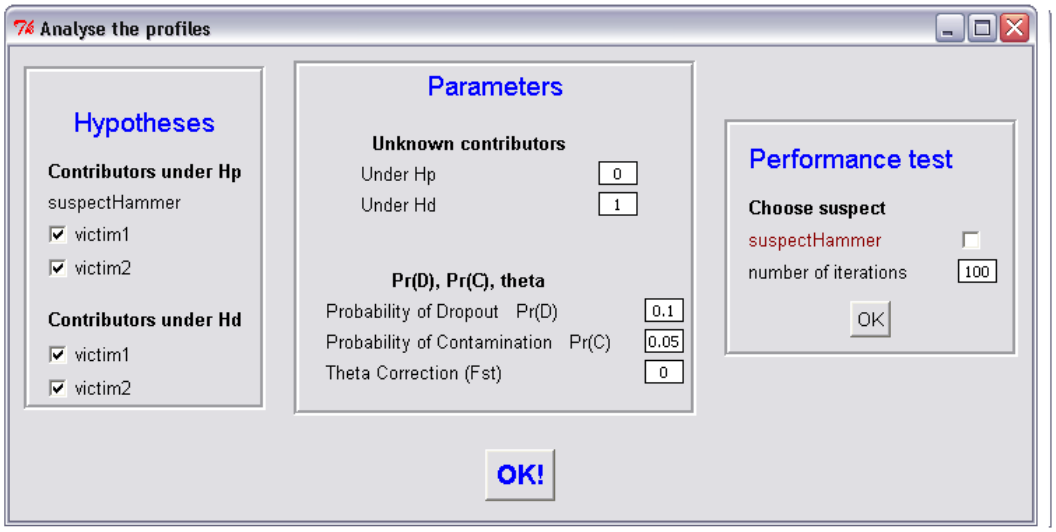

Figure 9: Analysing the DNA profiles from the Hammer case.

This interface allows you to define the hypotheses that you want to evaluate in the likelihood ratio. By default the model selects the suspect and the victim (if provided) as the contributor(s) under Hp, and the victim(s) as the contributors under Hd. The suspect is automatically non-contributor under Hd. Note that you cannot unselect the suspect under Hp, but you can choose to add the victims as contributors under either Hd or Hp. If you provide more than one suspect, all suspects will be considered under Hp. The unknown numbers of contributors must also be specified under each hypothesis. Finally the probabilities of dropout (PrD) and drop-in must be specified, default values are 0.1 and 0.05 respectively. The theta correction is set to zero by default. The OK button launches the computations and the results are displayed in a separate window. The LR is given per locus and overall loci by multiplying the per-locus values (Figure 10).

Likelihood ratios

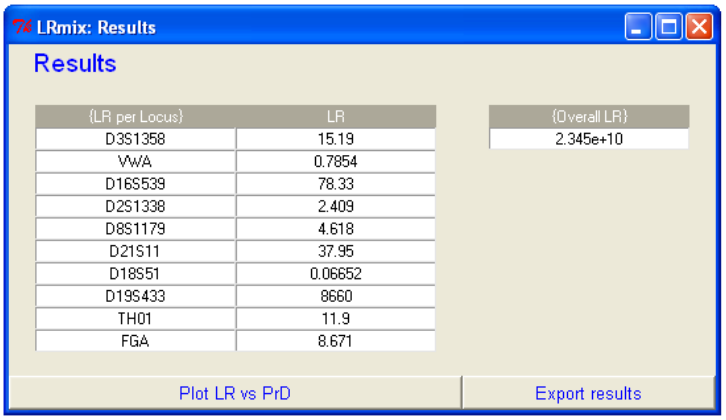

Figure 10: Likelihood ratios of hypotheses Hp and Hd, as specified in Figure 9.

LRmix displays the LRs in a separate window. The results can be saved into a text file (button Export results). The user can also choose to carry on the analysis with a 'sensitivity analysis', this is the exploration of the sensitivity of the likelihood ratios when the dropout probability varies between 0.01 and 0.99. The sensitivity analysis takes longer than the simple evaluation of LRs when a single value of PrD is given. You can follow the progress of the calculations in the R window.

Given the hypotheses and the parameters given in the Analysis window, LRmix tries to find plausible ranges for the probability of dropout following the Monte-Carlo simulation method described in Gill et al. (2007). This qualitative method derives the most plausible ranges of PrD, based on the total number of alleles in the sample profiles. Conditioned on the genotypes specified under each hypothesis, the program simulates a large number of mixtures that have the same composition in alleles than the questioned sample, and looks for the levels of dropout that could have generated a sample with the same number of alleles. Because the method relies on the hypothesised contributors under each hypothesis, the estimation is carried out separately under Hp and under Hd. The minimum and the maximum values obtained across the two analyses, are reported on the sensitivity analysis plot. The results of the sensitivity analysis can be exported as a text file (Export results button), the range of drop-out are given in at the bottom of the file, as the 5% and the 95% percentiles of the empirical distributions of the probabilities of dropout, under Hp and under Hd.

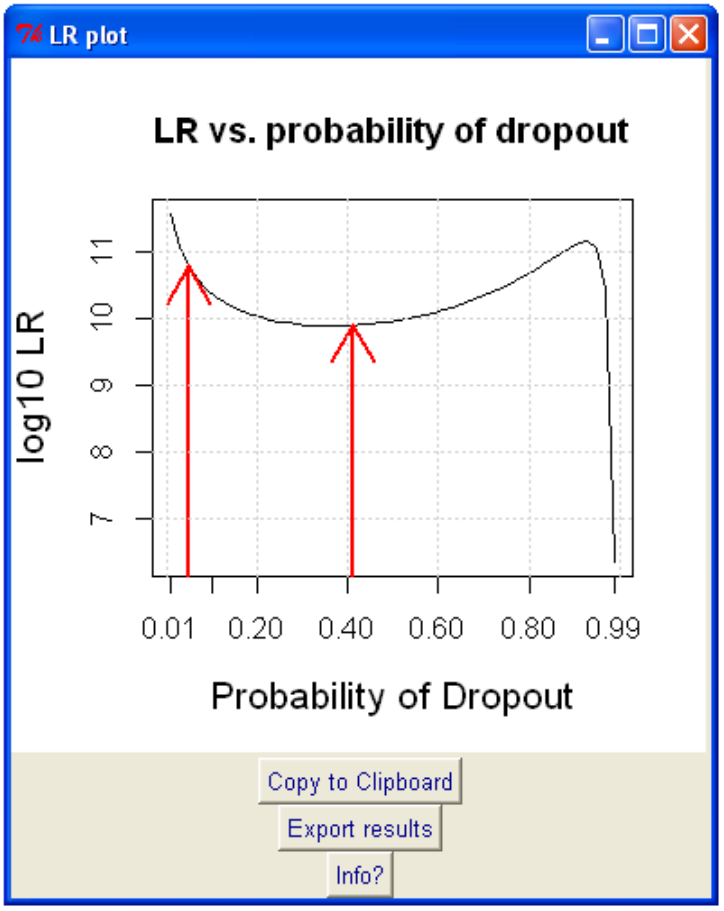

Figure 11: Sensitivity analysis of the LR to variations in the dropout probability. The red arrow correspond to the most plausible ranges for the probabilities of dropout, derived via Monte-Carlo simulations.

3.6 Performance plots

LRmix offers the possibility to carry out performance studies, on a per-case basis, using the principles outlined by Gill et al. (2008). Performance plots are implemented to enable the evaluation of likelihood ratios when the suspect is substituted with a random man, simulated by randomly drawing alleles from the allele frequencies provided by the user. Only one suspect can be evaluated at a time, thus, if multiple suspects are evaluated, the user has to choose which suspect has to be replaced by a random man. The parameters and hypotheses specified in the Analysis window are the one used in the LRs calculated in the Performance plot module.

The user can choose the number of iterations, which corresponds to the number of simulations (i.e. the number of random men to simulate) the program has to run to build the distribution of LRs. Increasing this number, which is set by default to 100, increases the computation time and may slow down the program. The progress of the computation can be followed in the R console.

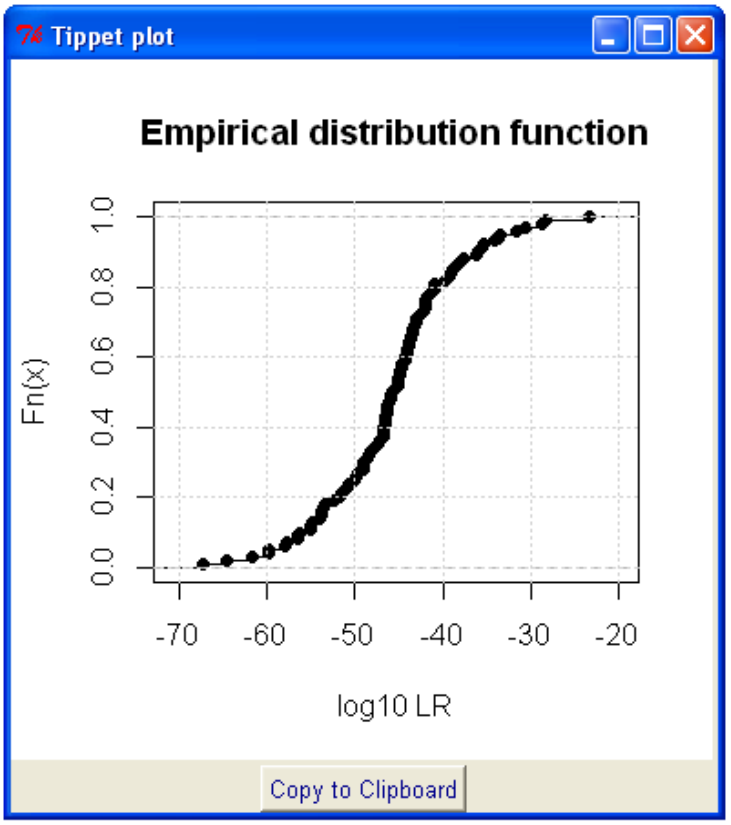

Figure 12: Performance plot generated by LRmix in the Hammer case.

4 Debugging

4.1 Use this checklist before you load your data into LRmix

If you are using CSV files

- Check that the column names of your files are: SampleName, Marker, Allele1, Allele2, Allele3...
- Check that there are no spaces in the column names
- If you are uploading sample profiles, you can add as many alleles as you want, if you are uploading reference profile, you must only add Allele1 and Allele2
- The field separator must be the comma ‘,’
- The decimal separator must be the dot ‘.’
- Do not provide the Amel locus
- Make sure that the marker names in your files are consistent with the markers provided in the allele frequencies files (see blow)

If you are using txt files Please check that:

- Check that the column names of your files are: SampleName, Marker, Allele1, Allele2, Allele3...
- Check that there are no spaces in the column names
- If you are uploading sample profiles, you can add as many alleles as you want, if you are uploading reference profile, you must only add Allele1 and Allele2
- The field separator is the 'tab', as typically obtained from an Excel file.
- The decimal separator must be the dot '.'

Note that it does not matter to LRmix whether there are quotes in your files or not.

4.2 Common errors

- To avoid format errors, refer to the example files given on Forensim website. Typical errors consist in using the comma ',' instead of the dot '.' as a decimal separator in the data files. If you are encountering problems uploading your files into LRmix, open your files under a text editor (e.g. Notepad++) and display the spaces, this may help find where the errors are.
- Another common error is that the allele provided in the data files are not recovered in the allele frequencies files uploaded by the user, so make sure that in the allele frequencies files, all relevant alleles are listed.
- LRmix notifies the user if it fails to determine the dropout ranges. Keep in mind that this qualitative approach depends on the hypotheses, thus if n alleles are observed and only one contributor is hypothesised under a given hypothesis, the program may fail to derive the ranges, which means that no occurrences of n alleles were found in the Monte-Carlo simulations with the assumed contributor. In this case, reconsider the hypotheses (if relevant) and rerun the porgram.

5 Workshop

Several cases are explored during the practical sessions, and participants are encouraged to analyse their own cases during the course. The Hammer case is provided as an example on Forensim website. The case files are provided both in CSV and txt formats. The case profiles are provided in two zipped folders (the txt and CSV formats). To get the files, simply unzip the folders. It is recommended that you create a working folder for the course, and start R in that folder. Windows users can simply copy the R blue icon in the working folder (shortcut for R), and start R by a double-click. To make sure that R starts in the working folder, right-click on the blue icon, and make sure the "start in" entry is left blank.

During the course, only the LRmixTK module is used, but you can read more about R, for example:

- "An Introduction to R" http://cran.r-project.org/doc/manuals/R-intro.pdf
- "Using R for Data Analysis and Graphics - Introduction, Examples and Commentary"http://cran.r-project.org/doc/contrib/usingR.pdf

References

D. J. Balding and R. A. Nichols. DNA profile match probability calculation: how to allow for population stratification, relatedness, databse selection and single bands. *Forensic Sci. Int.*, 64:125–140, 1994.

J. Butler, R. Schoske, M. Vallone, J. W. Redman, and M. C. Kline. Allele frequencies for 15 autosomal str loci on u.s. caucasian, african american, and hispanic populations. *Journal of Forensic Sciences*, 48(8):908–911, 2003.

J. M. Curran, P. Gill, and M. R. Bill. Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure. *Forensic Sci. Int.*, 148:47–53, 2005.

P. Gill, A. Kirkham, and J. Curran. LoComatioN: A software tool for the analysis of low copy number DNA profiles. *Forensic Sci. Int.*, 166(2-3):128–138, 2007.

P. Gill, J. M. Curran, C. Neumann, A. Kirkham, T. Clayton, J. Whitaker, and J. Lambert. Interpretation of complex dna profiles using empirical models and a method to measure their robustness. *Forensic Science International: Genetics*, 2: 91–103, 2008.

H. Haned. Forensim: an open source initiative for the evaluation of statistical methos in forensic genetics. *Forensic Sci. Int. Genet.*, 5(4):265–268, 2011.