ISFG Educational Workshop 2011

Interpretation of Complex STR Results Using the Forensim Package

Peter Gill

Hinda Haned

Contents

1 The forensim package

1.1 Overview. Documentation

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 allows to calculate likelihood ratios for complex STR profiles.

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`.

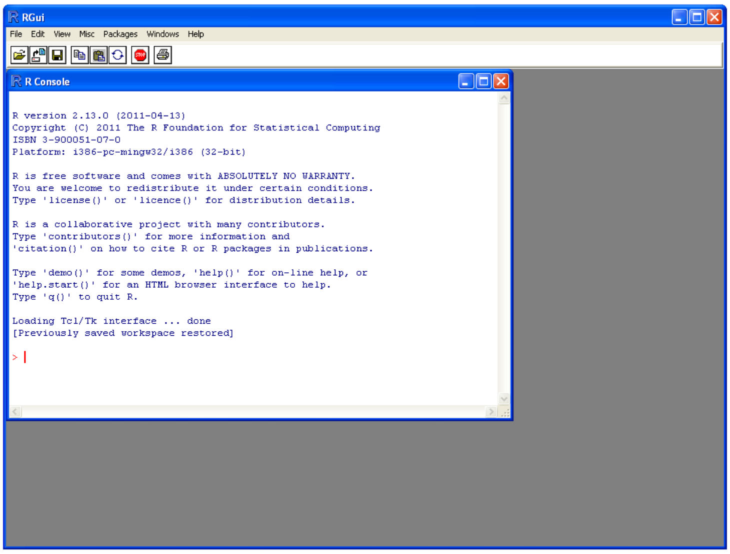
1.2 Software installation

Before we start, make sure you have installed R properly.

1.2.1 Install the R software

The R software is available from the Comprehensive R Archive Network (CRAN). Hereafter we explain how the software can be installed:

- Go to `http://www.r-project.org/`
- In the Getting Started tab, go to : DownloadR
- Choose a CRAN mirror (preferably one close to where you live)
  - Argentina: `http://mirror.fcaglp.unlp.edu.ar/CRAN/`
  - Netherlands: `http://cran.xl-mirror.nl/`
- Dependent on which operating system you use, click on one of the links: Linux, MacOS X or Windows. For Windows:
  - Click the "base" link
  - Click the link "Download R 2.13.1 for Windows", run the file and the installation program will start.
- Click on R-2.13.1.exe to install the set-up file
- After installation, a blue colored icon on your desktop, click on the icon to launch an R session (Figure 1).



```
R version 2.13.0 (2011-04-13)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-pc-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Loading Tcl/Tk interface ... done
[Previously saved workspace restored]

> 
```

Figure 1: R session (Windows)

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

1.2.2 Install the Forensim package

Forensim and its dependencies can be found on the CRAN website http://www.r-project.org. In the left menu, under the "Download, Packages" tab, click on the CRAN link. Choose a CRAN mirror, ideally one that is close to where you live. In the new window, 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_2.0.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. gdata
2. gtools
3. combinat
4. MASS
5. mvtnorm
6. genetics
7. tcltk2
8. tkrplot.

1.2.3 Install and load packages in R

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_2.0.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 2).

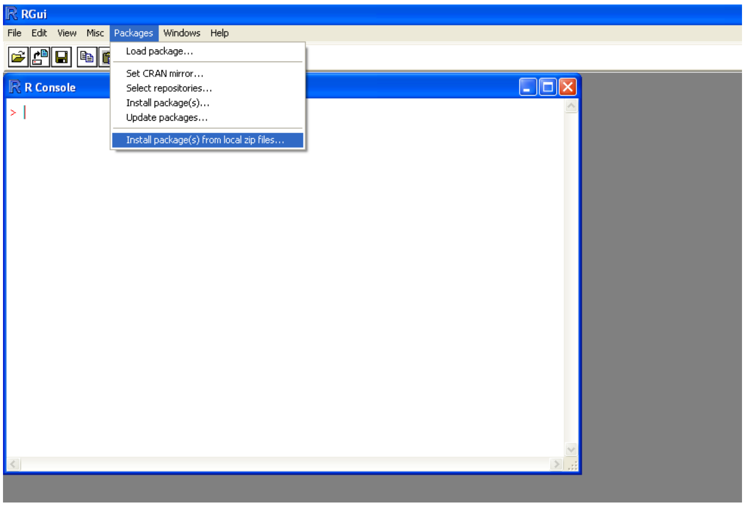


Figure 2: Package download in (Windows system).

The present document serves to introduce a particular functionality of the Forensim package, the LRmix module.

2 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 (2010).

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).

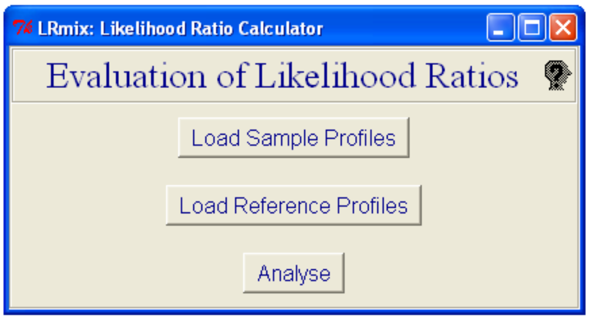

LRmix: Likelihood Ratio Calculator
Evaluation of Likelihood Ratios
Load Sample Profiles
Load Reference Profiles
Analyse

Figure 3: LRmix main graphical user interface

2.1 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. To launch the module, you have to simply type the following R commands into the R session:

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

```
> library(forensim)


   ### forensim 2.0 is loaded ###
```

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()
```

This launches a window that is the main interface to the LRmix module (Figure 3). 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, interpret the evidence using likelihood ratios.

2.2 Load sample Profiles

Pressing this button launches a window that allows you to select the files that contain the profiles of the evidence (figure 4).

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

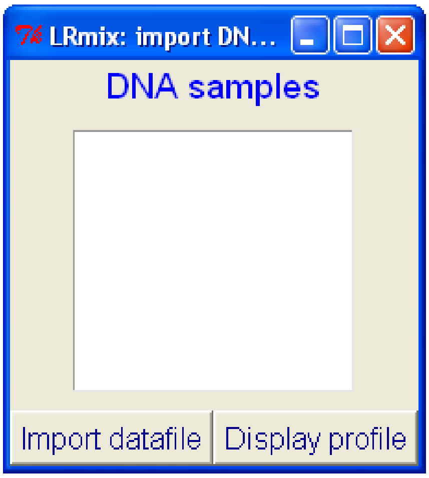


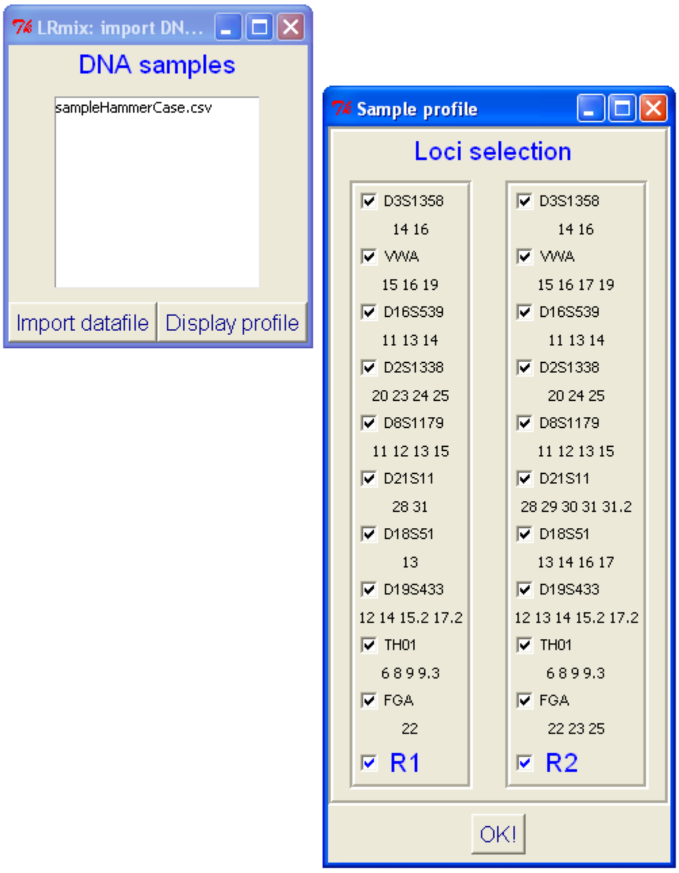
Figure 4: LRmix file upload window for the evidence profile.

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	Allele5
R1	AMEL	X	Y			
R1	D3S1358	14	16			
R1	VWA	15	16	19		
R1	D16S539	11	13	14		
R1	D2S1338	20	23	24	25	
R1	D8S1179	11	12	13	15	
R1	D21S11	28	31			
R1	D18S51	13				
R1	D19S433	12	14	15.2	17.2	
R1	TH01	6	8	9	9.3	
R1	FGA	22				

Table 1: Required format for the input file for the evidence profile(s).

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 5). Note that for the purpose of the course, only four replicates can be analysed simultaneously.



sampleHammerCase.csv

☑ D3S1358
14 16
☑ VWA
15 16 19
☑ D16S539
11 13 14
☑ D2S1338
20 23 24 25
☑ D8S1179
11 12 13 15
☑ D21S11
28 31
☑ D18S51
13
☑ D19S433
12 14 15.2 17.2
☑ TH01
6 8 9 9.3
☑ FGA
22
☑ R1

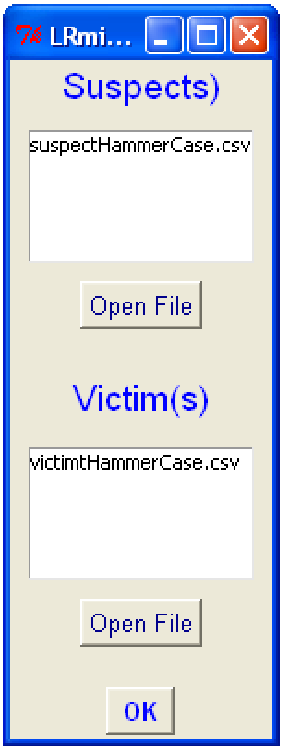
☑ D3S1358
14 16
☑ VWA
15 16 17 19
☑ D16S539
11 13 14
☑ D2S1338
20 24 25
☑ D8S1179
11 12 13 15
☑ D21S11
28 29 30 31 31.2
☑ D18S51
13 14 16 17
☑ D19S433
12 13 14 15.2 17.2
☑ TH01
6 8 9 9.3
☑ FGA
22 23 25
☑ R2

Figure 5: DNA profiles from the Hammer case.

Once your choice is made, simply press OK!, this will close the window. The program has recorded your preferences.

2.3 Load reference profiles

The next step now is to load the reference profiles, namely the suspect and the victim. Simply press OK when you finish uploading (figure 6).



suspectHammerCase.csv

victimtHammerCase.csv

Figure 6: Uploading the reference DNA profiles from the Hammer case.

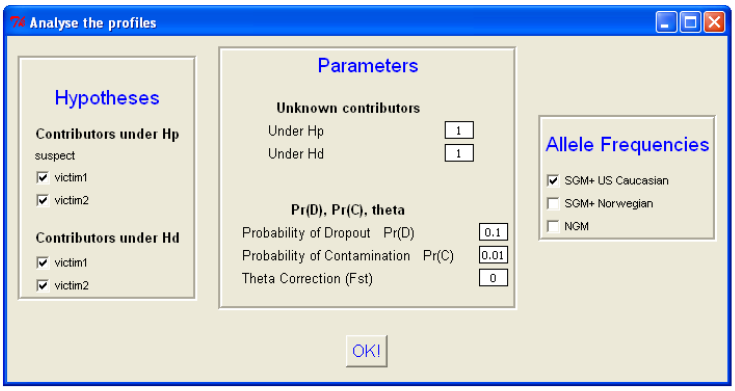
The selected files should be in the same format as the files used for the sample file (see Table 2). Any number of suspects and victims can be uploaded into the program.

SampleName	Marker	Allele1	Allele2
suspect	AMEL	X	Y
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).

2.4 Analysis

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



Analyse the profiles

Hypotheses

Contributors under Hp

suspect

☑ victim1

☑ victim2

Contributors under Hd

☑ victim1

☑ victim2

Parameters

Unknown contributors

Under Hp 1

Under Hd 1

Pr(D), Pr(C), theta

Probability of Dropout Pr(D) 0.1

Probability of Contamination Pr(C) 0.01

Theta Correction (Fst) 0

Allele Frequencies

☑ SGM+ US Caucasian

□ SGM+ Norwegian

□ NGM

Figure 7: Analysing the DNA profiles from the Hammer case.

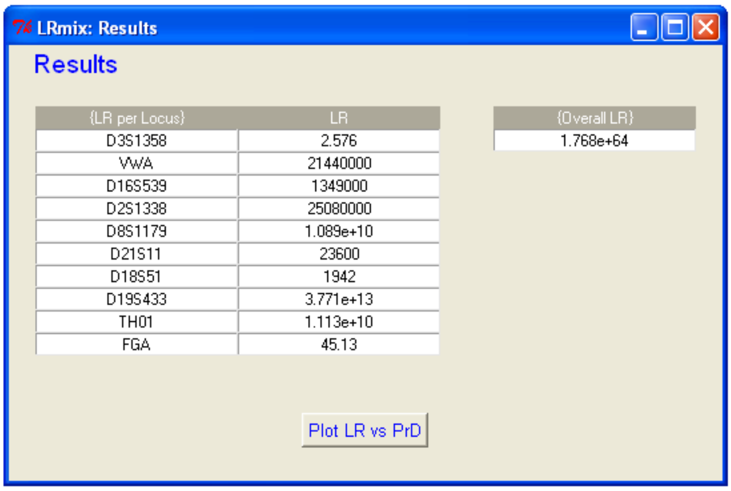
By default the model selects the suspect and the victim (if provided) as the contributor(s) under Hp, and the victim(s) as the contributors under Hd. The suspect is automatically non-contributor under Hd.

The unknown numbers of contributors must also be specified under each hypothesis. Finally the probabilities of dropout and drop-in must be specified, default values are 0.1 and 0.01 respectively. The theta correction is set to zero by default.

Different values of dropout probabilities are applied to homozygotes and heterozygotes. We denote D the probability of dropout for heterozygotes and D_2 the probability of dropout for homozygotes. Following Balding and Buckleton (2009): $D_2 = \alpha D^2$. In LRmix, $\alpha = \frac{1}{2}$. 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 8).

Allele frequencies The allele frequencies can be chosen among three datasets:

- "SGM+ US Caucasian": allele frequencies for the US Caucasian population (Butler et al., 2003).
- "SGM+ Norwegian": allele frequencies for the Norwegian population (Andreassen et al., 2007).
- "NGM": allele frequencies (Budowle et al., 2011).



{LR per Locus}	LR
D3S1358	2.576
VWA	21440000
D16S539	1349000
D2S1338	25080000
D8S1179	1.089e+10
D21S11	23600
D18S51	1942
D19S433	3.771e+13
TH01	1.113e+10
FGA	45.13

{Overall LR}
1.768e+64

Figure 8: Likelihood ratios obtained for the Hammer case.

Note that is also possible to plot the LR for varying values of the probability of drop-out. This functionality will be further explored during the practical session.

3 Application

Two cases are explored during the practical session, they both involve the analysis of mixed DNA stains. The data files containing the DNA profiles are available on Forensim website. Three CSV files are provided for each case:

- Sample: csv file containing the tabulated profile(s) of two PCR amplifications of the crime scene samples. The sample is analysed with the SGM+ kit.
- Suspect: csv file containing the tabulated profil(e) of the suspect(s).
- Victim: csv file containing the tabulated profile(s) of the victim(s).

The tabulated profiles for the two cases are given in the folders Case 1 and Case 2, available as zipped files on Forensim website. Further information about each case will be given during the course.

These profiles are provided as CSV files in two zipped folders. 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

Andreassen, R., Jakobsen, S. and Mevaag, B. (2007), “Norwegian population data for the 10 autosomal STR loci in the AMPFlSTR(R) SGM Plus(TM) system”, *Forensic Sci. Int.* , Vol. 170(1), pp. 59–61.

Balding, D. and Buckleton, J. (2009), “Interpreting low template DNA profiles.”, *Forensic science international. Genetics* , Vol. 4, pp. 1–10.

Balding, D. J. and Nichols, R. A. (1994), “DNA profile match probability calculation: how to allow for population stratification, relatedness, databse selection and single bands”, *Forensic Science International* , Vol. 64, pp. 125–140.

Budowle, B., Ge, J., Chakraborty, R., Eisenberg, A., Green, R., Mulero, J., Lagace, R. and Hennessy, L. (2011), “Population genetic analyses of the NGM STR loci”, *International Journal of Legal Medicine* , Springer, pp. 1–9.

Butler, J., Schoske, R., Vallone, M., Redman, J. W. and Kline, M. C. (2003), “Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations.”, *Journal of Forensic Sciences* , Vol. 48(8), pp. 908–911.

Curran, J. M., Gill, P. and Bill, M. R. (2005), “Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure”, *Forensic Science International* , Vol. 148, pp. 47–53.

Haned, H. (2010), “Forensim: an open source initiative for the evaluation of statistical methos in forensic genetics”, *Forensic Science International Genetics* .