

# Tutorial for EuroForMix v1

Author: Øyvind Bleka <Oyvind.Bleka.at.fhi.no>

Date: 12-04-2014

## Part 1: Installation and running program:

- 1) Install and run R ( $\geq 3.0.1$ ) in Windows, Linux or MAC (<http://cran.r-project.org/>).
  - a. Note that this is only tested on a Windows 7 OS (at current moment).
- 2) Copy and run these commands in the R-software to install the required packages:

```
install.packages("gWidgets")
install.packages("gWidgetstcltk")
install.packages("forensim")
install.packages("cubature")
install.packages("gammadnamix", repos="http://R-Forge.R-project.org")
```

- 3) Run these commands in the R-software to start the GUI EuroForMix

```
library(gammadnamix)
efm()
```

## Part 2: How to use EuroForMix

### Get started:

- 1) Find the R-installation folder:
  - a. Find the name of the folder where you installed your R software.
    - i. For instance (in Windows 7), my R version was installed at "C:\Program Files\R\R-3.1.2\" or "C:\Programfiler\R\R-3.1.2\" (for the Norwegian OS version).
- 2) Get GUI as top layer (in Windows):
  - a. Set the EuroForMix GUI as top layer by using ALT+TAB (at keyboard).
- 3) Selecting the Working Directory to access the tutorial data:
  - a. Click on **File** and then *Set Directory* at the Toolbar.
  - b. Select the folder "C:\Program Files\R\R-3.1.2\library\gammadnamix\tutorialdata"
  - c. Press **OK**.

## The basics:

### 1) Import Population frequencies:

- a. Press **1) Select directory** button.
  - i. Find back the folder you selected as Working Directory. Click on the folder and click on the folder *FreqDatabases*. Then press **OK** and go back to the GUI.
- b. Press **2) Import from directory** button
  - i. Now all population frequency files in the selected folder *FreqDatabases* are loaded into the software.

### 2) Import Evidences and References:

- a. Press **Import evidence** button.
  - i. Click on *stain.txt* and then press **Open**.
    1. The stain evidence is now loaded into the software.
- b. Press **Import reference** button.
  - i. Click on *refs.csv* and then press **Open**.
    1. The reference profiles are now loaded into the software.

### 3) View the evidence:

- a. Check/select *evid1*.
- b. Press **View evidence**
  - i. An EPG of the evidence is shown as a plot in the R software.
  - ii. The allele-names and corresponding peak heights are printed out to the R-terminal.
  - iii. Notice that marker set corresponds with the SGMPlus kit.

### 4) View references and their matching summary against the selected evidence.

- a. Check/select *Victim* and *Suspect*.
- b. Press **View references**
  - i. A table with the genotypes of the references is printed out to the R-terminal.
  - ii. Since the *evid1* stain is selected, a matching summary table between the references and the stain is also printed out to the R-terminal.
    1. MAC is number of matching alleles in to reference profiles to the stain.
    2. nLocs is number of valid markers which are used for the match.

### 5) Selecting a kit and population frequency to use in the evaluations.

- a. Go to the **Select kit:** - drop-down list and select *SGMPlus*.
- b. Go to the **Select population:** - drop-down list and select *UK*.
  - i. The frequencies of the SGMPlus typed UK population is now selected.

- 6) View frequencies and get the false positive match probabilities.
  - a. Press **View frequencies**
    - i. The imported and selected population frequencies are shown in an own GUI window.
    - ii. Go to the R-software frame.
      1. Since *evid1* stain is selected, the software calculates and shows the probability that a random man in the population matches the stain with at least “k” alleles (i.e.  $\Pr(\text{MAC} \geq k)$ ). The plot shows this probability as a function of “k” number of matching alleles.
      2. Victim has MAC=20, which means the random match probability of the victim becomes  $2.2 \times 10^{-8}$ .
      3. Suspect has MAC=16, which means the random match probability of the suspect becomes  $6.9 \times 10^{-3}$ .
- 7) Saving project for later restoring.
  - a. Under **File** in Toolbar, press **Save project**. Name a filename (e.g. “proj”), select a folder where you want to save the project-file and press save.
    - i. The project (with all imported data and evaluations) can now be restored by pressing **Open project** under **File** in Toolbar at any time.
    - ii. This is useful to quickly restore a project session again after the GUI has been closed.

## Weight evidence with a continuous model:

- 8) Check/select *evid1* and *Suspect* and then press **Weight-of-Evidence**
  - a. You then come to the Model specification GUI page.
- 9) First, specify the contributors under hypotheses Hp and Hd.
  - a. Hp : "The suspect profile and 1 unknown individual contributes to *evid1*"
  - b. Hd : " 2 unknown individuals contributes to *evid1*"
- 10) Specify model parameters:
  - a. If peak heights are imported in the evaluated evidence, the **Detection threshold** should first be specified as the lower peak height limit for the imported evidence. Let this be 150 as default.
  - b. The imported evidence is here not applied with any (n-1) stutter-filter (n is as allele name), and hence we need to assume a (n-1) stutter ratio in the model.
    - i. To assume that the (n-1)-stutters in the model has an unknown ratio, the box of **Stutter ratio (xi)** must be empty (as default).
    - ii. If the stutter ratio is known, the box of **Stutter ratio (xi)** can contain this known ratio.
    - iii. The **Q-assignment** means that non-present alleles in the evidence is assigned as allele Q (i.e. "99") with the allele frequency as sum of all the non-present allele frequencies. This will speed things up if checked, but may lead to some inaccuracy when taking account for (n-1)-stutters.
  - c. Leave all values as default.
- 11) Calculate **Continuous LR (Maximum Likelihood based)**:
  - a. Press **Continuous LR (Maximum Likelihood based)**.
    - i. The user is now re-directed to the MLE fit GUI page.
    - ii. The software now optimizes the Likelihood (under each hypothesis) as a function of the unknown parameters in the continuous model:
      1. **mx=(mx1,..., mxC)**: mixture proportion for contributor 1,...,C.
      2. **mu**: amount of DNA
      3. **sigma**: coefficient of variance
      4. **xi**: (n-1)-stutter ratio
  - b. Press the **Optimize model more** button (under *Further evaluation*) to be sure that the Likelihood functions are optimized.
    - i. The optimized likelihood values are given under *Maximum Likelihood value*.
  - c. Press **Model validation** under *Further action* for each of the hypothesis to test whether the fitted model is not adequate with the observed peak heights.
    - i. A Goodness-of-fit test reports a p-value for the test.
      1. This is printed out to terminal.

- ii. A large p-value indicates that the fitted model fits the observed peak heights very well, while small p-values indicates that the fitted model does not fit the observed peak heights very well.
- d. The LR values under *Weight-of-evidence* are based on the optimized likelihood functions. More specific, the **Joint LR** values are the ratio between the optimized likelihood value under  $H_p$  and the optimized likelihood value under  $H_d$ .
  - i. LR for each locus is also conditioned on the optimized parameters.
- e. In order to take the uncertainty of the parameter estimates into account to the LR value, the user may press **Simulate LR distribution** (under *Further evaluation*) in order to simulate 10000 from the LR distribution over the posterior space of the unknown parameters in each of the hypothesis.
  - i. This could take a while, depending on number of samples (this can be changed under MCMC in Toolbar).
  - ii. A density smoothed plot is given the R-software together a range of quantiles printed out to the terminal.
  - iii. Note:
    1. The user could report the 5% quantile as a conservative LR value (I got  $\log_{10}=2.63$ ).
    2. Note that the **Joint LR** values should lie around the mode of the simulated density. Otherwise, the optimization procedure hasn't reached the maximum likelihood value.

## 12) Calculate **Continuous LR (Integrated Likelihood based)**:

- a. Press **Continuous LR (Integrated Likelihood based)**.
  - i. The software now integrates out the unknown parameters in each of the likelihood functions to make a marginalized calculates Likelihood Ratio weight-of-evidence which are independent of the unknown parameters.
    1. A flat prior is considered on all the unknown parameters in the continuous model (see vignette for more details).
  - ii. Note:
    1. The integration depends on a relative error parameter which gives to accuracy of the integral. This is default 0.005 but can be changed under **Set relative error requirement** under *Integration* at the Toolbar.
    2. This calculation can be done directly from the Model specification GUI page as well.

## Deconvolution:

- 1) If you followed the steps in Weight evidence with a continuous model:
  - a. The user should now be at the MLE fit GUI page.
  - b. Remember our hypothesis  $H_p$ : "The *suspect* profile and 1 unknown individual contributes to *evid1*", where there is 1 unknown individual in the hypothesis.
  - c. Assume that we have gathered extra information to say that we know that the *suspect* reference is a true contributor to *evid1*.
  - d. To do deconvolution on the unknown individual based on the fitted model, press **Deconvolution** under the section *Estimates under  $H_p$* .
  - e. The user is now re-directed to the Deconvolution GUI page which shows a ranked table of the most probable unknown jointly genotype profiles for the unknown.
    - i. Allele "99" means that the allele is not presented in the evidence.
    - ii. The **posterior** value is the posterior probability of the jointly combined genotypes presented at each row conditioned on the maximum likelihood estimated parameters (see vignette for more details).
- 2) If you did not follow the steps in Weight evidence with a continuous model:
  - a. Check/select *evid1* and *Suspect* and then press **Deconvolution**
    - i. You then come to the Model specification GUI page.
  - b. First, specify the contributors under hypotheses  $H_d$ .
    - i.  $H_d$  : "The suspect profile and 1 unknown individual contributes to *evid1*"
  - c. Specify model parameters:
    - i. If peak heights are imported in the evaluated evidence, the **Detection threshold** should first be specified as the lower peak height limit for the imported evidence. Let this be 150 as default.
    - ii. The imported evidence is here not applied with any (n-1) stutter-filter (n is as allele name), and hence we need to assume a (n-1) stutter ratio in the model.
      1. To assume that the (n-1)-stutters in the model has an unknown ratio, the box of **Stutter ratio (xi)** must be empty (as default).
      2. If the stutter ratio is known, the box of **Stutter ratio (xi)** can contain this known ratio.
      3. The **Q-assignment** means that non-present alleles in the evidence is assigned as allele Q (i.e. "99") with the allele frequency as sum of all the non-present allele frequencies.
        - C) This should always be used when doing deconvolution.
    - iii. Leave all values as default.
  - d. Press **Continuous LR (Maximum Likelihood based)**.
    - i. The user is now re-directed to the MLE fit GUI page.

- ii. The software now optimizes the Likelihood (under Hd) as a function of the unknown parameters in the continuous model:
    - 1. **mx**=(**mx1**,..., **mxC**): mixture proportion for contributor 1,...,C.
    - 2. **mu**: amount of DNA
    - 3. **sigma**: coefficient of variance
    - 4. **xi**: (n-1) stutter ratio
- e. Press the **Optimize model more** button (under *Further evaluation*) to be sure that the Likelihood function is optimized.
  - i. The optimized likelihood values are given under *Maximum Likelihood value*.
- f. Press **Model validation** under *Further action* for each of the hypothesis to test whether the fitted model is not adequate with the observed peak heights.
  - i. A Goodness-of-fit test reports a p-value for the test.
    - 1. This is printed out to terminal.
  - ii. A large p-value indicates that the fitted model fits the observed peak heights very well, while small p-values indicates that the fitted model does not fit the observed peak heights very well.
- g. To do deconvolution on the unknown individual based on the fitted model, press **Deconvolution** under the section *Estimates under Hd*.
- h. The user is now re-directed to the Deconvolution GUI page which shows a ranked table of the most probable unknown jointly genotype profiles for the unknown.
  - i. Allele "99" means that the allele is not presented in the evidence.
  - ii. The **posterior** value is the posterior probability of the jointly combined genotypes presented at each row conditioned on the maximum likelihood estimated parameters (see vignette for more details).

## Weight evidence with a qualitative model:

- 1) Follow the import steps under The basics.
  - a. You should then have the evidence *evid1* and the references *Victim* and *Suspect* imported to the software.
  - b. You should have selected the *UK* population frequencies for the *SGMPlus* kit.
- 2) First, specify the contributors under hypotheses Hp and Hd.
  - a. Hp : "The *victim* and *suspect* profile contributes to *evid1*"
    - i. Change number of unknowns under Hp to 0.
  - b. Hd : "The victim profile contributes and 1 unknown individual contributes to *evid1*"
    - i. Under Hd, check/select the *victim* and change number of unknowns under to 1.
- 3) Specify model parameters:
  - a. We will consider a qualitative model to evaluate the evidence.
  - b. If peak heights are imported in the evaluated evidence, the **Detection threshold** should first be specified as the lower peak height limit for the imported evidence. Let this be 150 as default.
  - c. The Qualitative Model Parameters
    - i. We let **Probability of Dropin** and **fst-correction** be set as default (0.05 and 0.02 respectively).
- 4) Press **Qualitative LR (semi-continuous)** which re-directs the user to the Qual. LR GUI.
- 5) Press **Sensitivity**
  - a. A plot in the R-software shows the Likelihood Ratio (Weight-of-evidence) as a function of probability of allele dropout (equal for same contributors).
    - i. Number of ticks and max probability can be changed under *Qual LR* at the toolbar.
- 6) Press **Conservative LR**
  - a. The 5% and 95% quantiles in the distribution of "allele dropout probability given number of total observed alleles in the evidence" are estimated using 2000 samples.
    - i. Number of samples and significance level for quantiles can be changed under *Qual LR* at the toolbar.
  - b. The estimated quantiles are printed out at the terminal in the R-software.
  - c. The reporting LR under Weight-of-Evidence uses the allele drop-out probability which gives the smallest LR (to make it conservative in favor of the defendant).
- 7) Press **Do tippet**



- a. The “tippet” reference *Suspect* is exchanged with a random man in the selected population. This is sampled 1e6 times (default) and used to show a cumulative density of these samples by plotting in the R-software.
  - i. The blue line is the Joint log10LR under Weight-of-evidence.

## Database search:

- 8) Follow the import steps under The basics.
  - a. You should then have the evidence *evid1* and the references *Victim* and *Suspect* imported to the software.
  - b. You should have selected the *UK* population frequencies for the *SGMPlus* kit.
- 9) Press **Import database**, select the file *databaseESX17.txt* in the *tutorialdata*-folder and press **Open**.
  - a. A database with 77 ESX17 typed reference profiles are then imported to the software.
  - b. The output on the R-terminal shows that allele 7 in D18S51 and allele 19.2 was in FGA was missing in the population frequencies, but are each assigned as the minimum observed allele frequency.
    - i. The allele frequencies are after normalized to 1.
  - c. Tips:
    - i. If a database file contains millions of references, it is very useful to split the file up and import each (split) files separately into the software.
      - 1. Avoid the limitation of Computer Memory
    - ii. If the importing process of a reference database takes long time, the user should save the session as a project (use **Save project** under *File*) to avoid the need of importing the same reference database again.
      - 1. Stores big databases very efficient.
- 10) Check/select *evid1*, *Victim* and *databaseESX17* and then press **Database search**
  - a. We will now assume that *Victim* is a true contributor to the evidence.
  - b. You then come to the Model specification GUI page.
- 11) First, specify the contributors under hypotheses *Hp* and *Hd*.
  - a. *Hp* : “The Database-reference and the victim profile contributes to *evid1*”
    - i. Change number of unknowns under *Hp* to 0.
  - b. *Hd* : “The victim profile contributes and 1 unknown individual contributes to *evid1*”
    - i. Change number of unknowns under *Hd* to 1.

12) Specify model parameters:

- a. If peak heights are imported in the evaluated evidence, the **Detection threshold** should first be specified as the lower peak height limit for the imported evidence. Let this be 150 as default.
- b. The imported evidence is here not applied with any (n-1) stutter-filter (n is as allele name), and hence we need to assume a (n-1) stutter ratio in the model.
  - i. To assume that the (n-1)-stutters in the model has an unknown ratio, the box of **Stutter ratio (xi)** must be empty (as default).
  - ii. If the stutter ratio is known, the box of **Stutter ratio (xi)** can contain this known ratio.
  - iii. The **Q-assignment** means that non-present alleles in the evidence is assigned as allele Q (i.e. "99") with the allele frequency as sum of all the non-present allele frequencies. This will speed things up if checked, but may lead to some inaccuracy when taking account for (n-1)-stutters.
- c. We will now assume that allele drop-in can occur with a given peak height model.
  - i. Change **Probability of Dropin** to 0.001 (i.e. the probability of having a allele drop-in event to a particular marker) under *Continuous Model parameters*.
  - ii. Change **Dropin peak height hyperparam (lambda)** to 0.014 (i.e. the parameter to a shifted exponential density starting from the detection threshold 150 rfu) under *Advanced parameters*.
    1. Only small peaks up to around 400 rfu are probable.

13) Press **Continuous LR (Integrated Likelihood based)**.

- a. The software now integrates out the unknown parameters in each of the likelihood functions to make a marginalized calculates Likelihood Ratio weight-of-evidence which are independent of the unknown parameters.
  - i. Note:
    1. A flat prior is considered on all the unknown parameters in the continuous model (see vignette for more details).
    2. The integration depends on a relative error parameter which gives to accuracy of the integral. This is default 0.005 but can be changed under **Set relative error requirement** under *Integration* at the Toolbar.
- b. The user is now re-directed to the Database search GUI page which shows a sorted table of the references in the imported database.
  - i. The sorted table can be based on the continuous LR, qualitative LR, Number of Matching Alleles (MAC) or number of evaluated loci (nLocs).
- c. Go back to the Model specification GUI page (by clicking).

14) Press **Continuous LR (Maximum Likelihood based)**.

- a. Press the **Optimize model more** button (under *Further evaluation*) to be sure that the Likelihood function is optimized.
  - i. The optimized likelihood values are given under *Maximum Likelihood value*.

- b. Press **Model validation** under *Further action* for each of the hypothesis to test whether the fitted model is not adequate with the observed peak heights.
  - i. A Goodness-of-fit test reports a p-value for the test.
    - 1. This is printed out to terminal.
  - ii. A large p-value indicates that the fitted model fits the observed peak heights very well, while small p-values indicates that the fitted model does not fit the observed peak heights very well.
- c. Press **Search Database** under *Further evaluation* to do the database search.
- d. The user is now re-directed to the Database search GUI page which shows a sorted table of the references in the imported database.
  - i. The sorted table can be based on the continuous LR, qualitative LR, Number of Matching Alleles (MAC) or number of evaluated.