MANUAL FOR EUROFORMIX V4.0

Author: Øyvind Bleka <oyvble.at.hotmail.com> and Peter Gill

Date: 03-01-2023

Version used: 4.0.1

Data used "*Data set 0: Tutorialdata*" from www.euroformix.com/datasets

See webpage www.euroformix.com for details for installation and guides for how to create startup icon for EuroForMix on your desktop.

Mathematical details of methods used in EuroForMix:

[1] Ø. Bleka, G. Storvik, P. Gill; EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts; Forensic Sci. Int. Genet., 21:35-44, 2016.

[2] Ø. Bleka, P. Gill, L. Prieto; CaseSolver: An investigative open source expert system based on EuroForMix; Forensic Sci. Int. Genet., 41:83-92, 2019 (supplementary materials)

[3] Ø. Bleka, C.C. Benschop, G. Storvik, P. Gill; A comparative study of qualitative and quantitative models used to interpret complex STR DNA profiles.;Forensic Sci Int Genet.,25:85-96, 2016

[4] P. Gill, Ø. Bleka, O. Hansson, C.C. Benschop, H. Haned; Forensic Practitioner's Guide to the Interpretation of Complex DNA Profiles; Academic Press, 2020 (Chapter 7, 8, Appendix B).

[5] P. Gill, Ø. Bleka, O. Hansson, A.E. Fonneløp; Limitations of qPCR to estimate DNA quantity: An RFU method to facilitate inter-laboratory comparisons for activity level, and general applicability; Forensic Sci. Int. Genet., 61:102777, 2022

Table of Contents

1 Toolbar

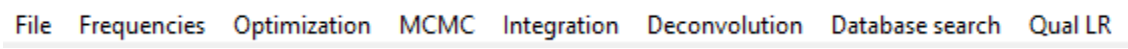


***Figure 1**: The Toolbar contains configurations and advanced model parameters for different kinds of analyses.*

1.1 File

1.1.1 Set directory

The user may select the working directory for the program.

1.1.2 Open project

The user may open an earlier project which is saved in a file in the form: “projectname.Rdata”.

1.1.3 Save project

The user may save the existing project into a file with name: “projectname”.

- Extension ‘.Rdata’ is added automatically to project name.
- All data imported to the program and resulting calculations are stored into a single project-file which may be opened at any time in the program.
- Large reference databases are stored efficiently (the required space for the database is drastically reduced).

It is strongly recommended to use “save project” because it saves a lot of time if you need to re-evaluate the analysis. All the data are conveniently stored and can be reloaded instantly.

1.1.4 Settings

The user can set calibration settings for the model: Detection threshold, drop-in model and prior distributions of the stutter parameters. The saved settings are restored after GUI is closed. The settings are also stored into the report.

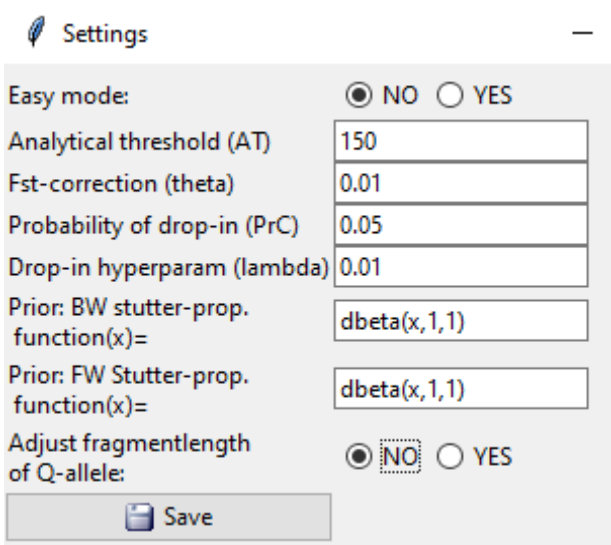


***Figure 2**: The setting window under "**File->Settings**" used to set advanced model parameters.*

- **Easy mode**: NO gives all functionalities. YES introduces a reporting mode for calculating the Likelihood Ratio (disables buttons to guide the reporter).

Easy mode is strongly recommended for routine casework because it guides the user to carry out interpretation using a recommended path.

The following predefined parameters (calibrated hyperparameters) can be specified here:

- **Analytical threshold (AT):** [1,->)
  - The analytical threshold is used to define whether an allele is present in the evidence or not.
  - If peak heights in the evidence are lower than the specified threshold, the corresponding alleles (and peak heights) below threshold **are** automatically **removed**.
  - Not considered if no peak heights are provided in the evidence (qualitative analysis).
- **Fst-correction (theta)**: [0,1] Assumed co-ancestry parameter (theta) assigned in the genotype probability for each contributor in the hypotheses. See references for more details.
- **Probability of drop-in (PrC):** [0,1]
  - Assumed probability of an allele drop-in to the evidence at a given locus. See references for more details.
  - If **PrC**>0 when considering 'Quantitative LR', the user needs to specify **Drop-in Hyperparam (lambda)**>0.
- **Drop-in hyperparam (lambda)**: (0,1]
  - Only used for 'Quantitative LR' if **PrC**>0.
  - Assumed hyper-parameter to model the peak height of the dropped in allele caused by a 'random allele drop-in' (**Figure 3**).

  A default of 0.01 is suggested if the user has no data (although it is recommended that the user calculates this parameter from his/her own data). The parameter is calculated automatically when the user imports a dataset with the "Fit dropin data" button.
- **Prior: BW/FW Stutter-prop. function(x)= dbeta(x,1,1)**:
  - A prior density function for the backward/forward stutter proportion parameter **BW/FW Stutter-prop.**
  - The user can design his own density function over [0,1], or define his own boundary.
  - Default is a flat prior (specified through the beta distribution).
- **Adjust fragment length of Q-allele**: Whether to use an alternative fragment length of Q-allele.
  - **NO:** The maximum defined fragment length for marker (as in earlier EFM versions)
  - **YES:** A weighted average of fragment length of non-observed allele frequencies[1]
- **Save**: Click to save settings permanently*: The values are stored and loaded in the next session. This means that you can close EuroForMix and open it again without the loss of values.

  **A loaded project restores settings used in the corresponding project.*

[1] Introduced to be consistent with method used by DNAStatistX (https://www.forensicinstitute.nl/research-and-innovation/international-projects/dnaxs)

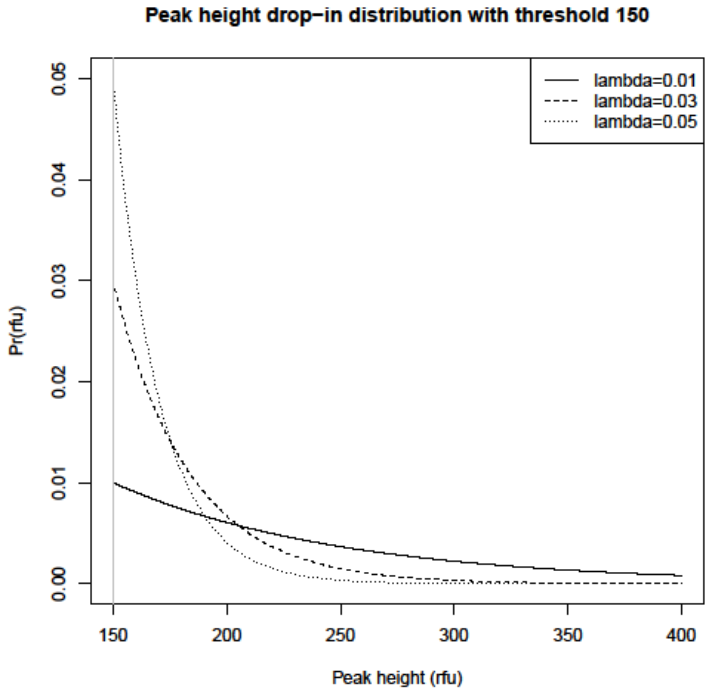


***Figure 3**: The figure shows the allele peak height drop-in distribution for three values of the lambda hyper-parameter. The distribution is (shifted) exponential beginning from defined from detection threshold. The higher the value the steeper the curve.*

1.1.5 Marker settings

The user can specify marker (or dye) specific settings for Detection threshold (AT), Dropin model, and Fst-correction. It requires that *allele frequencies* have been imported. Selecting a **kit** enables the 'Dye' specific information; If selected, the markers which overlaps in the selected kit and imported allele frequencies will be shown (**Figure 4**).

Marker specific settings

Restore | Set to default | Fill out dye info | Empty all | Save settings

Marker	Analyt. thresh (AT)	Dropin prob. (pC)	Hyperparam (lambda)	Fst-correction (theta)	Dye (color)
D3S1358	150	0.05	0.01	0.01	blue
TH01	150	0.05	0.01	0.01	blue
D21S11	150	0.05	0.01	0.01	blue
D18S51	150	0.05	0.01	0.01	blue
D10S1248	150	0.05	0.01	0.01	green
D1S1656	150	0.05	0.01	0.01	green
D2S1338	150	0.05	0.01	0.01	green
D16S539	150	0.05	0.01	0.01	green
D22S1045	150	0.05	0.01	0.01	yellow
VWA	150	0.05	0.01	0.01	yellow
D8S1179	150	0.05	0.01	0.01	yellow
FGA	150	0.05	0.01	0.01	yellow
D2S441	150	0.05	0.01	0.01	red
D12S391	150	0.05	0.01	0.01	red
D19S433	150	0.05	0.01	0.01	red
SE33	150	0.05	0.01	0.01	red

***Figure 4**: The 'Marker settings' window under "File->Marker Settings" used to set advanced model parameters.*

- **Restore**: Click to restore the values as obtained when the 'Marker settings' button was clicked.

- **Set to default:** Click to automatically insert the default settings as specified in 'Settings'.

- **Fill out dye info:** If a kit is selected, the Dye specific information will be shown: The user can then insert values for the 1st marker of a specific dye - and clicking this button will fill in these values for the remaining markers (of that specific dye). This is very useful in order to save time when manually filling in values.

- **Empty all:** Click to automatically empty the value in all cells.

- **Save settings:** Click button to save marker specific settings permanently*: The values are stored and loaded in the next session. This means that you can close EuroForMix and open it again without the loss of values.

  **A loaded project restores settings used in the corresponding project.*

1.1.6 Quit project

When button is pushed, the user is given a question about saving project before terminating the GUI.

1.2 Frequencies

1.2.1 Set size of frequency database

User may specify number of individuals 'N' used to create the population frequencies.

- When new alleles, i.e. not in the frequency database, from imported files are found, these are assigned as freq0.
  - If N=0 (this is default), freq0 is equal to the minimum imported allele frequency (see 1.2.2).
  - If N>0, freq0='5/(2N)'.
- New alleles are updated to the population frequency database:
  - When a reference database is imported
  - Frequencies will always be normalized when importing reference databases.
  - When interpretations are carried out ('Generate sample', Deconvolution, Weight-of-Evidence or 'Database search')
- The allele frequencies used for an analysis will be presented in the stored report (see 4.5.2).

1.2.2 Set minimum frequencies

The user can specify the allele frequency for new alleles. See details above.

Note: This option is not to be confused with a strict "minimum frequency rule" used for all alleles in the frequency file.

1.2.3 Set whether to normalize frequencies

The user can specify whether to normalize the allele frequencies after new alleles are included to the population frequency database. Value *1* means *Yes*, and value *0* means *No*. Default is ***Yes***.

Notes about capability of other software to normalize allele frequencies when new alleles are added:

- DNAStatistix: **Yes**
- EuroForMix v3: **Yes**
- EuroForMix v2: **No**
- LRmix Studio: **No**

- EuroForMix v1: **Yes**

1.2.4 Set number of wildcards in false positive match

The user may specify the number of 'wildcards' in the random match probability statistics, which are applied when the user has imported and selected an evidence profile together with the population frequencies (click View frequencies).

1.2.5 Set URL for STRidER import

The user can modify the URL address used when importing frequency data from the STRidER webpage.

1.3 Optimization

1.3.1 Set number of successful optimizations

The user may set required *accepted* number of equal* maximum optimizations to be obtained. At least two identical optimizations are recommended to reduce the risk that the maximum point is not global (i.e. to avoid risk of not obtaining the Maximum Likelihood Estimate **MLE)**. Default is **3**.

If the calculations are very comprehensive e.g. 4 or more contributors, the user may select value equal 1 to begin with. The reason to do this is because the calculations may be very time consuming (several hours), hence it is convenient to shorten the period as much as possible. Ideally a high-spec. computer should be used to facilitate the calculation speed. Once optimization has been completed, save project and repeat the optimization again ensure that the same results are obtained (if **seed of randomizer** *is* ***set****, be sure to select different seed for different optimizations).*

- An optimization is *accepted* if a valid maximum point has been obtained (i.e., the covariance matrix is positive definite yielding positive variance of the estimators).
- Separate optimizations **are no longer** carried out in parallel (parallelization) - hence each optimization is always carried out in serial.

*** A tolerance of |x-y|<0.01 is used when comparing the log-likelihood values. The value "0.01" can be adjusted in Optimization-> Set 'difference tolerance'**

1.3.2 Set variance of randomizer

The user may set the variance parameter used for the random generation of startpoints used in optimizer. Default is **1.**

Note: Previous versions used value 10 as default, which for this version (since version 3) would produce poor startpoints. If projects saved from earlier version are loaded, with value 10 used, the value will be automatically modified to 1.

1.3.3 Set difference tolerance

Change the log-likelihood value tolerance criterion for when two maximum log-likelihood values are the same (see the asterisk footnote in 1.3.1).

1.3.4 Set seed of randomizer

The user may set a seed number to make the optimization process reproducible. Default is no value (empty), which makes each optimization random (non-reproducible) since no seed is chosen.

Note: The report will indicate whether a seed was set for the optimization, and eventually its value.

1.3.5 Set accuracy of optimization

Adjusts the "steptol" argument used in the nlm R-optimizer function. Notes:

- Smaller value will give better numerical accuracy of the maximum likelihood value.
- Default is 1e-3 for faster return from optimizing (significant time difference to 1e-6 which is default in R).

- Too large values may cause convergence problems for situations when the magnitude of the log-likelihood values is small (typically with few data).

1.3.6 Set significance level of validation

The user may set the significance level for model validation. Default is 0.01.

1.3.7 Set maximum threads for computation

The user may set the maximum number of threads allowed to be executed in the C++ parallelization. Value equal 0 means that all threads will be utilized (no limitation). Default is 0.

1.4 MCMC (Markov Chain Monte Carlo)

1.4.1 Set number of samples

The user may set the number of MCMC samples drawn from the posterior distribution of the parameters: applied when clicking 'MCMC simulation'/'LR sensitivity'. Default is **x**=2000 samples*****. The minimum recommended number samples are shown to the user when clicking "LR sensitivity". The user can click "LR sensitivity" again to provide **x** additional samples to the MCMC chain. A trace plot is provided for the "LR sensitivity" module used for checking convergence.

*****It may be useful to carry out a small preliminary experiment with just a few samples in order to make an estimate of the time it will take to run a larger sample.

1.4.2 Set variance of randomizer

The user may set the variance parameter scalar used in the 'Markov Chain Monte Carlo (MCMC) random walk Metropolis'. Default is 2. This parameter is automatically further tuned* when applying "LR sensitivity" to ensure an optimizes** MCMC procedure.

*The tuning is based on 100 MCMC samples under Hp.

**The MCMC procedure is optimized if acceptance rate of the sampler is around 0.25 (tolerance is [0.15-0.35]).

1.4.3 Set quantile

The user may set the conservative quantile for the "conservative LR approach". The estimated quantile (and an approximative 95% CI) is the output from the "LR sensitivity".

1.4.4 Set seed of randomizer

The seed used for the 'MCMC simulation'/'LR sensitivity' can be changed. Default is 1, which gives reproducible results.

- For the "LR sensitivity", the seed for the Hd simulations is 'seed value'+999. The Hp and Hd chains must be randomly independent.
- The report will indicate the seed that was used to create the conservative LR estimate.

1.5 Integration

1.5.1 Set relative error requirement

The user may set the required estimated relative error used in the integration function adaptIntegrate (cubature). See references for more details. Default is 0.1. A relative error of LR is provided as an interval. The smaller the value the better the precision. If the relative error interval contains invalid numbers (such as NA), the user should increase the maximum number of evaluations (see 1.5.2).

1.5.2 Set maximum number of evaluations

The user may set the maximum number of evaluations for calculating the integral. This number will override the relative error requirement if selected greater than 0 (which means no limitations). Default is 20000. Notes:

- The user may need to increase the number (or set to zero) to avoid non-convergence of results (this is typically indicated with a relative error interval containing invalid numbers).
- When this value is greater than 0, the progress bar will be shown.

1.5.3 Deviation scale

The user may set a deviation **scale** number which helps define the boundary of the integral: **scale***2*SE apart from MLE, where SE and MLE is obtained from the MLE optimization. Default is 3. A too small **scale** may cause cutting off the posterior and lead to inaccurate results. Also, a too high scale may cause numerical issues leading to inaccurate estimates.

1.6 Deconvolution

1.6.1 Set required summed probability

The user may set the required summed posterior genotype-probability which the deconvolution lists must contain. Default is 0.99.

1.6.2 Set max listsize

The user may set the maximum number of genotypes shown in the console. Default is 20.

1.7 Database search

1.7.1 Set maximum view-elements

Used for very large reference databases: The user may set maximum number of individuals to show from the reference-database. Default is **n**=10000.

- The greater **n** is, the more time-consuming it will become to show the GUI table.
- Note that the results table from the database search shows only the top **n**-ranked elements.

1.7.2 Set drop-in probability for qualitative model

When searching a database with quantitative LR model, the qualitative LR model is also considered with a specific drop-in probability parameter given here. Default is 0.05.

1.7.3 Set number of non-contributors

The user may specify the number of random non-contributor samples in the non-contributor analysis. Default is 10.

- The user should take note of the time usage before increasing this number (in case of quantitative model).
- Assuming no theta-correction (fst=0) causes the Hd-evaluation to only be excecuted ones.

1.8 Qual LR

1.8.1 Set upper range for sensitivity

The user may specify the maximum allele dropout-probability in the sensitivity plot (for a qualitative model). Default is 0.6.

1.8.2 Set nticks for sensitivity

The user may specify number of grids of the allele dropout-probability in the sensitivity plot (for a qualitative model). Default is 31.

1.8.3 Set required samples in dropout distr.

The user may specify number of required allele drop-out probability samples used to estimate the quantiles or median for the distribution of the '*allele drop-out probability given number of observed alleles*'. Default is 2000.

1.8.4 Set significance level in dropout distr.

The user may specify the significance level in the conservative LR calculation (i.e. the quantile for the distribution of the '*allele drop-out probability given number of observed alleles*'). Default is 0.05.

2 Importing data

- Notes about all data files:
  - The extension (denotes file-type) of the file names does not matter. It may also have no extension at all.
  - All imported files must be either comma, semi-colon or tab-separated (‘,’,’;’,’\t’).
  - The program automatically converts the marker names to upper case letters (case insensitive).
  - A subset of the imported data is shown to R-console. From this, the user may check that the columns are correctly separated.

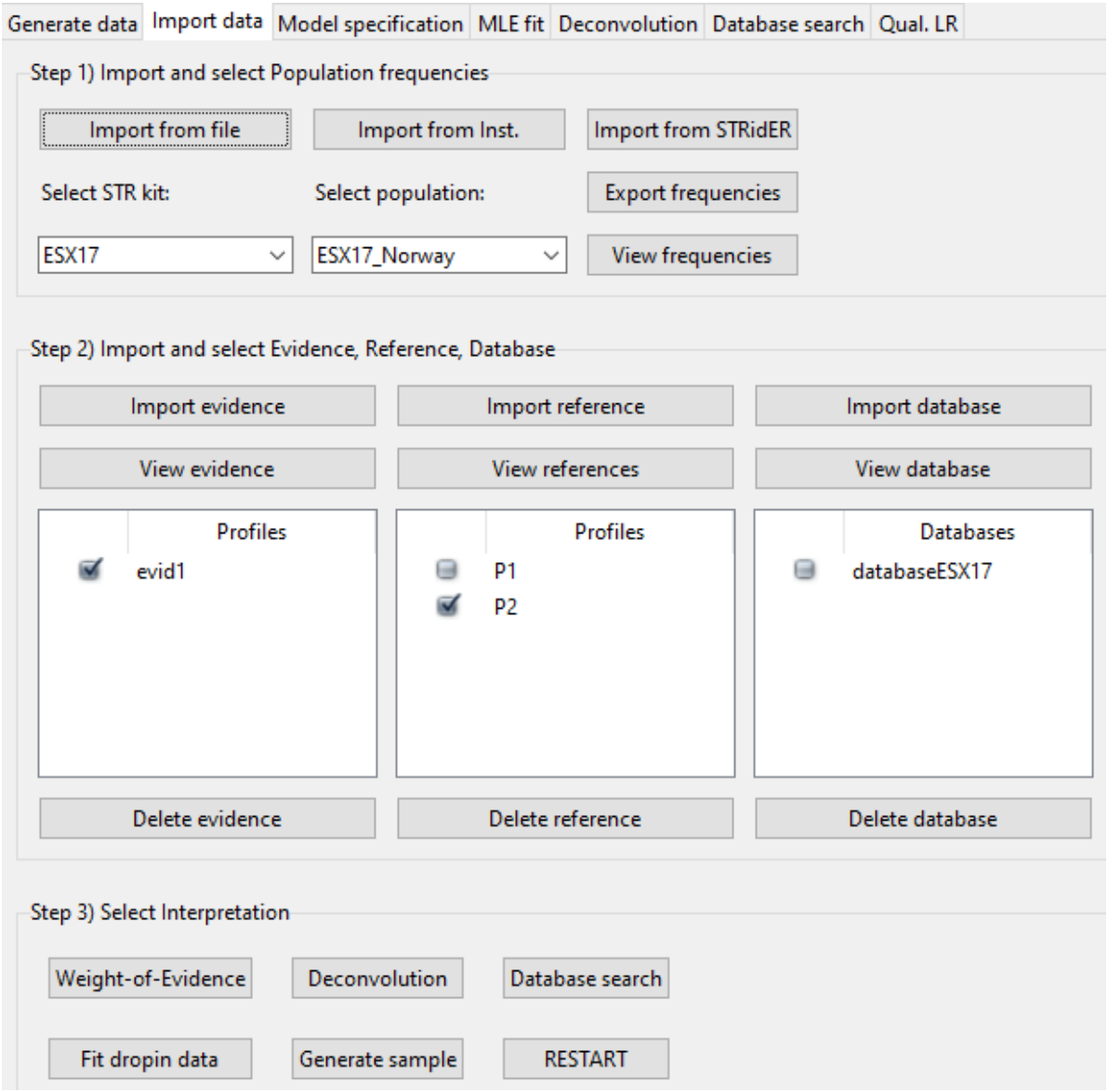


***Figure 5**: The figure shows the Import data page where the user can import population frequencies, evidence profiles, reference profiles and reference databases.*

2.1 Population frequency import

In EuroForMix, “frequencies” are defined as the **relative frequencies**, meaning that the values are assumed to be already normalized before input. However, raw counts can in practice be used if frequency normalization is conducted (see 1.2.3). Notice that another frequency file cannot be imported after a Reference database is imported (2.3.2).

2.1.1 Import from file

The user can select a frequency file from the local system. The name of the selected file will be present under “Select population”.

- Hover the button with the mouse to see the directory of example files (*FreqDatabases*) in the R installation location of euroformix.
- Required file format:
  - First column must contain allele-designations (header-name may be anything).
  - Other columns are frequency-information (header-name denotes the locus name and these will be converted to capital letters).
- Requirement for allele frequency values
  - The values in the file **cannot be zero** (instead keep the cells empty).
  - The values must sum up to one for each column: Gives a warning if this is not the case, see figure 6. The warning can be ignored by the user.

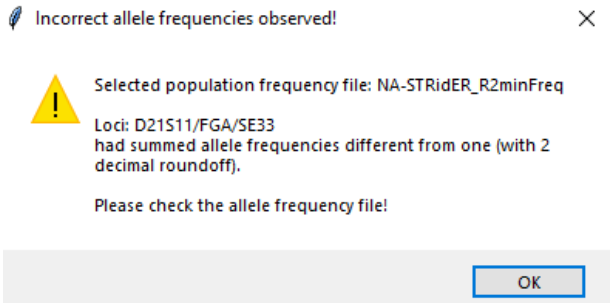


***Figure 6**: The figure shows the warning provided to the user if the frequencies does not sum to one.*

2.1.2 Import from Inst.

- Click to import population freq. data directly from installation folder.
  - Hover with mouse to obtain path of the folder.
  - This folder can **only** contain frequency-files.
- Filename: There is no longer any requirement for the name of the frequency-files.

2.1.3 Import from STRidER

- This will import frequency tables found at the STRidER webpage.
- Hover with mouse to see the URL of where data are imported from (can be modified, see 1.2.5).

2.2 Select kit and frequencies

2.2.1 Select STR kit

- The user can at any time select the relevant kit in the drop-down menu. Here the kits are given by its *short name*.
  - This will be the same names as obtained when running *getKit()* in the R-console (after loading *euroformix*).

- EuroForMix uses the kit info found in ~euroformix\extdata\kit.txt. (installation folder).
  - If the relevant kit is not found, the user can **contact help support** or try to create a new kit.txt file.

2.2.2 Select population (frequencies)

The user must select one of the populations in the drop-down menu. Then corresponding allele frequencies will be used in all analyses.

2.3 Profile import

2.3.1 Import Evidence/Reference profiles

See **Figure 6**

- **Multiple** evidence or reference profiles are **allowed** in each file.
- Required/optional headers (all are capital insensitive):
  - "**sample**" is required header for sample(s) name(s).
    - The sample names are NOT capital invariant.
    - If more than one header name contains "**sample**", it will select the header name which in addition contains "**name**" in the same string.
  - "**marker**" is required header for marker name(s).
    - Marker names are capital invariant.
    - If no header is found, the header containing "**loc**" will be used if found.
  - "**allele**" is required header(s) for allele-information.
    - This may be a vector ("alleleX1",…,"alleleX10") of any length denoting allele(s) to a given marker for a given sample. Here X can be anything.
    - References must have exactly two allele columns.
  - "**height**" optional header(s) for peak height-information.
    - This may be a vector ("heightX1", …,"heightX10") of any length denoting peak height to the corresponding allele(s) in "allele". Here X can be anything.

- In evidence files:
  - "height" header is required for analysis: 'Deconvolution', 'Weight-of-Evidence' (quantitative model) and 'Database search'. For 'Qualitative LR' this is not required.
  - Be sure that the number of alleles and corresponding peak heights are the same (error is not thrown).

- In reference files:
  - "height" header is optional but will not be used further in any analysis.
  - Homozygote genotype may have an empty allele under 'Allele 2'. The user will get a notification if this occurs.
  - Loci without any allele-information (i.e. empty or dropped out), will also be imported.

- For both evidence and reference:
  - A pop-up window is provided to the user if an "OL-allele" is detected: The user can choose to remove the situations with the OL-alleles. If the user chooses "No" the profile is not imported.

- MPS based data is supported (see section **9 Special for MPS data**). The alleles must be strings, and the column name for coverage/reads must still be "height".

Sample Name	Marker	Allele 1	Allele 2	Allele 3	Height 1	Height 2	Height 3
evid1	AMEL	X	Y	NA	2136	1015	NA
evid1	D3S1358	14	15	16.0	178	2405	1982
evid1	TH01	6	7	9.3	419	282	1871
evid1	D21S11	27	29	NA	1128	1750	NA
evid1	D18S51	15	17	NA	467	524	NA
evid1	D10S1248	13	14	15.0	1856	155	1045
evid1	D1S1656	12	15	16.0	1140	601	488
evid1	D2S1338	17	19	20.0	290	619	259
evid1	D16S539	9	10	11.0	217	312	743
evid1	D22S1045	15	16	NA	1017	610	NA
evid1	vWA	14	15	17.0	1250	440	1232
evid1	D8S1179	10	13	14.0	206	352	978
evid1	FGA	21	22	NA	664	714	NA
evid1	D2S441	9	10	11.0	200	3362	1168
evid1	D12S391	18	18.3	19.0	297	1446	751
evid1	D19S433	13	14	15.2	1157	781	922
evid1	SE33	29.2	30.2	33.2	221	473	570

SampleName	Marker	Allele1	Allele2
P1	D3S1358	16.0	15.0
P1	TH01	9.3	9.3
P1	D21S11	29.0	27.0
P1	D18S51	17.0	15.0
P1	D10S1248	15.0	13.0
P1	D1S1656	12.0	17.3
P1	D2S1338	23.0	19.0
P1	D16S539	11.0	12.0
P1	D22S1045	15.0	16.0
P1	VWA	14.0	17.0
P1	D8S1179	14.0	15.0
P1	FGA	22.0	21.0
P1	D2S441	10.0	14.0
P1	D12S391	18.3	22.0
P1	D19S433	13.0	15.2
P1	SE33	30.2	33.2

Figure 6*: The figure shows the table format for the imported evidence file (left) and reference file (right).*

2.3.2 Import Database

See Figure 7

- This is not to be confused with the allele frequency database (2.1).
- Exactly same format as reference files.
- Multiple database files may be imported (**must** be done one-at-a-time)
- **Requires** that population frequencies are imported and selected.
    - **WARNING**: Population frequencies may not be changed again after database importing!

Notes:

- Same samples within a database need to be in same block but markers within a sample can be in different orders.
- Some samples **may** have more/less markers than others (e.g. SGMplus profiles contra ESX17).
- **Missing markers** for a sample are given with NA.
- Only markers shared with selected population frequencies are imported.
- The imported database files may contain different markers.
- Homozygote genotype may have an empty allele under 'Allele 2'.

- Concerning large databases:
    - The database file may contain **any** number of individuals.
    - It is more time efficient to import several small databases rather than one large one.
    - Resource usage to import a database file with 17 markers:
        - 1e6 profiles takes about 131 seconds (requires ~1.3GB memory).
        - 5e6 profiles takes about 800 seconds (requires ~6.1GB memory).
    - Store the imported database efficiently by saving project file (See File under toolbar).

```
[1] "Raw fil import:"
                          Sample.Name   Marker Allele.1 Allele.2
1   00-JP0001-14_20142342311_NO-3241  D3S1358       14       15
2   00-JP0001-14_20142342311_NO-3241     TH01        7      9.3
3   00-JP0001-14_20142342311_NO-3241   D21S11       29       30
4   00-JP0001-14_20142342311_NO-3241   D18S51       13       17
5   00-JP0001-14_20142342311_NO-3241 D10S1248       12       13
6   00-JP0001-14_20142342311_NO-3241  D1S1656       11       14
7   00-JP0001-14_20142342311_NO-3241  D2S1338       17       19
8   00-JP0001-14_20142342311_NO-3241  D16S539       10       11
9   00-JP0001-14_20142342311_NO-3241 D22S1045       15       16
10  00-JP0001-14_20142342311_NO-3241      VWA       17       18
11  00-JP0001-14_20142342311_NO-3241  D8S1179       12       13
12  00-JP0001-14_20142342311_NO-3241      FGA       19       22
13  00-JP0001-14_20142342311_NO-3241   D2S441       11       10
14  00-JP0001-14_20142342311_NO-3241  D12S391       17       18
15  00-JP0001-14_20142342311_NO-3241  D19S433       13       14
16  00-JP0001-14_20142342311_NO-3241     SE33       15       21
17  00-JP0001-14_20142342311_NO-3241     AMEL        X        Y
18  00-JP0002-14_20142342311_NO-3242  D3S1358       15       18
19  00-JP0002-14_20142342311_NO-3242     TH01        6        9
20  00-JP0002-14_20142342311_NO-3242   D21S11       28     31.2
21  00-JP0002-14_20142342311_NO-3242   D18S51       13       18
22  00-JP0002-14_20142342311_NO-3242 D10S1248       13       13
23  00-JP0002-14_20142342311_NO-3242  D1S1656       15     18.3
24  00-JP0002-14_20142342311_NO-3242  D2S1338       25       25
25  00-JP0002-14_20142342311_NO-3242  D16S539       11       13
26  00-JP0002-14_20142342311_NO-3242 D22S1045       15       16
27  00-JP0002-14 20142342311 NO-3242      VWA       14       17
```

***Figure 7**: The figure shows the table format for the imported reference database file.*

2.2 View Data

2.2.1 View frequencies

See **Figure 8** for the Norwegian ESX17 population

- Creates a new window which shows the selected population frequencies (as imported) in a table.
- If any evidence profiles(s) are selected after evidence-import, the software makes an 'inclusion probability' plot for each of the selected profiles.
  - The plot (**Figure 9**) shows the exact probability that a random individual (from population) **('false positive probability'**) matching at least (2n-wildcardsize) up to 2n alleles with a **selected evidence** profile. Here **n** is number of considered loci (which are both in evidence and population frequencies) and wildcardsize is the number of allowed mismatches (default is wildcardsize =5).
  - wildcardsize can be changed under "Frequencies" in Toolbar by changing value **Set number of wildcards in false positive match.**

  - Notes:
    - Only the allele information in evidence-profile(s) is used.
    - New alleles which are not found in the selected population are assumed to have a **minimum allele frequency** (see under **Frequencies** in section **1 Toolbar**).
    - The 'Inclusion probability' is equivalent to the Random man not excluded (RMNE).

Population frequencies

Allele	D3S1358	TH01	D21S11	D18S51
5	-	0.00259844093543874	-	-
6	-	0.209274435338797	-	-
7	-	0.212472516490106	-	0.0008984
8	-	0.0836498101139316	-	-
8.2	-	-	-	-
9	-	0.140915450729562	-	0.0009983
9.3	-	0.344293423945633	-	-
10	0.000898652021967049	0.00589646212272636	-	0.0105820
11	0.00559161258112831	0.000899460323805717	-	0.0063891
11.3	-	-	-	-
12	-	-	-	0.1320754
13	0.00329505741387918	-	-	0.1278825
13.1	-	-	-	-
13.2	-	-	-	-
14	0.124113829256116	-	-	0.1812918

***Figure 8**: The figure shows the viewed frequencies for the Norwegian ESX17 population.*

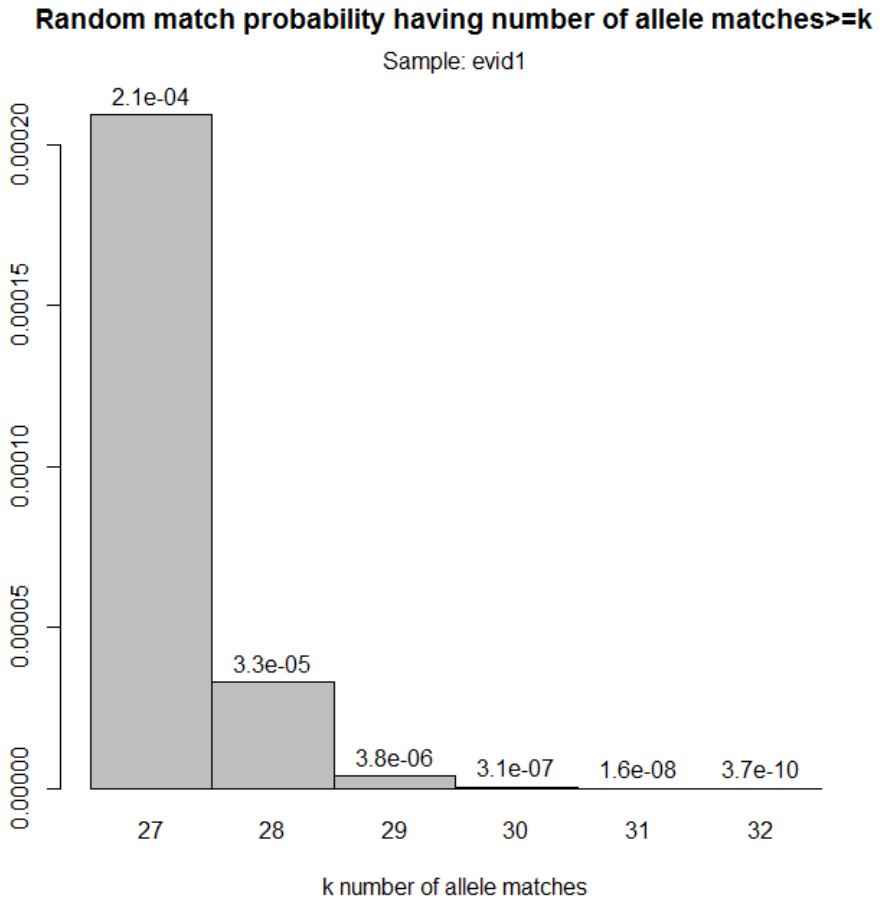


***Figure 9**: The figure shows the random probability of a match with at least k number of alleles (from a randomly chosen reference profile) compared with the observed alleles in evidence profile (wildcardsize=5).*

```
[1] "Samplename: evid1"
         Allele                Height
AMEL     "X/Y"                 "2136/1015"
D3S1358  "14/15/16"            "178/2405/1982"
TH01     "6/7/9.3"             "419/282/1871"
D21S11   "27/29"               "1128/1750"
D18S51   "15/17"               "467/524"
D10S1248 "13/14/15"            "1856/155/1045"
D1S1656  "12/15/16/16.3/17.3" "1140/601/488/155/1877"
D2S1338  "17/19/20/23"         "290/619/259/649"
D16S539  "9/10/11/12"          "217/312/743/619"
D22S1045 "15/16"               "1017/610"
VWA      "14/15/17"            "1250/440/1232"
D8S1179  "10/13/14/15"         "206/352/978/827"
FGA      "21/22"               "664/714"
D2S441   "9/10/11/14"          "200/3362/1168/3693"
D12S391  "18/18.3/19/21/22"    "297/1446/751/171/1370"
D19S433  "13/14/15.2"          "1157/781/922"
SE33     "29.2/30.2/33.2"      "221/473/570"
```

***Figure 10**: The figure shows the printed alleles and heights in the imported evidence.*

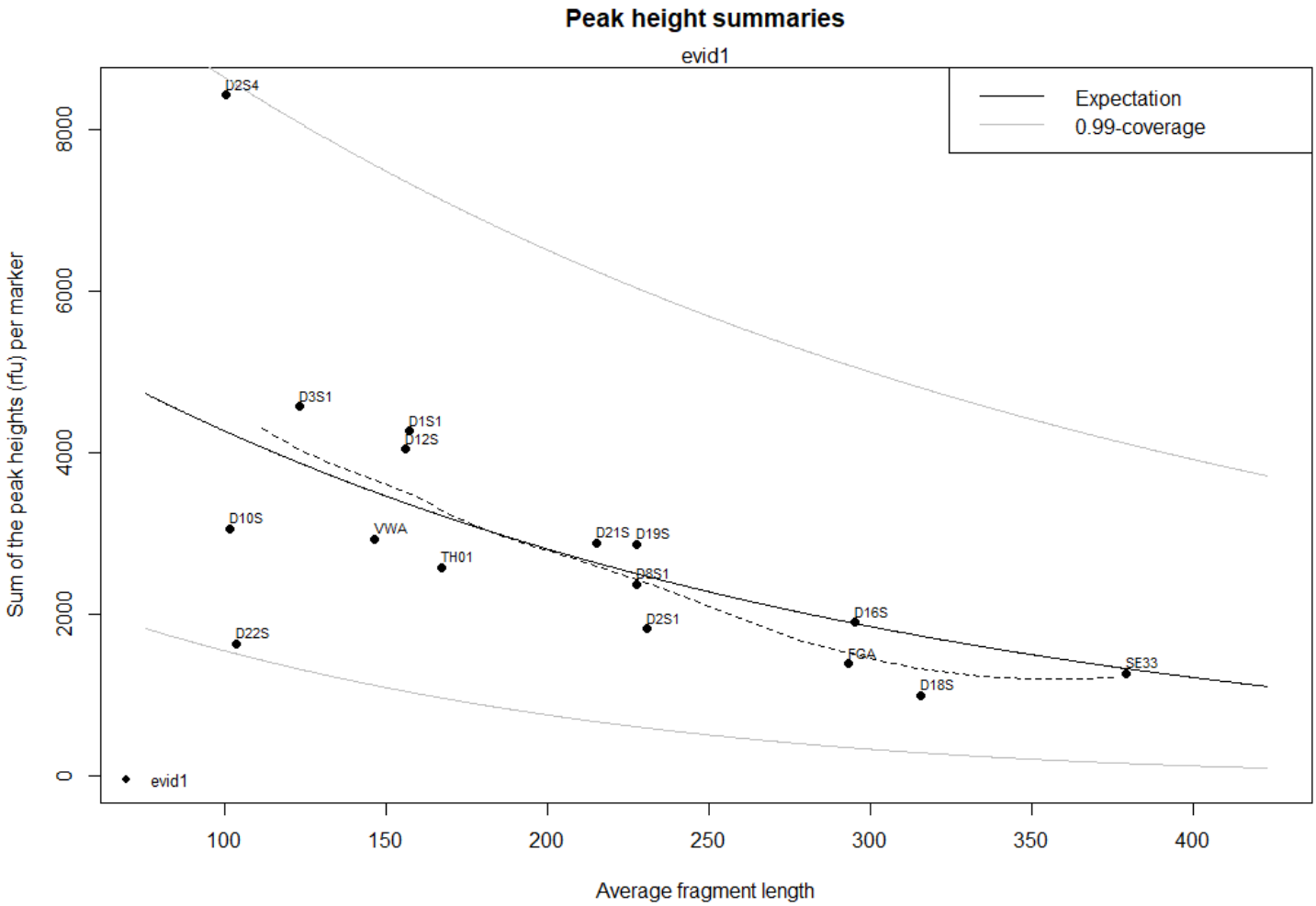


***Figure 11:** The figure shows a regression model with peak height sum per marker as the response (y), fitted against the 'average fragment length of observed alleles' (x). The model gamma(y,shape=2/sigma^2*beta^((x-125)/100),scale=mu*sigma^2) is fitted using maximum likelihood estimates for mu (P.H.expectation),sigma (P.H.variability) and beta (Degradation-slope.). The solid line is the expectation with corresponding 0.005- and 0.995-quantiles of the distribution. The dashed line is a smoothed curve of the points (loess function).*

2.2.2 View evidence

- Prints imported loci, along with alleles (and peak heights if any) for each selected evidence profile(s) (**Figure 10**).
- Plots EPG and degradation plot (**Figure 11**) for each selected evidence profile(s)
    - From version 2.3.0, EPG plots are also created in default browser using *plotly*.
    - The kit selected under '**Select STR kit**' denotes the EPG format.
    - Loci in evidence which are **inconsistent** with the ones in selected kit (or missing) are **not shown** in the EPG.
    - If reference profiles are imported and selected, they will be labeled together with the peak heights in the EPG plot.
    - The degradation plot shows points and a fitted regression line using sum peak heights at each marker (for the average fragment length). If multiple samples are provided, these will be shown in different colors for easy comparison.

2.2.3 View reference

- Prints imported genotypes for each selected reference profile(s) (**Figure 12**).
- If allele frequencies are imported and selected, the software will calculate the random match probability and its inverse, for each selected reference profile(s) (**Figure 13**).
    - Note: The corresponding selected fst value is used in the calculation.
- If any evidence profiles(s) are selected after evidence-import, the software counts number of matching alleles (MAC) for each locus of the selected reference profiles, for each selected evidence (**Figure 14**).
    - MAC = number of alleles for the reference which are included in the evidence.
    - nLocs = number of considered loci when counting MAC.
    - MatchRate = MAC/(2*nLocs), the proportion of matches.
    - nMissing = 2*nLocs – MAC, the number of allele-mismatches.

```
            P1          P2
"D3S1358"   "16/15"     "16/15"
"TH01"      "9.3/9.3"   "6/7"
"D21S11"    "29/27"     "29/35"
"D18S51"    "17/15"     "11/14"
"D10S1248"  "15/13"     "13/13"
"D1S1656"   "12/17.3"   "15/16"
"D2S1338"   "23/19"     "17/20"
"D16S539"   "11/12"     "9/10"
"D22S1045"  "15/16"     "15/15"
"VWA"       "14/17"     "15/17"
"D8S1179"   "14/15"     "10/13"
"FGA"       "22/21"     "22/25"
"D2S441"    "10/14"     "11/11"
"D12S391"   "18.3/22"   "18/19"
"D19S433"   "13/15.2"   "14/14"
"SE33"      "30.2/33.2" "27.2/29.2"
```

***Figure 12**: The figure shows the printed alleles of the imported reference profiles.*

```
[1] "Calculation of random match probability and its inverse for fst=0.01"
                          P1        P2
RMP          "RMP"        "1.3e-22" "0"
log10(1/RMP) "log10(1/RMP)" "21.9"    "Inf"
```

***Figure 13**: The figure shows the printed random match probabilities (and log10 of inverse) for each reference.*

```
            P1   P2
"AMEL"      NA   NA
"D3S1358"   "2"  "2"
"TH01"      "2"  "2"
"D21S11"    "2"  "1"
"D18S51"    "2"  "0"
"D10S1248"  "2"  "2"
"D1S1656"   "2"  "2"
"D2S1338"   "2"  "2"
"D16S539"   "2"  "2"
"D22S1045"  "2"  "2"
"VWA"       "2"  "2"
"D8S1179"   "2"  "2"
"FGA"       "2"  "1"
"D2S441"    "2"  "2"
"D12S391"   "2"  "2"
"D19S433"   "2"  "2"
"SE33"      "2"  "1"
"MAC"       "32" "27"
"nLocs"     "16" "16"
"MatchRate" "1"  "0.84"
"nMissing"  "0"  "5"
```

***Figure 14**: The figure shows number of matching alleles and total (MAC) between the imported references and selected evidence profile.*

2.2.4 View database

See **Figure 15** for selected database

- Creates a new window (for each selected database) which shows the genotypes for every reference in the database.
  - "-" means that the genotype of a reference was missing.
- If any evidence profiles(s) are selected after evidence-import, the software counts the number of matching alleles (MAC) for all references in the database against each of the selected evidence (**Figure 16**). The results are shown in a MAC-ranked table in a new window (for each selected database).
  - **MAC** is the total number of alleles for the reference which are included in the evidence.
  - **nMarkers** is the number of reference-loci which has been used to evaluate the MAC.
- Notes:
  - Max number of individuals to view in a database can be changed with selecting **Set maximum view-elements** under "Database search" in toolbar.
  - Only overlapping loci with the selected kit will be shown and used in further calculations.

References in imported database databaseESX17

Reference	D3S1358	TH01	D21S11	D18S51	D10S1248	D1S1656	D2S1338	D16S539	D22S1045
00-JP0001-14_20142342311_NO-3241	14/15	7/9.3	29/30	13/17	12/13	11/14	17/19	10/11	15/16
00-JP0002-14_20142342311_NO-3242	15/18	6/9	28/31.2	13/18	13/13	15/18.3	25/25	11/13	15/16
00-JP0003-14_20142342311_NO-3243	16/18	9.3/9.3	30/30	13/18	14/16	13/16	17/18	8/12	15/16
00-JP0004-14_20142342311_NO-3244	18/18	7/9.3	29/32.2	12/22	15/16	12/15	19/23	11/11	11/16
00-JP0005-14_20142342311_NO-3245	15/17	7/8	28/33.2	12/17	13/15	16/17.3	19/25	13/13	11/17
00-JP0006-14_20142342311_NO-3246	14/18	7/9.3	28/32.2	11/15	15/16	14/15.3	20/24	9/13	16/16
00-JP0007-14_20142342311_NO-3247	15/19	9.3/9.3	30/32	14/19	13/15	17.3/17.3	17/23	9/10	14/16
00-JP0008-14_20142342311_NO-3248	14/16	9/9.3	30/30.2	14/18	14/16	15.3/16.3	17/23	9/11	11/16
00-JP0009-14_20142342311_NO-3249	14/16	7/7	30/30	12/16	14/14	11/14	21/22	12/12	15/15
00-JP00010-14_20142342311_NO-32410	15/16	6/6	30/32	16/17	13/16	16/18.3	21/23	9/14	14/15
00-JP00011-14_20142342311_NO-32411	15/17	6/9	29/30	15/16	13/16	16/17	17/25	12/12	15/17
00-JP00012-14_20142342311_NO-32412	15/17	7/9.3	30/31.2	14/19	13/14	12/16	19/20	10/12	15/15
00-JP00013-14_20142342311_NO-32413	17/18	6/9	28/29	12/19	13/14	15/16.3	17/24	11/13	15/17
00-JP00014-14_20142342311_NO-32414	15/18	9/9.3	29/30	13/18	13/17	16/17.3	18/24	9/13	15/16
00-JP00015-14_20142342311_NO-32415	16/16	8/9.3	30/30	12/15	14/14	13/17.3	17/24	9/11	13/16
00-JP00016-14_20142342311_NO-32416	14/15	6/9.3	28/31	15/17	13/16	16/18.3	23/25	11/12	16/18
00-JP00017-14_20142342311_NO-32417	17/18	6/7	29/33.2	13/14	13/15	13/18.3	19/19	13/13	15/16

***Figure 15**: The figure shows the viewed references from the imported ESX17 database*

Number of sample matching alleles in ref...

Reference	evid1	nMarkers
00-JP00057-14_20142342311_NO-32457	25	16
00-JP00059-14_20142342311_NO-32459	24	16
00-JP00025-14_20142342311_NO-32425	23	16
00-JP00036-14_20142342311_NO-32436	23	16
00-JP00041-14_20142342311_NO-32441	22	16
00-JP00044-14_20142342311_NO-32444	22	16
00-JP00056-14_20142342311_NO-32456	22	16
00-JP0001-14_20142342311_NO-3241	21	16
00-JP00018-14_20142342311_NO-32418	21	16
00-JP00019-14_20142342311_NO-32419	21	16
00-JP00031-14_20142342311_NO-32431	21	16
00-JP00042-14_20142342311_NO-32442	21	16

***Figure 16**: The figure shows the sorted references (in the reference database) with respect to MAC (total number of matching alleles) compared to the selected evidence.*

2.2.5 Delete Data

Clicking Delete evidence/reference/database will remove the selected corresponding (imported) profile(s).

2.3 Interpretation

2.3.1 Weight-of-Evidence:

Weight-of-Evidence is carried out by calculating the Likelihood Ratio (LR) for the specified hypotheses Hp (prosecution) and Hd (defense) using the quantitative (or qualitative) model. The following modules are available:

1) 'Quantitative LR' (Maximum Likelihood based)
   - Optimizes (maximum) the model parameters in the continuous model.
2) 'Optimal quantitative LR' (automatic model search)
   - Performs automatic model selection and returns the optimal '1) 'Quantitative LR (ML based)' result by traversing several models (specified by user).

3) 'Qualitative LR' (semi-continuous) – Mirrors the LRmix module in addition to the maximum likelihood method.

*Note that 'Easy Mode' guides the user to use module 1) '*Quantitative *LR'.*

- Requirements:
  - Imported population frequencies, **at least one** evidence profile and **at least one** reference profile (e.g. a suspect) to weight evidence for. Additional reference profiles are optional to condition on in the hypotheses.
  - 'Quantitative LR' calculations require that evidence(s) includes peak heights, 'Qualitative LR' calculations only requires allele data (peak height information is ignored).
- Features:
  - The quantitative model handles replicates, allele drop-in, allele drop-out, fst-correction, degradation and backward/forward-stutters.
  - The qualitative model handles replicates, allele drop-in, allele drop-out (equal across contributors) and fst-correction.

2.3.2 Deconvolution

- Deconvolution performs ranking of the genotype profiles of unknown contributors under a **specific hypothesis**
  - Note: The relationship module can be used for deconvolution.
- The ranking is based on the posterior genotype probabilities conditioned upon based on maximum likelihood estimates (fitted quantitative model).
- Requires: Imported population frequencies and selection of at least one evidence profile with peak height information. References are optional to condition on in the hypothesis.
- Feature: Model may handle replicates, allele drop-in, allele drop-out, fst-correction, degradation and backward/forward-stutters.

2.3.3 Database search:

- Carries out 'weight-of-evidence' tests by comparing the Likelihood Ratio (LR) between the specified hypotheses Hj (reference j in database) and Hd (defense) using the quantitative model as given in the references.
- Modules:
  - 1) 'Quantitative LR' (Maximum Likelihood based)
  - 2) 'Qualitative LR' (Semi-continuous)

- The quantitative LR value is shown together with qualitative LR (fixed with dropout probability 0.1) and MAC.
- Requires: Imported population frequencies, **at least one** evidence profile with **peak height** information and **at least one** reference-database. Reference profiles are optional to condition on in the hypotheses.

- Feature: Model may handle replicates, allele drop-in, allele drop-out, fst-correction, degradation and backward/forward-stutters.

2.3.4 Fit drop-in data:

- When clicking the button the user must select a text-file which contains drop-in peak heights in a separated format (“;”,”,”,”\t”,”\n”).
- This will fit the **lambda** parameter for the shifted exponential drop-in function, and the value of the lambda parameter in the “Settings” will automatically be updated accordingly.
- A histogram of the peak heights will be present in a plot together with the fitted model.
- **WARNING**: Remember to specify the analytical threshold (AT) in “Settings” before doing this, since the estimated lambda depends on this.

2.3.5 Generate sample:

- Generates alleles using the population frequencies and draws peak heights for a specified hypothesis using the quantitative model.
- Requires: Imported population frequencies.
- Feature: All the parameters in the quantitative model.
- Notes:
    - The relationship module is ignored.
    - Only one replicate at a time can be generated

2.3.6 Restart

Simply restarts the program.

3 Model specification

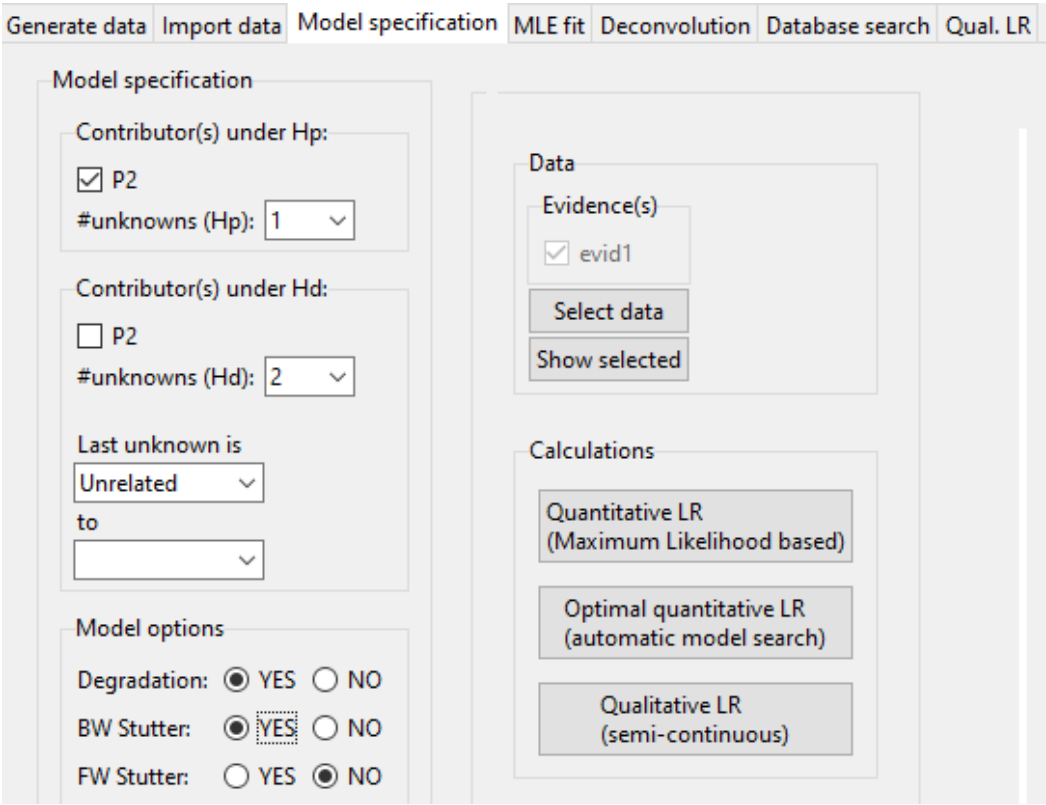


*Figure 17: The figure shows the Model Specification page for **Weight-of-Evidence***

3.1 Model specification

The model specification tab is invoked from several different routes. From the 'Import data' tab the options that can be followed are the buttons: **Generate sample**, **Weight of evidence**, **Database search** and **Deconvolution**. The effect and properties of each case are as follows:

3.1.1 Contributors under Hp

- Case: **Weight-of-Evidence** or **Database search:**
  - User may condition on selected references (from 'Import data') in the hypothesis Hp.
  - #unknowns (Hp): Denotes number of unknown contributors under the prosecution hypothesis Hp. Can be manually edited or selected from drop-down menu.
- Case: **Database search:** The individual in the reference-database is already included in the hypothesis Hp.
- Case: **Deconvolution** or **Generate sample**: This block is not considered, since Deconvolution only considers the model under Hd, and sample generation is carried out only under a specific hypothesis.

3.1.2 Contributors under Hd

The Hd hypothesis specification is similar for **all cases (Weight-of-Evidence, Database search, Deconvolution, Generate sample**).

- The user may condition on selected references (from 'Import data') in the hypothesis Hd.
- #unknowns (Hd): Denotes number of unknown contributors under the defense hypothesis Hd. Can be manually edited or selected from drop-down menu.

- **Relationship module** (included from version 2):
  - The user can specify the relationship between the last unknown contributor (only under Hd) to an imported reference sample.
  - Supported relationships:
    - Unrelated (this is default).
    - Parent/Child, Sibling, Uncle/Nephew, Grandparent/Grandchild, Half-sibling, Cousin.
  - The relationship model will be used for the following features:
    - The quantitative LR calculations (also for database searching).
    - Non-contributor tests: The non-contributors will be **random unknown individuals**, possibly related to an imported reference, as specified under Hd.
    - LR sensitivity, MCMC simulations, Deconvolution, Model validation, Model fitted P.H.
  - The relationship model is not (yet) implemented for the following features:
    - Qualitative LR and its non-contributor analysis.
    - The "Generate data" module.
  - Note:
    - If a relationship is specified, the user must select which of the imported reference profile that the unknown individual is related to.
    - The specified relationship will be indicated in the "report".
    - Theta/fst correction is taken into account for this module.
    - If the related reference profile contains any empty markers, these will be inserted with an unrelated unknown as substitute (this is also done for unrelated situations).

- Case: **Weight-of-Evidence** or **Database search**: References which are conditioned under Hp but not under Hd, will be assumed to be **known non-contributors** under Hd (this is relevant when fst>0).

3.1.3 Model options

- **Degradation**: Boolean (yes/no) incorporation of a global degradation model where the slope is determined with parameter **Degrad.slope** (beta):
  - The expected peak height of a specific allele is modelled to be proportional to beta^{(f-125)/100}, where f is the fragment length of the corresponding allele.
- **Stutters (backward/forward):** Boolean (yes/no) incorporation of a model for (n-1) backward-stutters or (n+1) forward-stutters by including additional parameter **BW stutter-prop.** or **FW stutter-prop.**:
  - **BW stutter-prop.** is a parameter which denotes the expected fraction of the contribution at allele *a* to allele *a-1*. See references for more details about the backward stutter model.
  - **FW stutter-prop.** is a parameter which denotes the expected fraction of the contribution at allele *a* to allele *a+1*.

3.2 Data

3.2.1 Select data

- The user may select or unselect loci for each selected evidence(s) and reference(s) from "Import data".
- Click "Confirm" to finish selection.
- If a locus is missing or has been unselected for an evidence, the locus for that evidence will not be evaluated at all.

- If a locus is missing or has been unselected for a reference, the corresponding locus is substituted with an unknown.

3.2.2 Show selected

Prints the following to the R-console: The selected evidence sample(s), reference(s) and considered population frequencies which are used for further analysis.

3.2.3 Note about missing data

- Missing markers in evidence profiles (possibly present in references), will not be evaluated (this deactivates the selection ticks). However, EuroForMix handles markers that have fully dropped out in the evidence profile – **these markers should still be included in the evidence sample file (but with no alleles)**.
- Missing markers in the reference profiles (possibly present in references) will be evaluated (this deactivates the selection ticks). For such markers the program will substitute the reference with an unknown contributor.

3.2.4 Note about new alleles

If alleles that do not exist in the population allele frequency table occur in the imported evidence, the new alleles are assigned with allele frequency *freq0*. *freq0* can be specified in several ways:

1) *freq0* is equal to the minimum observed allele frequency in the population table if N=0, or *freq0*=5/(2N) otherwise where N is number of individuals used to create the imported frequency database. This can be changed manually under "Frequencies->**Set size of frequency database**" in Toolbar.

2) *freq0* is the frequency set by the user in "Frequencies -> **Set minimum frequency**".

WARNING: The population frequencies **are** by default normalized after adding new allele frequencies to the population frequencies. This can be changed under "Frequencies->**Set whether to normalize frequencies**" in Toolbar.

3.3 Calculations

3.3.1 'Quantitative LR (Maximum Likelihood based) '

Under case **Weight-of-Evidence**, **Deconvolution** and **'Database search'**

Maximizes the likelihood with respect to the unknown parameters in the quantitative model for the specified hypothesis Hd (and Hp in case of Weight-of-Evidence/'Database search').

- The optimizer should return a global maximum. However, it may sometimes just return a local maximum, by chance. Number of successfully obtained maximums should be sufficiently large to ensure that the optimizer has found the global maximum of the Likelihood function. This can be changed under "Optimization->Set number of successful optimizations" in Toolbar.
- After calculation, the page 'MLE fit' is visited to present results.

3.3.2 'Optimal quantitative LR (automatic model search) '

Under case **Weight-of-Evidence**. This feature is new from version 3, and only works if Hp contains a conditional reference profile (POI) which is replaced by an unknown under Hd.

- The user can specify an outcome of several models to compute the maximum likelihood for:
  - The outcome of the number of contributors.
  - Whether Degradation model should be applied (YES/NO).

        - Note: The user must turn on Degradation under “Model options” under “Model specification” in order to be allowed to select YES.
    - Whether any of the Stutter models (BW or FW) should be applied (YES/NO). Note that version 4 does not accept BW=NO when FW=YES, hence these situations are ignored.
    - Click **Evaluate** to proceed: The user must confirm the outcome that will be evaluated (**Figure 18**).

- An information table for each model outcome is shown and this contains the following (**Figure 19**):
    - **NOC**: The number of contributors (NOC), assumed equal under Hp/Hd.
    - **Boolean** (TRUE/FALSE) of whether DEG/BWstutt/FWstutt models are applied.
    - **logLik**: the optimized logLik under the Hd hypothesis.
    - **adjLogLik**: The logLik value minus number of ‘effective’ parameters (equivalent to AIC).
    - **log10LR**: This is the LR of the Person of Interest (POI).
    - **MxPOI**: The estimated Mixture proportion for POI.
    - **SignifHp/SignifHd**: The number of failed model validations (i.e., number of points falling outside the red envelope) under Hp/Hd based on the Bonferroni corrected 1% significant level (see **Model validation**).
    - Note: The table is stored when saving a report (using ‘Create report’).

- The optimal model will be automatically selected and shown in the **MLE fit** panel.
    - The **optimal model** is the one with largest adjLogLik - which is equivalent of using the Akaike information criterion (AIC). See references for details.
    -

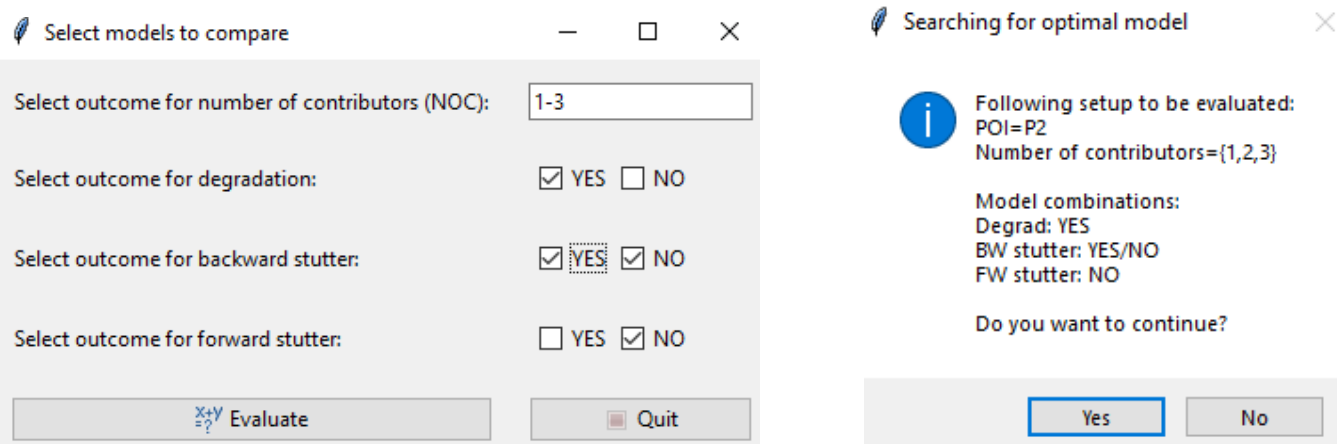


***Figure 18**: The user can select a model outcome for the analysis (left figure). The user must confirm the outcome to be evaluated (right figure).*

Model comparison results

NOC	DEG	BWstutt	FWstutt	logLik	adjLogLil	log10LR	MxPOI	SignifHp	SignifHd
1	TRUE	FALSE	FALSE	-538.38	-541.38	-74.26	1	29	13
1	TRUE	TRUE	FALSE	-500.65	-504.65	-66.48	1	29	10
2	TRUE	FALSE	FALSE	-475.76	-479.76	8.42	0.24	0	0
2	TRUE	TRUE	FALSE	-462.13	-467.13	10.05	0.23	0	0
3	TRUE	FALSE	FALSE	-468.26	-473.26	8.3	0.22	1	0
3	TRUE	TRUE	FALSE	-461.99	-467.99	9.99	0.23	0	0

***Figure 19**: The table shows an overview of results from fitting different models to the specified hypotheses: The number of contributors (NOC), whether DEG/BWstutt/FWstutt models are applied, the optimized logLik under Hd, and the #parameter-adjusted logLik, LR and estimated Mixture proportion of Person of Interest (POI), and number of failed model validations (points falling outside the red envelope) under Hp (SignifHp) and Hd (SignifHd).*

3.3.3 'Qualitative LR (semi-continuous)'

Under case **Weight-of-Evidence**

- Performs a semi-continuous procedure (mirrors the LRmix module) where the distribution of the 'allele drop-out probability given the number of observed alleles' are utilized to infer a "conservative" LR.
    - The model is purely qualitative which means that it is only based on allele-designation information.
    - It is also possible to obtain the LR based on the maximum likelihood method.
- Goes directly to the Qual. LR panel.

3.3.4 'Generate sample'

Under case **'Generate sample'**

- Push **'Generate sample'** button under the 'Import data' tab – this opens the Model specification tab.
- A dataset (an evidence sample and the contributing references) will be randomly simulated under the specified model under "Model specification" (relationship is not yet implemented).
- Reference profiles may be imported and selected as assumed known in the hypothesis.
- The (possibly marker specific) settings for Detection threshold, probability of drop-in and drop-in peak height hyper-param. lambda will be used in the simulation (fst-correction is not used).
- The unknown contributor profiles under the hypothesis will be randomly generated using the selected population frequencies (the relationship module is not used).
- The simulated peak heights of the generated evidence sample are entirely based on the quantitative model for assumed values of the model-parameters.
- Once the hypothesis of the model is specified, push button 'Generate sample' in the 'model specification' tab. The output goes directly to panel Generate data. Turn to section **8 Generate data** for a full description of this page.
- Click '**Plot EGP'** to show generated evidence profiles together with reference profiles in an EPG plot (if kit selected).
- Notice: This module does not include fst-correction or relationship module.

4 MLE fit: (Maximum Likelihood based)

This panel concerns the optimization of the likelihood function of the quantitative model (maximum likelihood).

Important notifications:

- The user can follow the progress of the optimization by following the progress bar in the R-console (note that this is only shown if the optimization if the Upper time expectation is at least 10 seconds).
- The user may experience difficulties in fitting the model if the model cannot explain the data. Examples:
  - Degradation model is applied when data does not show degradation.
  - Stutter model is applied when data does not show stutters (pre-filtered).
  - The solution is to turn off the corresponding model option which caused the problem.

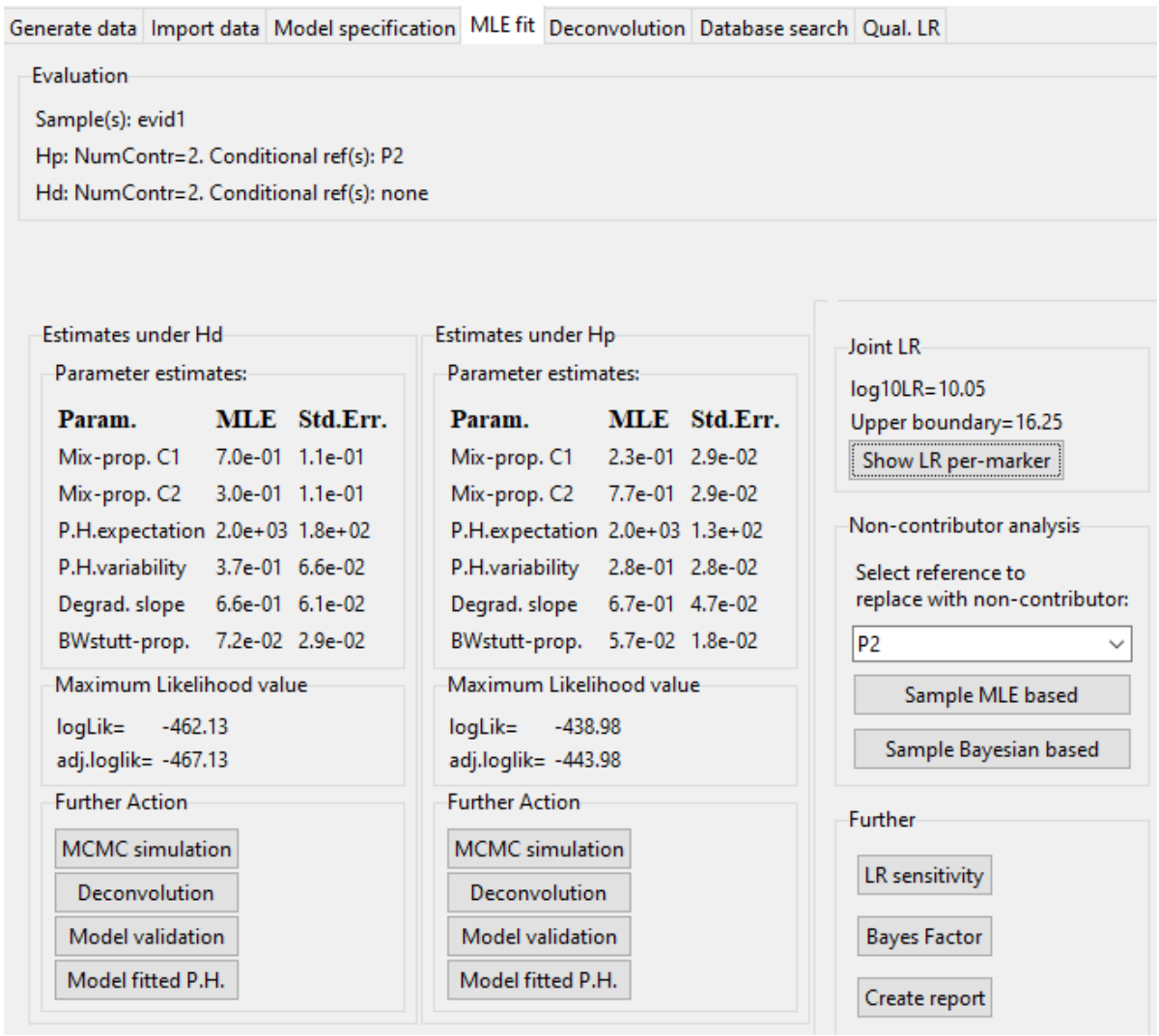


***Figure 20**: The figure shows the MLE-fit page after running the Quantitative LR (Maximum Likelihood based) calculation (maximizing the quantitative model with respect to the unknown parameters for each of the specified hypothesis in Figure 17) for Weight-of-Evidence.*

4.1 Evaluation

Lists the name of the evaluating evidence profiles and the defined hypotheses (including conditional references and number of contributors).

4.2 Estimates under Hd (and Hp for case: Weight-of-Evidence)

4.2.1 Parameter estimates

- **Param**: The unknown parameters in the model (see references for more details).
  - Mix-prop. C**k** (Mx**k**): Mixture-proportion for contributor '**k**'.
    - **The user can hover with mouse to obtain more information about the contributor names.**
    - **For the hovered information, RFU*Mx is given, where RFU is the average RFU across the evaluating markers (see [5]).**
  - P.H.expectation: Expectation of the peak height for a single heterozygote (Mix-prop=1) allele without degradation (at 125 bp or beta=1).
  - P.H.variability: Coefficient of variation of the peak height for a single heterozygote (Mix-prop=1) allele without degradation (at 125 bp or beta=1).
  - Degrad.slope (beta): A parameter related to the degree of decaying degradation global for all contributors.
    - From version 3.0.0, this value can never exceed one.
  - BWstutt-prop: A global parameter related to backward stutter. The expected proportion of peak height that are stutter.
  - BWstutt-prop: A global parameter related to forward stutter. The expected proportion of peak height that are stutter.
- **MLE**: The optimized[2] parameters in the model which attain a maximum point of the likelihood function.
- **Std.Err.:** The standard error of the parameter estimates in the model. These are based on the hessian matrix returned from the optimizer R-function *nlm*.

4.2.2 Maximum Likelihood value

- logLik: The logged (natural logarithm) value of the Likelihood value attained from the optimization[1].
- adjLogLik is the LogLik value minus the number of 'effective'[3] number of parameters, which is equivalent of the the Akaike Information Criterion.

4.2.3 MCMC simulation (Further Action)

See **Figure 21** for resulting output (explained below).

[2] This may be only a local maximum point, not the global maximum (i.e. the Maximum Likelihood Estimate)

[3] 'effective' means that the NOC-1 mixture proportion parameters are counted, since all sum to one

Important notification: It is important that the MCMC sampler is performing well. This is achieved by requiring an acceptance rate of around 0.25 (0.15-0.35). The acceptance rate can be tuned by modifying '**Set variance of randomizer**' under the MCMC toolbar. Notice that the acceptance rate is automatically calibrated for the 'LR-sensitivity' (see 4.5.1).

- Performs 'Markov Chain Monte Carlo (MCMC) Metropolis Hastings' sampling under the desired hypothesis.
    - A progress bar is shown in the R-console.
    - The MCMC proposal function is based on a normal distribution with the mode as and the covariance matrix attained from the optimization as parameters.
- Estimates the marginal likelihood using the GD-method (Gelfand and Dey 1994), which is further used to estimate Bayes Factor (LR for Bayesian approach). The estimated Bayes factor LR is shown in report if "LR sensitivity" is applied.
- The **first column** in the output shows the estimated posterior distributions for each of the unknown parameters in the model.
- The **second column** in the output monitors the parameter samples in the simulation (trace plot).
- After sampling, the **acceptance rate** of the sampler is printed out to the R-console: Acceptance rate = number of accepted samples divided by number of proposed samples.
- The user may change the **number of samples** in the simulation under 'MCMC->Set number of samples' (n) in toolbar. The number of total samples will be same as the one as specified here.
- The MCMC simulation can be used as an **exploratory tool** to show:
    - That the optimizer has found the global maximum.
    - To infer the posterior distribution of the parameters.
    - That the MCMC sampler used for 'LR sensitivity' **performs well**.

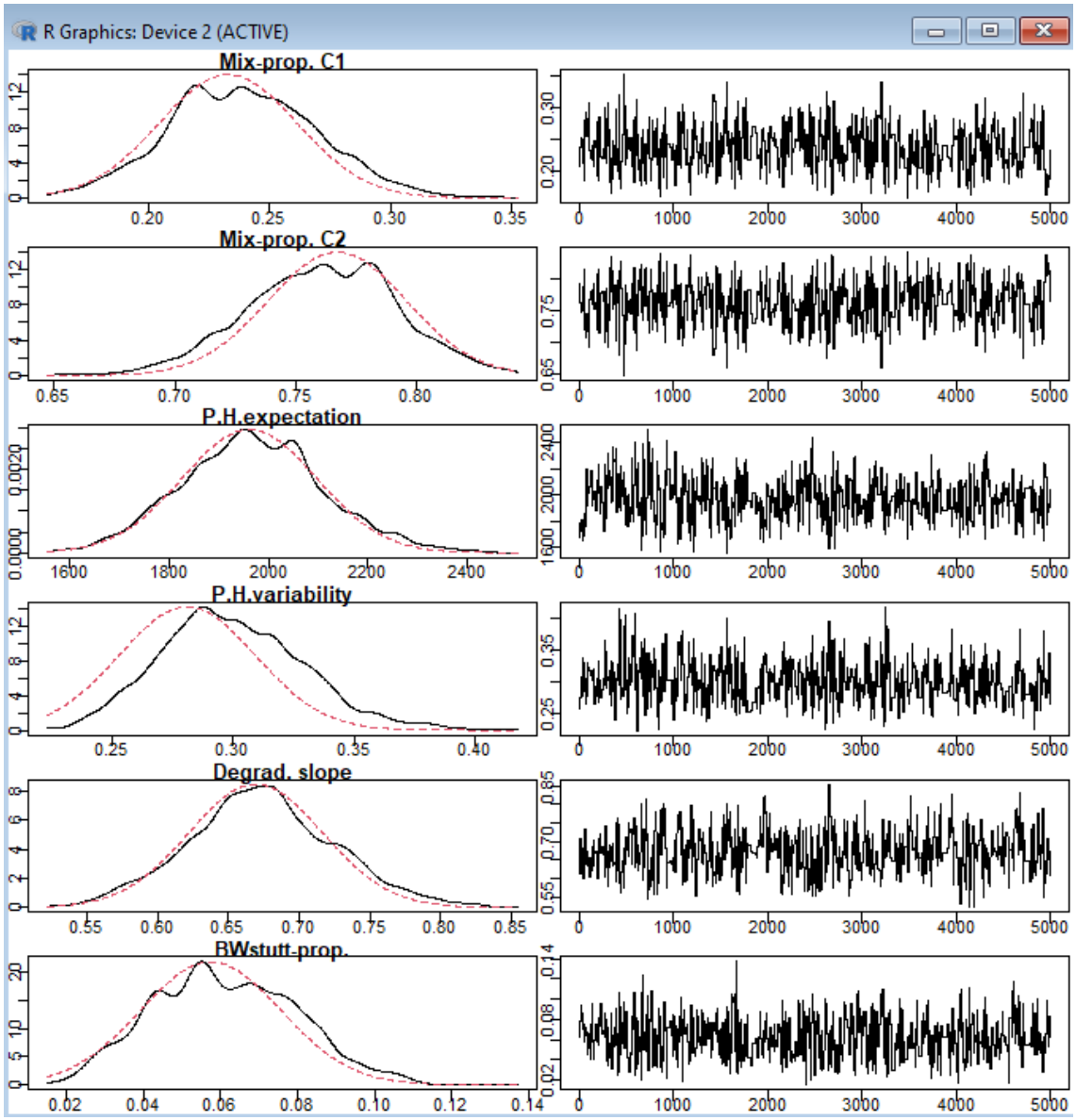


***Figure 21**: The figure shows the posterior density of the unknown parameters (first column) and corresponding iteration values (second column) from the MCMC method under the hypothesis Hp: "Reference P2+1 unknown individual contributes to evidence evid1", where 5000 number of samples were selected.*

4.2.4 Deconvolution (Further Action)

Performs "Deconvolution" under the desired hypothesis, where the genotypes for the unknown contributors are ranked with respect to the posterior probability (based on the quantitative likelihood function and allele frequencies). See section 5-Deconvolution.

*4.2.5 **Model validation** (Further Action)*

- Estimates the cumulative probability of the observed peak heights conditional on the other peak heights. These probabilities are compared with the theoretical underlying model (see references for more details).
- In theory the cumulative probabilities follow a uniform distribution, if the underlying density model is "reasonable" (null-hypothesis) – giving a straight line in the plot.
- The j-th largest (out of n alleles) observed probability (y-axis) is distributed as *beta(j, n - j + 1)* when observed probabilities are independently uniform(0,1).

- Quantiles of the beta-distribution are shown as the envelopes in **Figure 22.** The black lines are the 0.005 and 0.995 quantiles, while the red lines are the Bonferroni-adjusted 0.005/n and 0.995/n quantiles.
- The significance level can be set by the user in the Optimization toolbar. Default is 0.01.

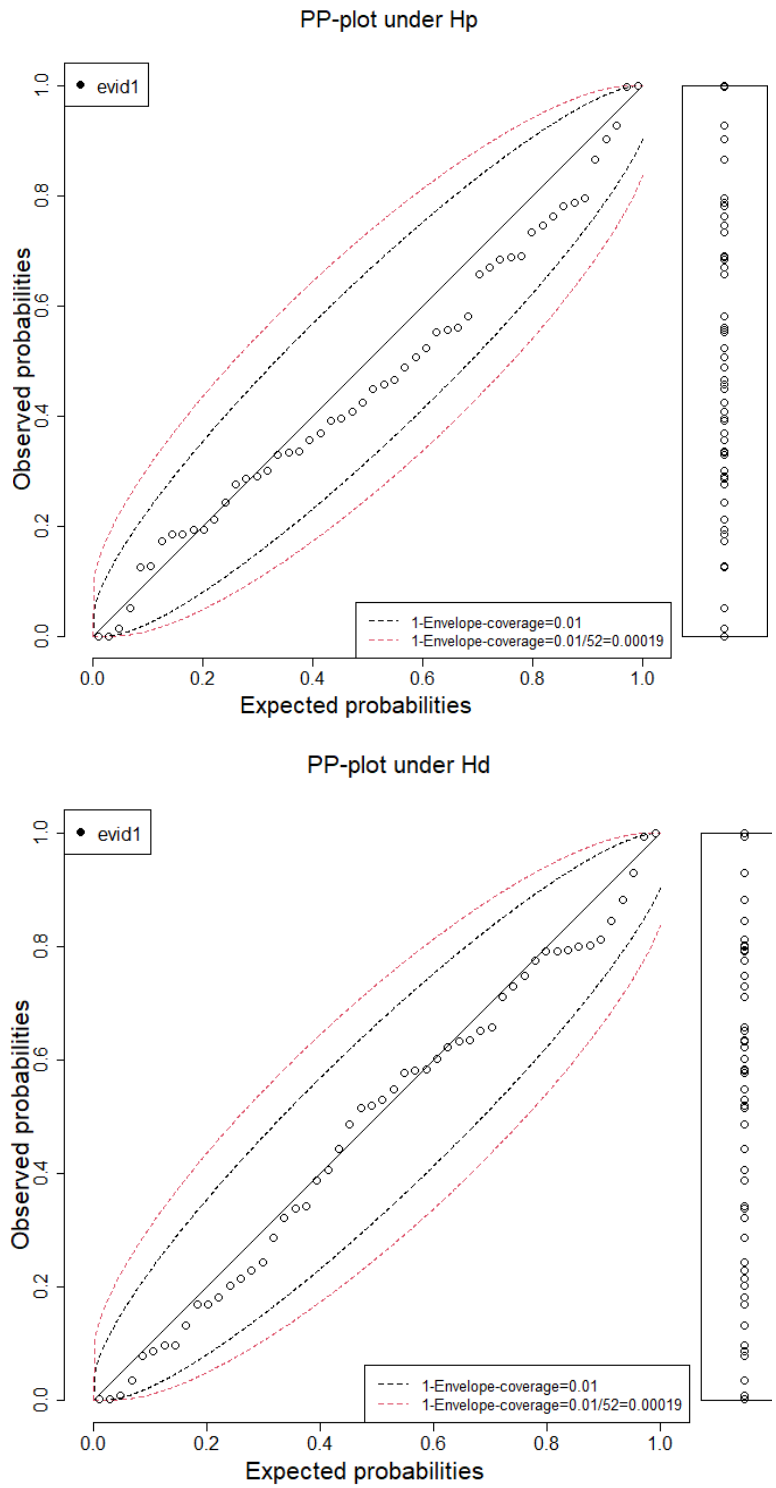
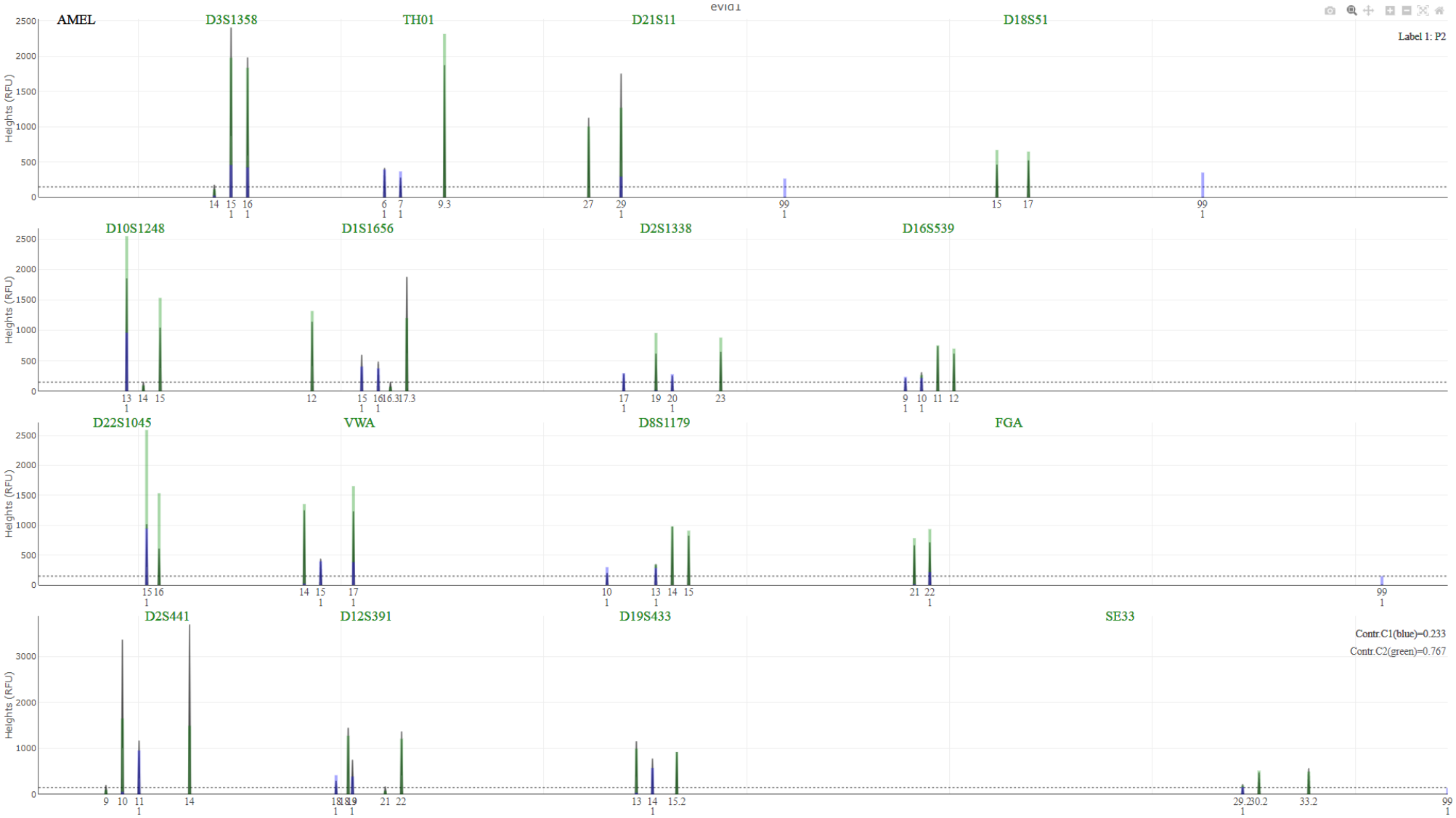


***Figure 22**: The figure shows "Model validation" under Hp (upper) and Hd (lower).*

***Figure 23**: The figure shows the results when clicking Model fitted P.H. under Hp.*


evid1
Label 1: P2
AMEL
D3S1358
TH01
D21S11
D18S51
D10S1248
D1S1656
D2S1338
D16S539
D22S1045
VWA
D8S1179
FGA
D2S441
D12S391
D19S433
SE33
Contr.C1(blue)=0.233
Contr.C2(green)=0.767
Heights (RFU)

*4.2.6 **Model fitted P.H.** (Further Action)*

See **Figure 23**

- This button gives a plot which presents the expected peak heights for each contributor in color bars, which are superimposed on top of the peak heights.
- The expectations are conditioned on the maximum likelihood estimates of the parameters and the most likely genotype for the unknown contributors (see section 5-Deconvolution).
- If the joint probability of the unknown genotypes is above 0.95 the locus name is colored green, and colored orange if between 0.9 and 0.95 and otherwise red.
- Drop-out alleles for contributors are presented as "99".
- The figure is shown in the browser if plotly is installed.

4.3 Joint LR

The LR value is calculated as the ratio between the maximized likelihoods of the two specified hypotheses Hp and Hd as specified in "Model specification". The likelihood function is based on the quantitative model as described in the references.

- log10LR: The ten-logged value of the LR.
- 'Upper boundary': The theoretical upper boundary of the ML based LR (given at log10 scale). Different scenarios:
  - When **one unknown** under Hd: Calculated as the inverse match probability (*IMP*) of the POI profile
    - Fst-correction and conditional reference profiles utilized in the Balding-Nichols sampling formula.
  - When **at least 2 unknowns** under Hd and fst/theta>0, *IMP* is scaled with following expression:
    - *(1+(3+2*nCond)*fst)/(1+(1+2*nCond)*fst)*(1+(4+2*nCond)*fst)/(1+(2+2*nCond)* fst)*
    - nCond is number of conditional references under Hd.

Important notification: Comparing the optimized LR against 'Upper boundary' LR is useful as a diagnostic of whether the maximum likelihood (under Hp or Hd) is obtained: The ML based LR should never exceed 'Upper boundary'.

- Show LR per-marker: Provides the calculated LR for each locus (based on MLE under Hp and Hd).

4.4 Non-contributor analysis

The user may calculate the LR values for random non-contributor profiles which replace **the selected reference** under the drop-down list (these are references considered under the Hp hypothesis but not Hd).

- Setting fst>0 may be very time-consuming since we require that the sampled non-contributor individual is a known non-contributor under Hd, and hence the likelihood value for Hd must be calculated for each sample.
- Supporting related non-contributors: Defined as the last **unknown individual** as specified under Hd.

4.4.1 The number of non-contributors

- Can be changed under 'Database search' in the toolbar (default is 10).
- It's recommended to increase this number after checking the run-time.

4.4.2 Select reference to replace with non-contributor

A drop-down list of references which are conditioned under Hp but not under Hd.

4.4.3 Sample MLE based

Sampled random individuals are calculated with the **Quantitative LR (Maximum Likelihood based)** method

4.4.4 Sample integrated based

Sampled random individuals are calculated with the **Quantitative LR (Bayesian based** using numerical integration**)** method.

4.4.5 Non-contributor results

- The mean, standard errors of LR, proportion of LR greater than zero and one, and log10LR-quantiles (50%, 95%, 99%, max) are provided in the plot.
- A plot of the cumulative distribution of log10LR is shown (**Figure 24**) where we applied fst=0 for faster run.

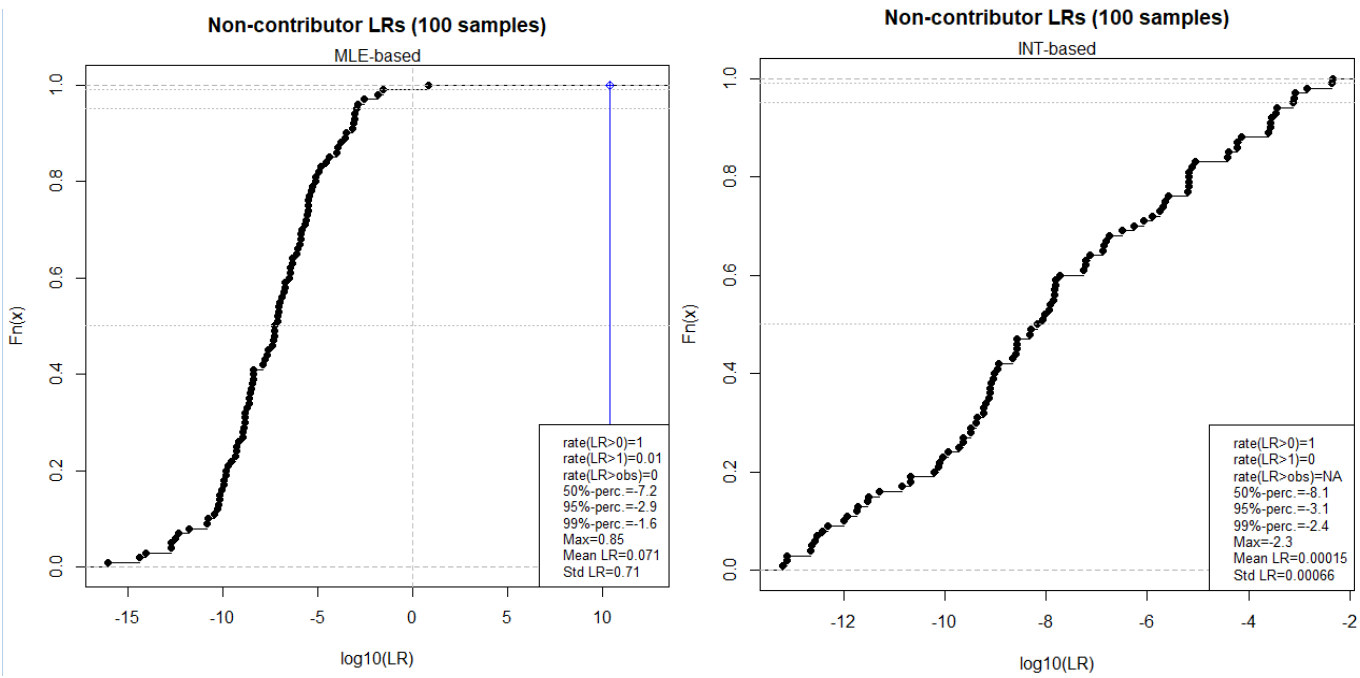


***Figure 24**: The plot shows the cumulative distribution of 100 non-contributing individuals replacing P2 in hypothesis Hp based on the fitted MLE based method(left) and for the Bayesian based method (right). The mean and standard errors of LR, proportion of LR greater than zero and one, and log10LR-quantiles (50%, 95%, 99%, max) based on the simulated non-contributors are given in the plot as well. A fst=0 was used to speed up the calculations.*

4.5 Further

4.5.1 LR sensitivity

Under case: '**Weight-of-Evidence**'.

MCMC simulation is applied under both Hp and Hd (independently) to provide Bayesian inference of the LR where the uncertainty of the parameters in the quantitative model is taken into account. Two methods are provided: 1) Conservative LR and Bayes factor.

1) Conservative LR: MCMC simulation is applied under both Hp and Hd (independently) with equally long chains. The posterior likelihood values from each chain are divided with each other's to produce a posterior distribution of LR. The 5% percentile of the distribution is calculated and used as the "conservative LR". User can change chosen percentile under "MCMC" -> "Set quantile". A 95% CI of the percentile is provided based on bootstrap re-sampling method where the effective sample size is used as a basis.

2) Bayes factor: An estimate of the marginal likelihood is calculated based on the GD-method (see **MCMC simulation section 4.2.3** for more details).

Notes:

- A recommended number of samples is provided to the user (based on the number of unknown parameters). The user can extend the number of samples by clicking LR-sensitivity multiple times which accumulates the results. The number of samples provided in a run can be changed with **Set number of samples** under MCMC in Toolbar (default is 2000 samples).
- The MCMC results will not vary between runs for the same data when same seed is used. However different seeds may yield different results. The variation of the obtained results decreases when the number of samples increases.
- The results are reproducible since it depends on the seed set under MCMC in Toolbar (**section 1.4.3**). The Hd-seed is fixed as "seed+999" in order to avoid dependent chains.
- A trace plot for the Bayes Factor and Conservative LR (from MCMC sampling) are provided in a separate plot (new from version 4), see left plot in **figure 26**.
- The software automatically calibrates the '**variance of randomizer**' to satisfy an acceptance rate of around 0.25 (0.15-0.35), carried out by drawing 100 samples under Hp.

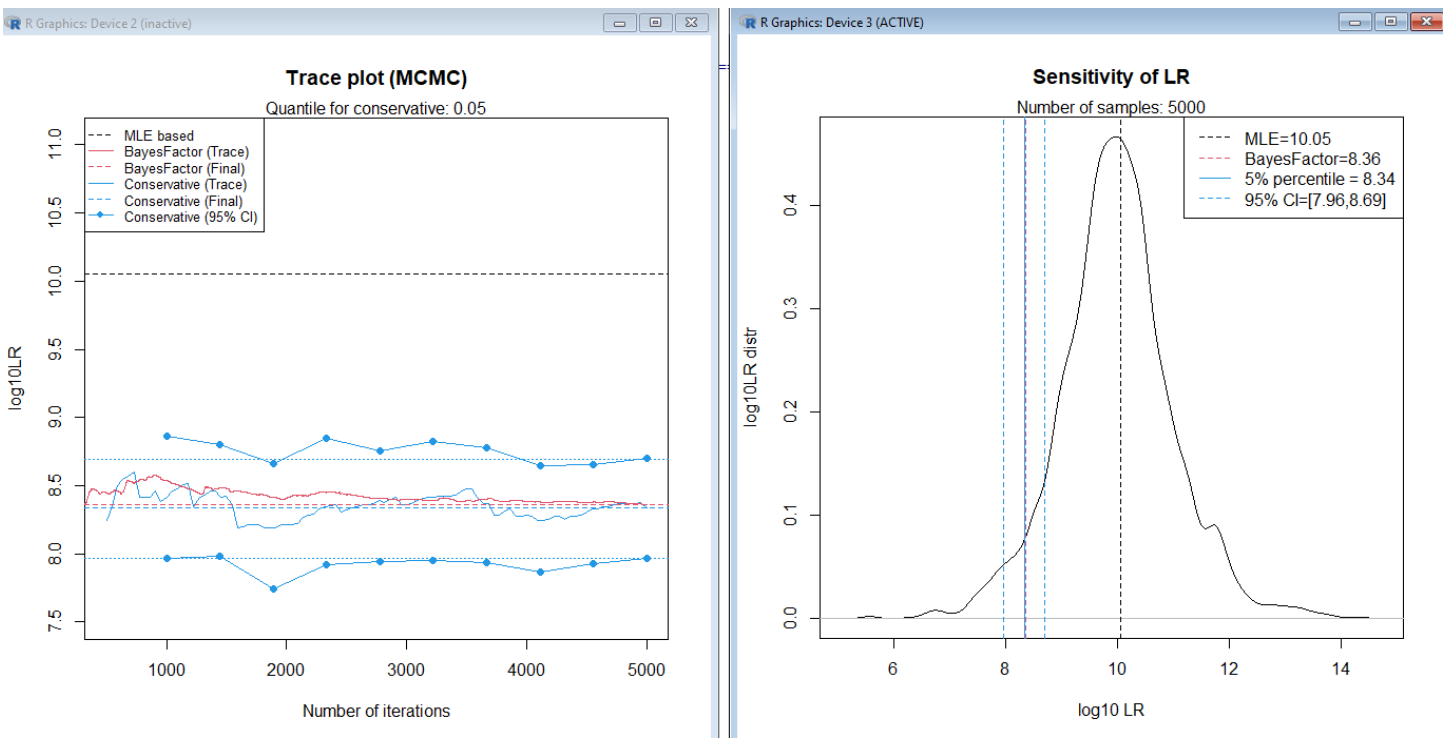


***Figure 25**: Resulting plots when calculating LR-sensitivity based on 5000 samples. Left plot shows the (accumulated) trace-plot of the calculated results (Bayes Factor and Conservative LR). Right plot shows the log10 LR distribution based on the posterior LR. Settings: Seed used is 1 and variation of randomizer is 2.*

4.5.2 Create report

Details of data with results from the analysis will be printed to a report file (**Figure 26**). Results from following analysis will be included if evaluated: 'LR sensitivity', 'Bayes Factor', 'Non-contributor analysis'.

```
This is a generated report from
EuroForMix version 4.0.1
R-version: R version 4.2.1 (2022-06-23 ucrt)
User: oyvbl
Created: 2022-12-30 14:12:52

-------Data-------
Selected STR Kit: ESX17
Selected Population: ESX17_Norway
Evidence(s)=evid1
Markers=D3S1358/TH01/D21S11/D18S51/D10S1248/D1S1656/D2S1338/D16S539/D22S1045/VWA/D8S1179/FGA/D2S441/D12S391/D19S433/SE33

-------Model options-------
Detection threshold=150
Fst-correction=0.01
Probability of drop-in=0.05
Hyperparam lambda=0.01
Degradation:YES
Backward Stutter:YES
Forward Stutter:NO
Backward Stutter prop. prior=function (x) dbeta(x, 1, 1)
Forward Stutter prop. prior=function (x) dbeta(x, 1, 1)
Adjusted fragmenth-length for Q-allele:NO
Rare allele frequency (minFreq):0.000896950368746264
Normalized after impute: Yes

-------Optimalisation setting-------
Required number of (identical) optimizations: 3
Accuracy of optimisations (steptol): 0.001
Seed for optimisations: NONE

-------Hypothesis Hp---------
Number of contributors: 2
Known contributors: P2

-------Hypothesis Hd---------
Number of contributors: 2
Known contributors:
Known non-contributors: P2

-------Estimates under Hp---------
Param.  MLE     Std.Err.
Mix-prop. C1    0.23297 0.02854
Mix-prop. C2    0.76703 0.02854
P.H.expectation 1961.4   134.5
P.H.variability 0.2810  0.0281
Degrad. slope   0.66917 0.04725
BWstutt-prop.   0.05741 0.01826

logLik=-438.9841
adj.logLik=-443.9841
Number of evals: 433
Time usage (sec): 3
```

```
-------Estimates under Hd---------
Param.  MLE     Std.Err.
Mix-prop. C1    0.6956  0.1122
Mix-prop. C2    0.3044  0.1122
P.H.expectation 1976.4   177.2
P.H.variability 0.36621 0.06563
Degrad. slope   0.65684 0.06055
BWstutt-prop.   0.07236 0.02862

logLik=-462.1304
adj.logLik=-467.1304
Number of evals: 552
Time usage (sec): 3

-------MLE based LR (all markers)------
LR=1.128e+10
log10LR=10.05
Upper boundary: log10LR=16.25

-------MLE based LR (per marker)------
D3S1358          5.9713
TH01             9.2776
D21S11           2.2346
D18S51           0.7275
D10S1248                 2.9514
D1S1656         33.2244
D2S1338         10.7613
D16S539         24.1694
D22S1045                 0.1064
VWA             17.3615
D8S1179         25.3853
FGA              1.6787
D2S441           0.1112
D12S391         26.9371
D19S433          3.6161
SE33             5.7607


---RESULTS BASED ON MCMC SAMPLING---
Conservative LR (5%): log10LR=8.341
95% CI of conservative LR (5%): log10LR=[7.963,8.693]
Bayes Factor (MCMC): log10LR=8.359
Number of MCMC samples (setting): 5000
Variation of randomizer (setting): 2
Tuned variation of randomizer (estimated): 1.1658
Seed of randomizer (setting): 1

-------Evaluating data---------
        evid1 | P2 | Freqs.

D3S1358
14      | 178   |        |0.12411
15      | 2405  | x      |0.27099
16      | 1982  | x      |0.23155
99      |       |        |0.37334

TH01
6       | 419   | x      |0.20927
7       | 282   | x      |0.21247
9.3     | 1871  |        |0.34429
```

***Figure 26**: The stored information in the text file generated from **'Create report'**.* Notice the seed settings (with other settings) for Optimization and MCMC for obtaining reproducible results.

4.5.3 Database search

Under case: '**Database search**'.

Search database with the specified quantitative model applied. See section 6 Database search for details.

4.5.4 'Quantitative LR (Bayesian based)'

Under case **Weight-of-Evidence** and **'Database search'**

Instead of maximizing the likelihood of the unknown parameters, a **numerical integration** over the unknown parameters is applied under both Hp and Hd hypotheses. The ratio becomes the estimated LR value, named *Bayes Factor* (marginalized with respect to the model parameters). A message with the estimated LR, along with the relative errors given in brackets, pops up after calculation (**Figure 27**). The user can choose to include the result to the report.

- The accuracy of the integrals depends on the specified '**relative error requirement'** (see EuroForMix paper [1] for details). Value can be changed under "Integration" in Toolbar. Default is 0.1.

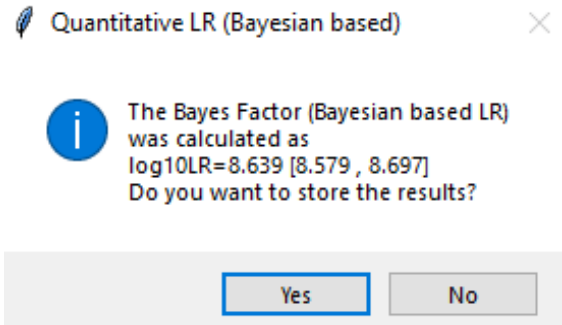
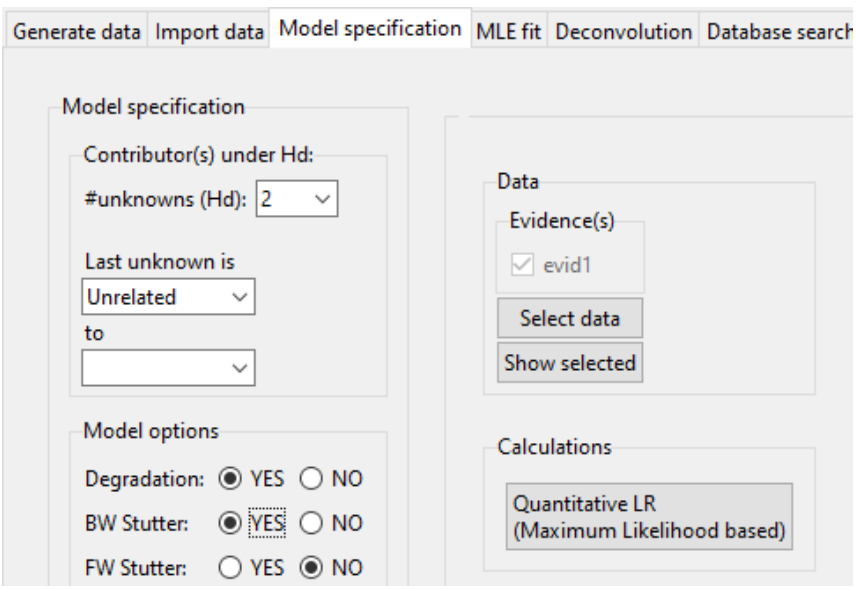


***Figure 27**: The figure shows the calculated Weight-of-Evidence based on the Bayesian based quantitative LR for the specified model in Figure **17**.*

5 Deconvolution

- Deconvolution is applied under a specific hypothesis (for instance Hd as shown in **Figure 28**). Click "Quantitative LR (Maximum Likelihood based)" to proceed.
- For a given optimized model, either Hd or Hp (**case Weight-of-Evidence**), the user must click on "Deconvolution" under "Further Action": See section **4.1 Estimates under Hd.**
- The deconvolution conditions on the optimized parameters (i.e. the MLE fit in **Figure 28**) for the quantitative model. Hence the deconvolution may handle multiple replicates, allele drop-in, drop-out, backward-stutter, degradation, theta-correction and relationship (if considered).
- The results tables will show different types of probabilities which are useful for deconvolution (**Top Marginal (Figure 29), All Joint (Figure 30), All Marginal (G) (Figure 31), All Marginal (A) (Figure 32)**). The probabilities give quantifications of how "certain" different genotypes/alleles are for the different contributors at different loci".
- See the supplementary of the CaseSolver paper [2] describing how the marginal probabilities are derived.



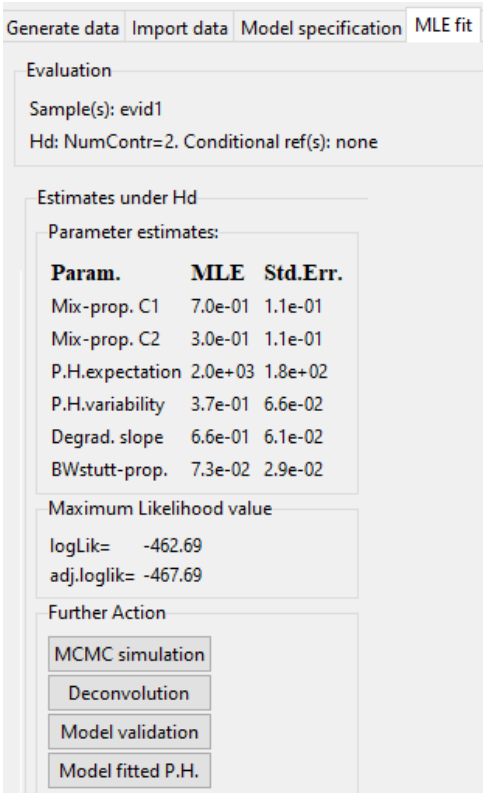


***Figure 28**: The left figure shows the Model Specification page for doing **Deconvolution**. We don't condition on any of the contributors, and we assume two unknowns in the hypothesis. We turn both degradation and backward stutter model options on. The right figure shows the optimized parameters (i.e. the MLE fit) for the quantitative model. The fitted model has the same "Further Action" possibilities as for "Weight-of-Evidence" and "Database search" in order to optimize the model.*

5.1 Result tables

- Notes:
  - Maximum length of table is 20 by default. Can be changed under 'Deconvolution-> Set max listsize' in toolbar.
  - The allele named as 99 represents alleles which are not in the evidence

5.1.1 Top marginal

Figure 30: Gives the top genotype with corresponding probability (most likely) marginalized for each contributor and each locus. "TopGenotype_Ck" gives the most likely genotype for contributor k (same order as in **MLE fit**), with corresponding probability under "probability_Ck"). The "ratioToNextGenotype_Ck" column gives the ratio of the largest probability (i.e. probability_Ck) to the second largest probability. The probabilities become one for known contributors.

5.1.2 All Joint

Figure 31: A ranked table of the combined genotype profiles for all contributors (C1,…,CK) with corresponding probabilities, given for each locus. The probabilities become one for known contributors.

5.1.3 All Marginal (G)

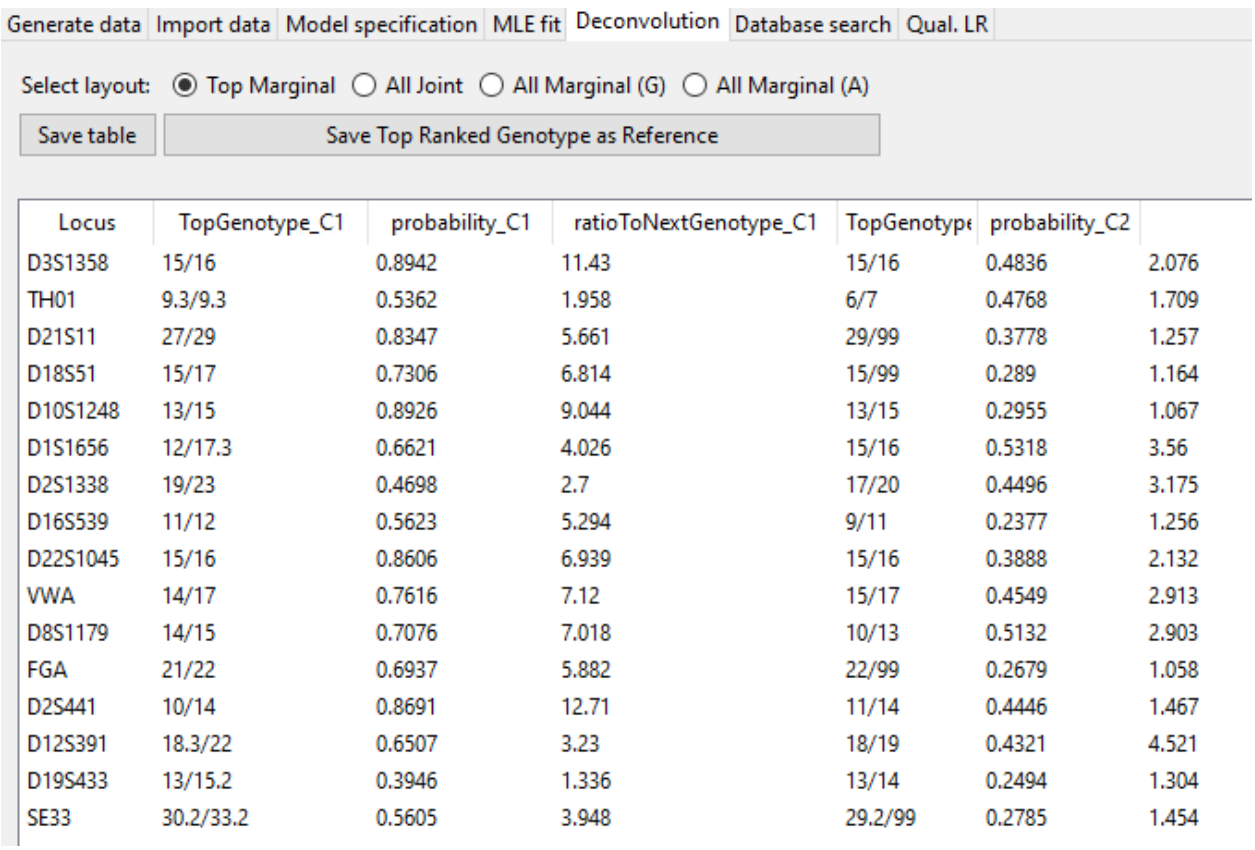
Figure 32 (left): A ranked table of the genotype profiles for each of the contributors, for each locus. The probabilities become one for known contributors.

5.1.4 All Marginal (A)

Figure 33 (right): A the ranked table of single alleles for each of the contributors for each locus. The probabilities become one for known contributors.

5.1.5 Save results

- **Save table**: The corresponding table will be exported to a tabulate-separated text-file.
- **Save Top Ranked Genotype as Reference**: Creates a reference table of the top ranked genotypes into a tabulate-separated text-file.



Locus	TopGenotype_C1	probability_C1	ratioToNextGenotype_C1	TopGenotype	probability_C2	
D3S1358	15/16	0.8942	11.43	15/16	0.4836	2.076
TH01	9.3/9.3	0.5362	1.958	6/7	0.4768	1.709
D21S11	27/29	0.8347	5.661	29/99	0.3778	1.257
D18S51	15/17	0.7306	6.814	15/99	0.289	1.164
D10S1248	13/15	0.8926	9.044	13/15	0.2955	1.067
D1S1656	12/17.3	0.6621	4.026	15/16	0.5318	3.56
D2S1338	19/23	0.4698	2.7	17/20	0.4496	3.175
D16S539	11/12	0.5623	5.294	9/11	0.2377	1.256
D22S1045	15/16	0.8606	6.939	15/16	0.3888	2.132
VWA	14/17	0.7616	7.12	15/17	0.4549	2.913
D8S1179	14/15	0.7076	7.018	10/13	0.5132	2.903
FGA	21/22	0.6937	5.882	22/99	0.2679	1.058
D2S441	10/14	0.8691	12.71	11/14	0.4446	1.467
D12S391	18.3/22	0.6507	3.23	18/19	0.4321	4.521
D19S433	13/15.2	0.3946	1.336	13/14	0.2494	1.304
SE33	30.2/33.2	0.5605	3.948	29.2/99	0.2785	1.454

***Figure 29**: The figure shows **Top Marginal,** the top ranked genotypes (TopGenotype) for each contributor per loci, with corresponding probabilities **probability_Ck,** for each of the contributors, k=1,…,K. ratioToNextGenotype is the ratio of the largest probability (i.e. **probability_C**) to the second largest probability.*

Generate data | Import data | Model specification | MLE fit | Deconvolution | Database search | Qual. LR

Select layout: ○ Top Marginal ◉ All Joint ○ All Marginal (G) ○ All Marginal (A)

Save table | Save Top Ranked Genotype as Reference

Locus	Ranl	C1	C2	Probability
D3S1358	1	15/16	15/16	0.4686
D3S1358	2	15/16	15/15	0.2101
D3S1358	3	15/16	16/16	0.07225
D3S1358	4	15/15	16/16	0.06121
D3S1358	5	15/16	14/15	0.04542
D3S1358	6	15/16	14/16	0.04141
D3S1358	7	15/16	15/99	0.03417
D3S1358	8	16/16	15/15	0.02277
D3S1358	9	15/16	16/99	0.01913
D3S1358	10	15/15	15/16	0.01315
D3S1358	11	15/16	14/99	0.002845
TH01	1	9.3/9.3	6/7	0.4643
TH01	2	6/9.3	7/9.3	0.2364
TH01	3	7/9.3	6/9.3	0.1325

***Figure 30**: The figure shows **All Joint,** the ranked table of the combined genotype profiles for all the contributors combined (here, C1 and C2) for each locus.*

Generate data | Import data | Model specification | MLE fit | Deconvolution | Database search | Qual. LR

Select layout: ○ Top Marginal ○ All Joint ◉ All Marginal (G) ○ All Marginal (A)

Save table | Save Top Ranked Genotype as Reference

Con	Locus	Genotype	P
C1	D3S1358	15/16	0.89420
C1	D3S1358	15/15	0.07822
C1	D3S1358	16/16	0.02733
C2	D3S1358	15/16	0.48360
C2	D3S1358	15/15	0.23290
C2	D3S1358	16/16	0.13350
C2	D3S1358	14/15	0.04778
C2	D3S1358	14/16	0.04319
C2	D3S1358	15/99	0.03457
C2	D3S1358	16/99	0.02124
C1	TH01	9.3/9.3	0.5362
C1	TH01	6/9.3	0.2739
C1	TH01	7/9.3	0.1831
C2	TH01	6/7	0.476800
C2	TH01	7/9.3	0.279000
C2	TH01	6/9.3	0.144400
C2	TH01	9.3/9.3	0.035970
C2	TH01	7/7	0.020240
C2	TH01	7/99	0.018780
C2	TH01	6/6	0.010120
C2	TH01	6/99	0.008746

Generate data | Import data | Model specification | MLE fit | Deconvolution | Database search | Qual. LR

Select layout: ○ Top Marginal ○ All Joint ○ All Marginal (G) ◉ All Marginal (A)

Save table | Save Top Ranked Genotype as Reference

Con	Locus	Allel	Proba
C1	D3S1358	15	0.9725
C1	D3S1358	16	0.9217
C2	D3S1358	15	0.79890
C2	D3S1358	16	0.68150
C2	D3S1358	14	0.09399
C2	D3S1358	99	0.05880
C1	TH01	9.3	0.9942
C1	TH01	6	0.2796
C1	TH01	7	0.1888
C2	TH01	7	0.79490
C2	TH01	6	0.64020
C2	TH01	9.3	0.46520
C2	TH01	99	0.03335
C1	D21S11	29	0.99000
C1	D21S11	27	0.84470
C1	D21S11	99	0.01257
C2	D21S11	29	0.8367
C2	D21S11	99	0.5140
C2	D21S11	27	0.2856

***Figure 31**: The left figure shows **All Marginal (G),** the ranked table of the genotype profiles for each of the contributors for each locus. The right figure shows **All Marginal (A),** the ranked table of single alleles for each of the contributors for each locus.*

6 Database searching

- The database to search must be loaded first from the Import data page.
- Click the database search button from the Import data page which takes you to the Model specification page
- The 'Database search' is very similar to the Weight-of-Evidence (**Figure 32**) with the only difference is that each individual in the reference-database is assumed to be a contributor in the hypothesis Hp. For each individual 'j' in reference-database we calculate a LR-value LRj.
- The user can utilize the peak heights in a **'Quantitative LR' (Maximum Likelihood based**) calculation or ignoring the peak heights in a 'Qualitative LR' calculation.

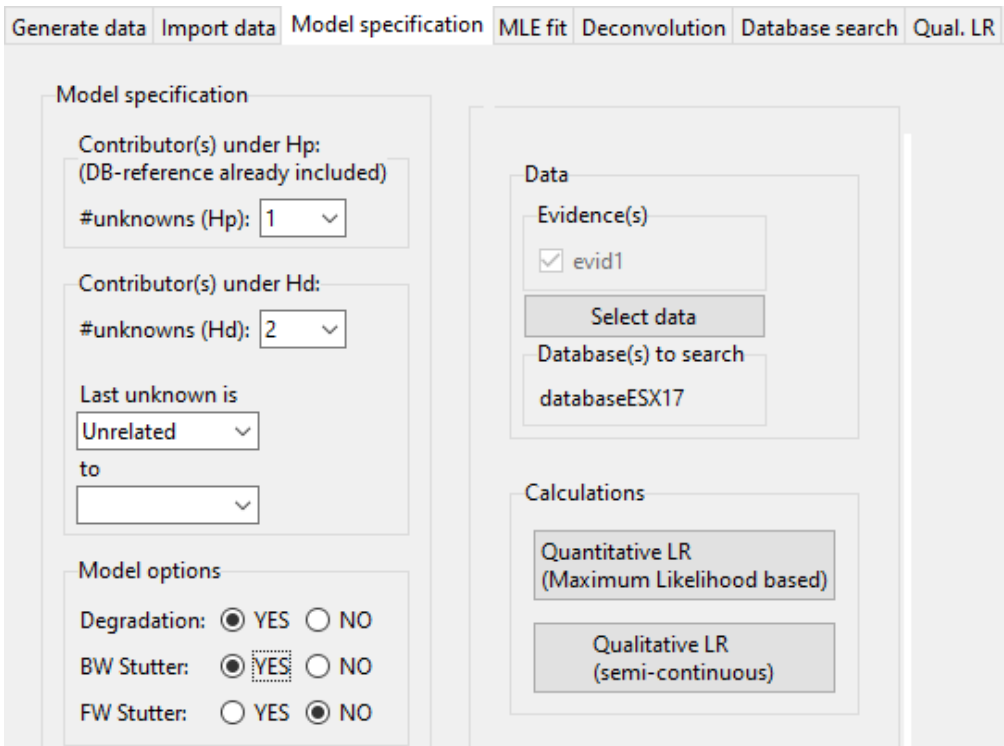


***Figure 32**: The figure shows the page of the model specification for doing database search on the database file "databaseESX17".*

6.1 Searching with Quantitative LR

When selecting 'Quantitative LR': (Leads to the MLE fit page as seen in **Figure 33**). See section **4.1 Estimates under Hd** for Further Action

- A 'Qualitative LR' is always calculated along with the 'Quantitative LR' values (**Figure 34**).
  - The qualitative model assumes an allele drop-out parameter as 0.1 fixed and fst=0.
  - The allele drop-in parameter in the qualitative model is set as default 0.05, but can be changed with "**Set drop-in probability for qualitative model**" under 'Database search' in the Toolbar.

- If "Quantitative LR (Maximum Likelihood based)" calculation is used, the optimized parameters under the Hd - hypothesis are first shown (**Figure 33**). We applied fst=0 for faster calculations.
  - The button **Search Database** must be clicked to proceed.
  - Further actions can be carried out based on the fitted model under Hd.

- The reason for showing the MLE fitted parameters under Hd (**Figure 33**) for "Quantitative LR (Maximum Likelihood based)" calculation is that the user has the possibility to check if the parameter estimates under Hd seems reasonable so he/she can go back and change the model specification.
- Notes:
    - The 'Quantitative LR' calculation is based on the quantitative **model** as given in the references and can handle allele drop-in, drop-out, degradation, backward-stutter and relationship under Hd (if applied).

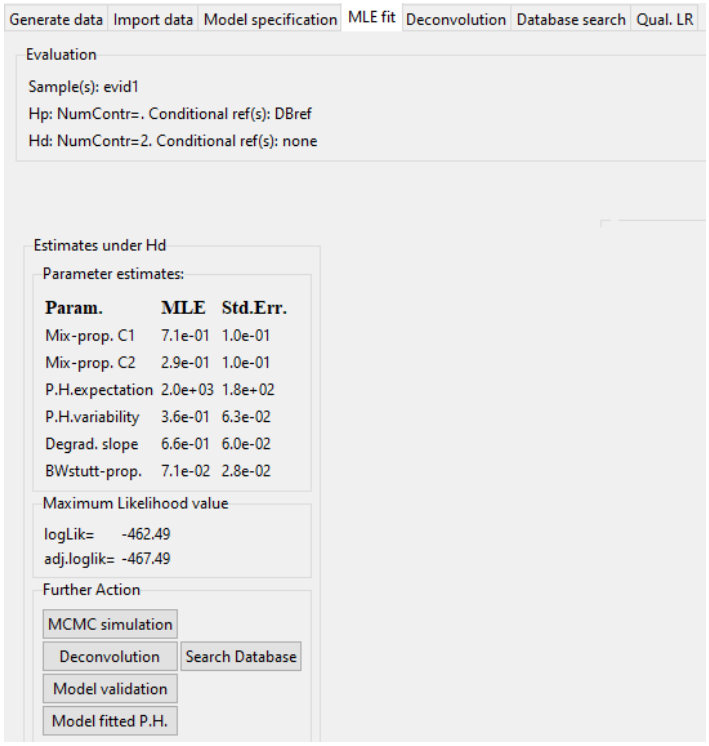


***Figure 33**: The figure shows the optimized parameters (i.e. the MLE fit) for the quantitative model under Hd (with specifications as given in **Figure 32**). The fitted model has the same "Further Action" possibilities as for "Weight-of-Evidence" and "Deconvolution". The user must push the "**Search Database**" button to carry out the actual database searching.*

6.2 Searching with Qualitative LR

When selecting 'Qualitative LR' from the 'database search page:

- The "**Set drop-in probability for qualitative model**" under 'Database search' in the Toolbar is ignored.
- The qualitative model assumes an allele drop-out parameter which is estimated using median of the '*allele drop-out probability given number of observed alleles*' distribution.
- The 'Quantitative LR' calculation is ignored.

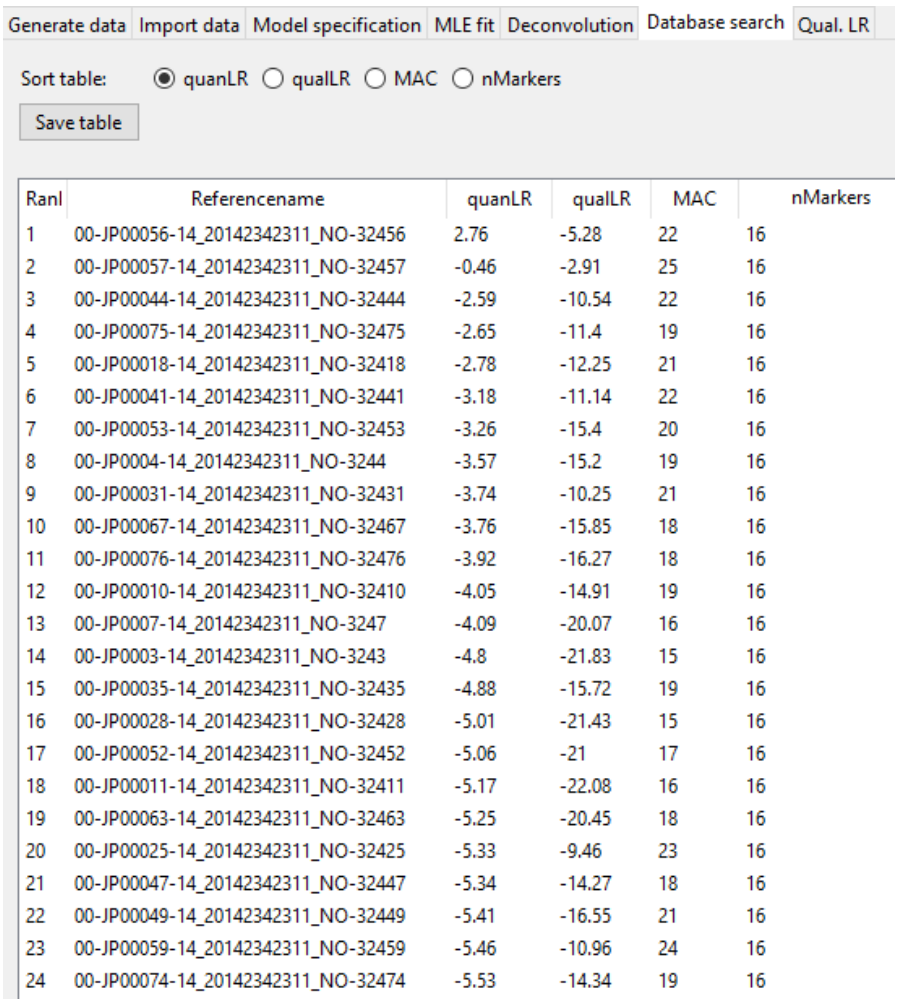
6.3 Search result tables

Database search table (**Figure 34**):

- '**Reference name'** is name of individuals given in the reference-database.
- The table shows the ranked individuals in the database based on the quantitative LR values (**quanLR**), qualitative LR values (**qualLR**), number of matching alleles (**MAC**) or number of evaluating loci (**nLocs**). The LR values are given on log10 scale.
- **qualLR** (Qualitative LR (semi-continuous model))
- Parameter for dropout probability is based on the median of 2000 samples from the 'distribution of dropout-probability'.
    - Number of required samples may be changed under 'Qual LR' in toolbar.

- Dropout probability is fixed to 0.1 when searched with “Quantitative LR”.

- For multiple evidences, the mean of the median is used as the dropout probability parameter.
- Assumes drop-in probability 0.05 as default. Can be changed under ‘Database search’ in toolbar.
- **MAC** (Matching allele counter) is number of alleles in the reference-profile which matches the evidence.
  - MAC is summed over the considered evidences (replicates).
- **nLocs** is number of loci in the reference-profile which are used to calculate the contLR,qualLR and MAC.
  - Some references in the database may have missing loci which are presented in the evaluated evidence.
- Maximum number of elements to view a ‘Database search’ result table is 10000. This can be changed in toolbar ‘Database search->Set maximum view-elements’.
  - Setting fst>0 may be very time-consuming since we require that individual ‘j’ is a known non-contributor under Hd, and hence the likelihood for Hd is calculated for each individual in database.
- **Save table:** The full table will be exported to a tabulator-separated text-file.



Ranl	Referencename	quanLR	qualLR	MAC	nMarkers
1	00-JP00056-14_20142342311_NO-32456	2.76	-5.28	22	16
2	00-JP00057-14_20142342311_NO-32457	-0.46	-2.91	25	16
3	00-JP00044-14_20142342311_NO-32444	-2.59	-10.54	22	16
4	00-JP00075-14_20142342311_NO-32475	-2.65	-11.4	19	16
5	00-JP00018-14_20142342311_NO-32418	-2.78	-12.25	21	16
6	00-JP00041-14_20142342311_NO-32441	-3.18	-11.14	22	16
7	00-JP00053-14_20142342311_NO-32453	-3.26	-15.4	20	16
8	00-JP0004-14_20142342311_NO-3244	-3.57	-15.2	19	16
9	00-JP00031-14_20142342311_NO-32431	-3.74	-10.25	21	16
10	00-JP00067-14_20142342311_NO-32467	-3.76	-15.85	18	16
11	00-JP00076-14_20142342311_NO-32476	-3.92	-16.27	18	16
12	00-JP00010-14_20142342311_NO-32410	-4.05	-14.91	19	16
13	00-JP0007-14_20142342311_NO-3247	-4.09	-20.07	16	16
14	00-JP0003-14_20142342311_NO-3243	-4.8	-21.83	15	16
15	00-JP00035-14_20142342311_NO-32435	-4.88	-15.72	19	16
16	00-JP00028-14_20142342311_NO-32428	-5.01	-21.43	15	16
17	00-JP00052-14_20142342311_NO-32452	-5.06	-21	17	16
18	00-JP00011-14_20142342311_NO-32411	-5.17	-22.08	16	16
19	00-JP00063-14_20142342311_NO-32463	-5.25	-20.45	18	16
20	00-JP00025-14_20142342311_NO-32425	-5.33	-9.46	23	16
21	00-JP00047-14_20142342311_NO-32447	-5.34	-14.27	18	16
22	00-JP00049-14_20142342311_NO-32449	-5.41	-16.55	21	16
23	00-JP00059-14_20142342311_NO-32459	-5.46	-10.96	24	16
24	00-JP00074-14_20142342311_NO-32474	-5.53	-14.34	19	16

Figure 34: *The figure shows the table from the database search with specifications as given in* ***Figure 32*** *based on ‘****Quantitative LR’ (Maximum Likelihood based)”*** *calculations. The references are sorted due to the quantitative LR values. An fst=0 was applied.*

7 Qual. LR: 'Qualitative model'

- From 'Import data' page, check evidence evid1 and reference P2, and press 'Weight-of-Evidence' button which leads to the 'Model specification' page. Under model specifications, construct Hp: 'P2+2 unknown contributors' vs Hd: '3 unknown contributors' as shown in **Figure 35 (left)**. Then select the 'Qualitative LR' button which leads to the 'Qual. LR' page shown in **Figure 35 (right)**.
- This module samples from the distribution of the '*allele drop-out probability given number of observed alleles*' to evaluate the qualitative LR automatically.
    - Note: the model will not fit the data if there are too many alleles compared to the number of contributors – always check that the model specification is reasonable
- Also, a sensitivity plot as a function of allele-dropout probability and a non-contributor sampling analysis is implemented (**Figure 36**).
- Marker specific Drop-in model and Fst can be used.

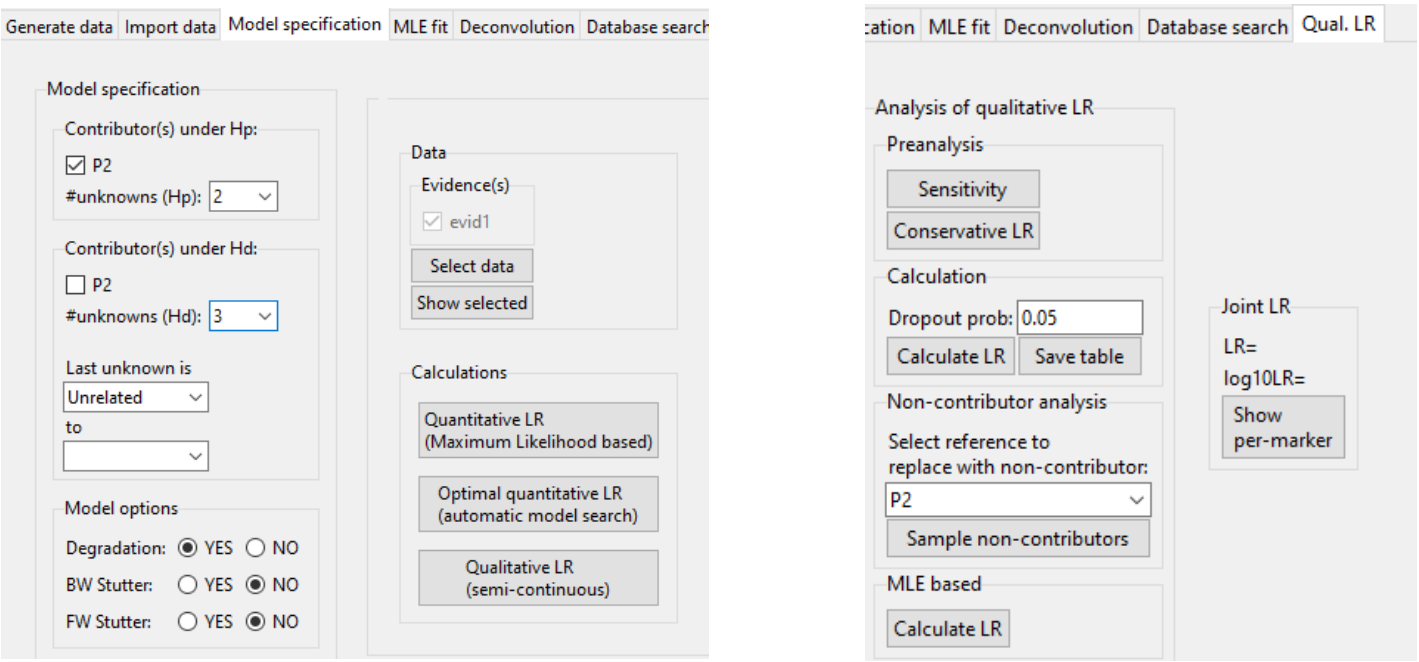


***Figure 35**: The left plot shows the assumed model specification. The right plot shows the page where the weight-of-evidence evaluation based on the qualitative model is carried out.*

7.1 Pre-analysis

7.1.1 Sensitivity

- Plots the log10LR as a function of allele-dropout probability (**Figure 36**).
- This is the same function as in LRmix Studio.
- The upper probability range and number of ticks can be changed under 'Qual LR' in the toolbar.
- Lower dropout probability limit in sensitivity is 1e-6 (something small).

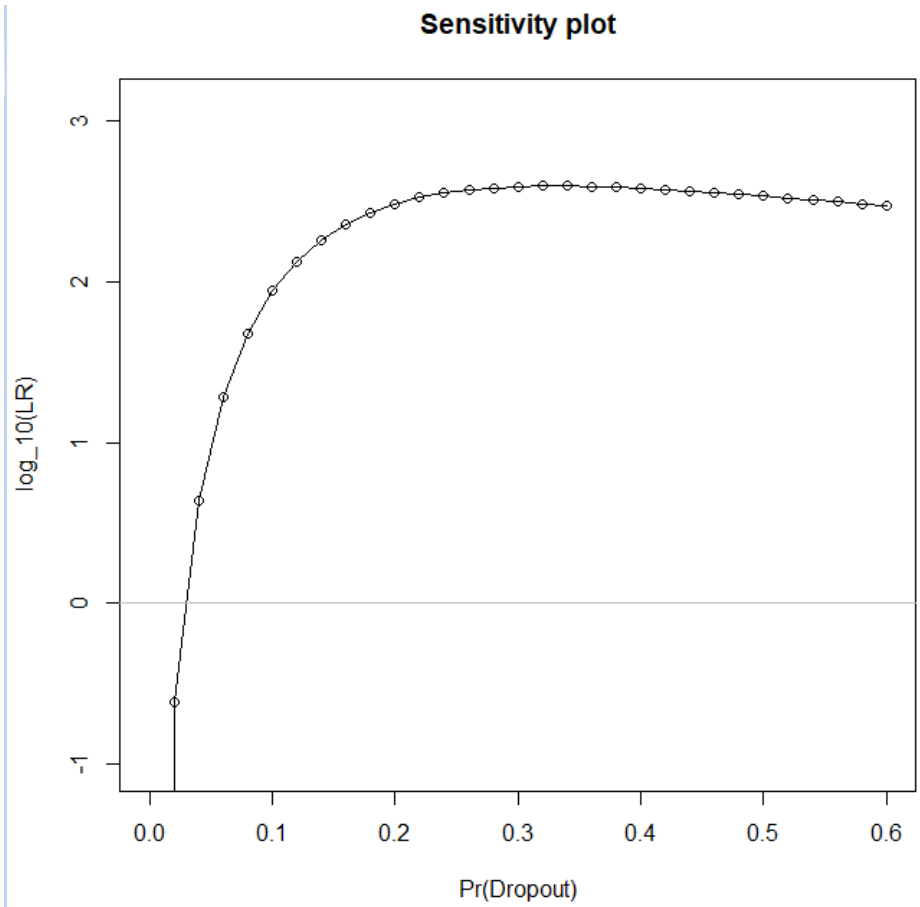


***Figure 36**: The figure shows the plot of Weight-of-evidence (Likelihood Ratio) as a function of allele drop-out probability.*

7.1.2 Conservative LR

- By sampling from the "*allele drop-out probability given number of observed alleles in the evidence"*- distribution for the hypothesis Hp and Hd, the most 'conservative' LR (i.e. smallest from the 5 or 95 percentile) is automatically calculated and printed (see **Figure 37** and **Figure 38**).
- The most "conservative" LR is found by following:
    - Take out the "alpha" and "1-alpha"-quantiles from the simulated 'allele-dropout probability distribution' under both Hp and Hd.
    - The quantile (under both Hp and Hd) which gives the lowest LR is the "conservative LR".
- The significance level "alpha" is given 0.05 as default.
    - This will give similar results as in LRmix Studio.
    - This can be changed under 'Qual LR' in the toolbar.
- The number of required samples from the 'allele-dropout probability distribution' is given 2000 as default.
    - This can be changed under 'Qual LR->Set required samples in dropout distr.'
- If no samples are accepted from the allele-dropout probability distribution', an error-message is provided to the user.
- When more evidence samples are imported (replicates), the most 'conservative LR' over all samples is considered.
    - The dropout probability quantiles are estimated for each of the evidence samples.

```
[1] "Total number of observed alleles for sample(s):"
       x
evid1 52
[1] "For evidence evid1:"
[1] "Estimating quantiles from allele dropout distribution under Hp..."
             x
5%  0.1185368
50% 0.2381719
95% 0.3575913
[1] "Estimating quantiles from allele dropout distribution under Hd..."
             x
5%  0.1454642
50% 0.2682626
95% 0.3843844
       5%  95%
[1,] 0.12 0.36
[2,] 0.15 0.38
```

Figure 37*: The plot shows the sampled 5%, 50% and 95% quantiles of the distribution of the 'allele drop-out probability given number of observed alleles' for each of the hypotheses.*

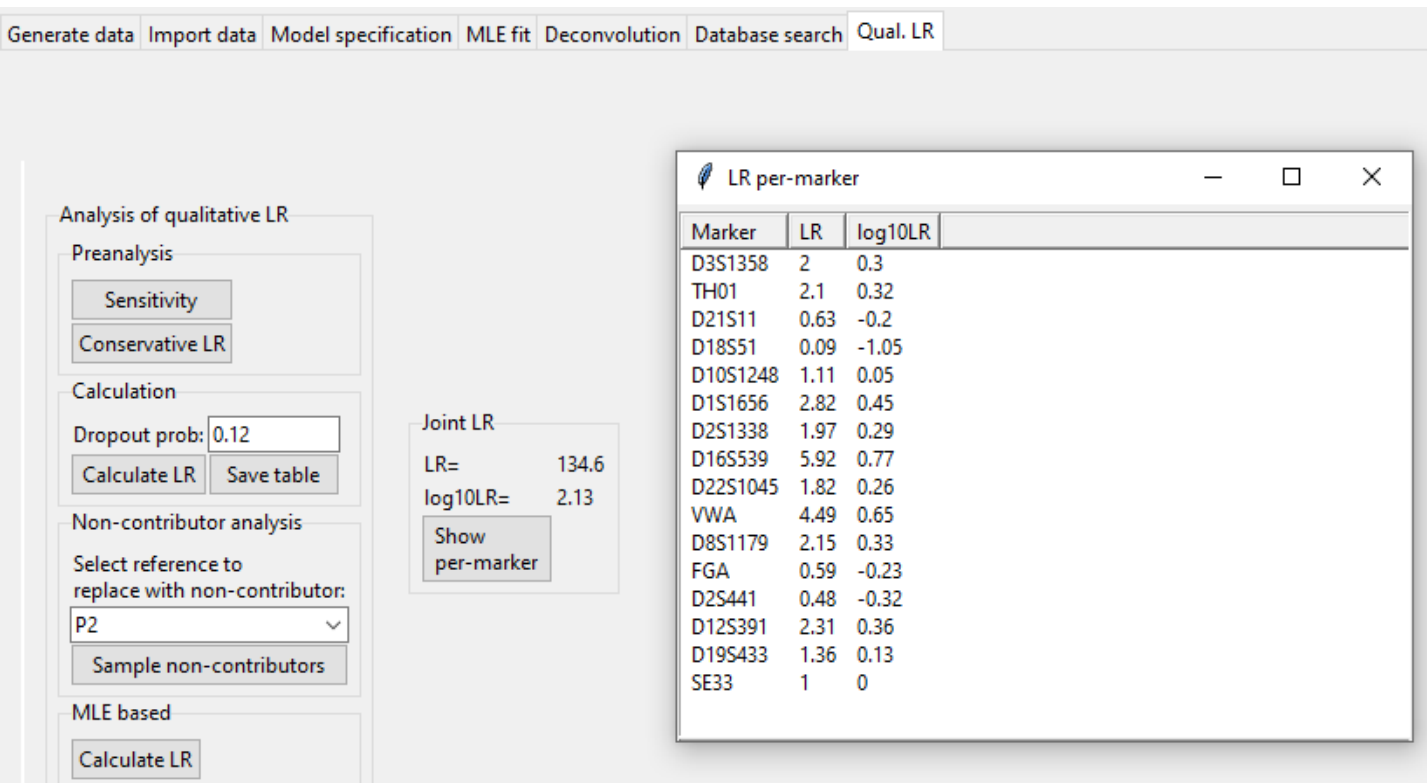


Figure 38*: The plot shows the conservative Weight-of-Evidence values (Likelihood Ratios) after pushing the "**Conservative LR**" button. The most conservative estimated allele drop-out probability-quantile from **Figure 37** was the 5% quantile under Hd which gave 0.12. Hence the table in this plot shows the LR inserted for this value.*

7.1.3 Calculation

- **Dropout prob:**
  - The user may specify the assumed value of the allele dropout-probability.

- **Calculate LR**
  - Instantly calculates the LR for the given user-specified allele dropout probability in "**Dropout prob**".

- **Save table:**
  - Saves the weight-of-evidence calculated LR results to a selected file.

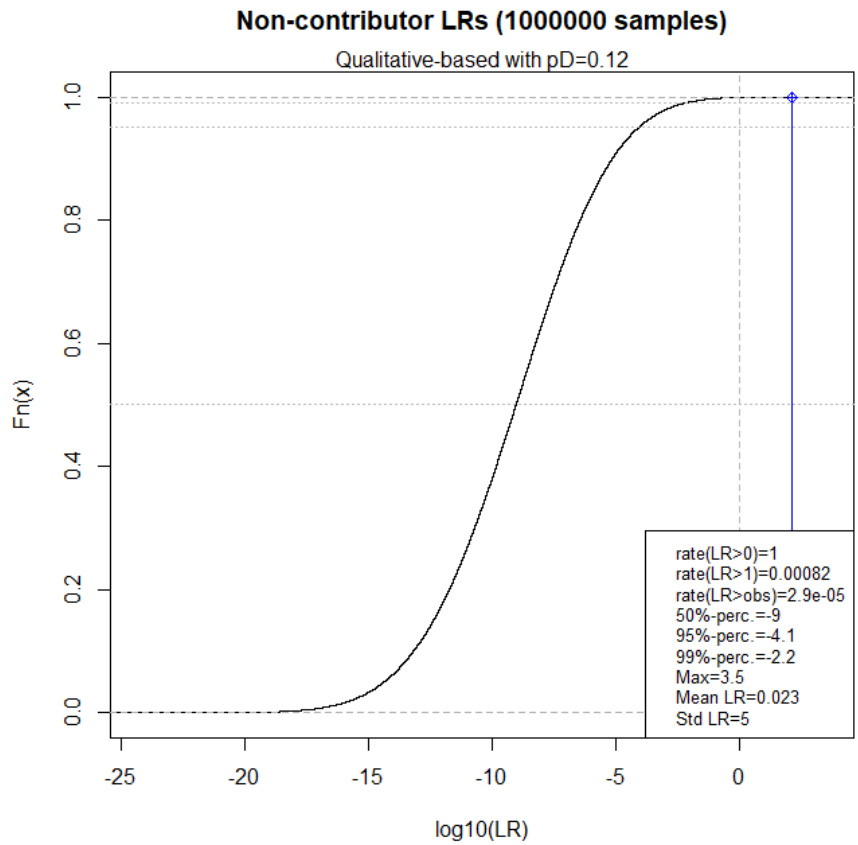


***Figure 39**: The figure shows a cumulative distribution of 1e6 log10LR of non-contributors, where each sample is based on replacing the "P2" in hypothesis Hp with a random man from the population. The reporting LR for the replaced reference (i.e. "P2 in this case) is superimposed as a blue line to the plot. The mean and standard errors of LR, proportion of LR greater than zero and one, and log10LR-quantiles (50%, 95%, 99%, max) based on the simulated non-contributors are given in the plot as well. In upper left box, the proportion of non-contributors LR exceeding the reported LR (v) is given.*

7.2 Non-contributor analysis (post-analysis)

- **Select reference to replace with non-contributor:**
  - A drop-down list of references which are conditioned under Hp but not under Hd.
- **Sample non-contributors:**
  - Random non-contributor samples are provided by replacing the selected reference (under the drop-down list under the Hp hypothesis) with a random individual from the population and then calculate his/her LR.

        - The mean, standard errors of LR and log10LR-quantiles (50%, 95%, 99%, max) are printed out to R-console.
        - A plot of the cumulative distribution of log10LR will be shown (**Figure 39**).
        - Number of non-contributors can be changed under 'Database search->Set number of non-contributors' in the toolbar.
    - If weight-of-evidence has been calculated:
        - The reporting LR for the "replaced reference" is superimposed as a blue line to the plot (**Figure 39**).
        - The discriminatory metric (log10LR-q99%) is printed out to R-console.
    - Note: Precalculations are always carried out previous to the non-contributor sampling, therefore the number of non-contributors is only limited to make the plot.

7.3 Qualitative MLE-based approach (alternative analysis)

This functionality will follow the maximum likelihood approach for estimating the dropout probability for each of the hypotheses.

- **Calculate LR**
    - The LR value based on the maximum likelihood estimates, with the corresponding estimated dropout probabilities for each of the hypotheses, is given in a pop-up window (**Figure 40**).

- **Export**:
    - Export to a text-file with the information.

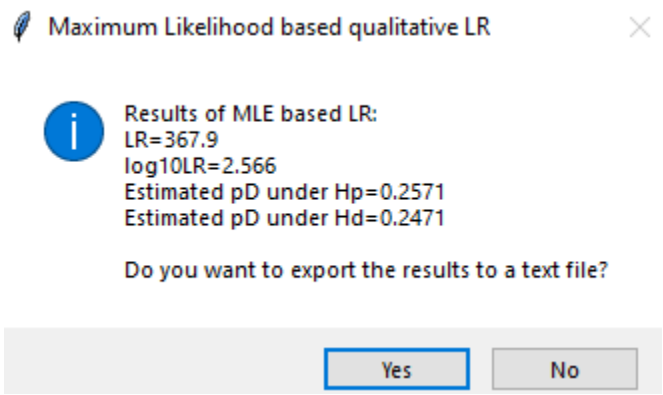


***Figure 40**: The calculated LR value based on the maximum likelihood estimates, with the corresponding estimated dropout probabilities for each of the hypotheses indicated.*

8 Generate data: 'from the quantitative model'

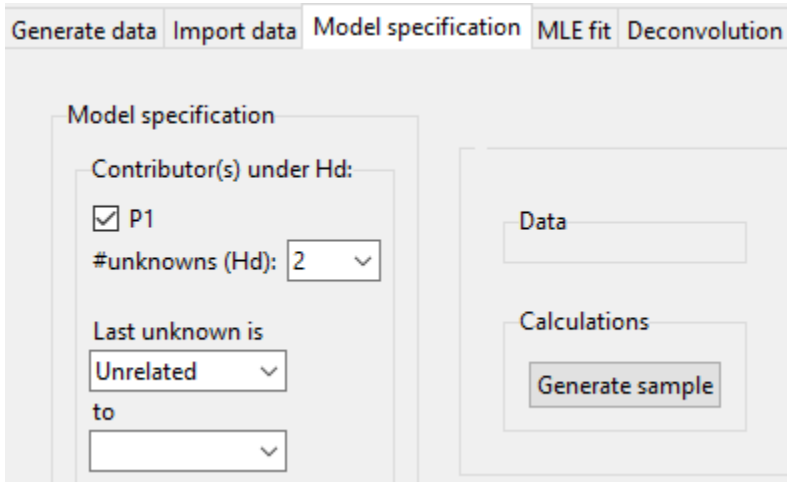


***Figure 41**: The figure shows the Model specification page for generating allele with corresponding peak heights from the quantitative model for a given specified model. From here we will generate data which are contributed P1 and an unknown individual.*

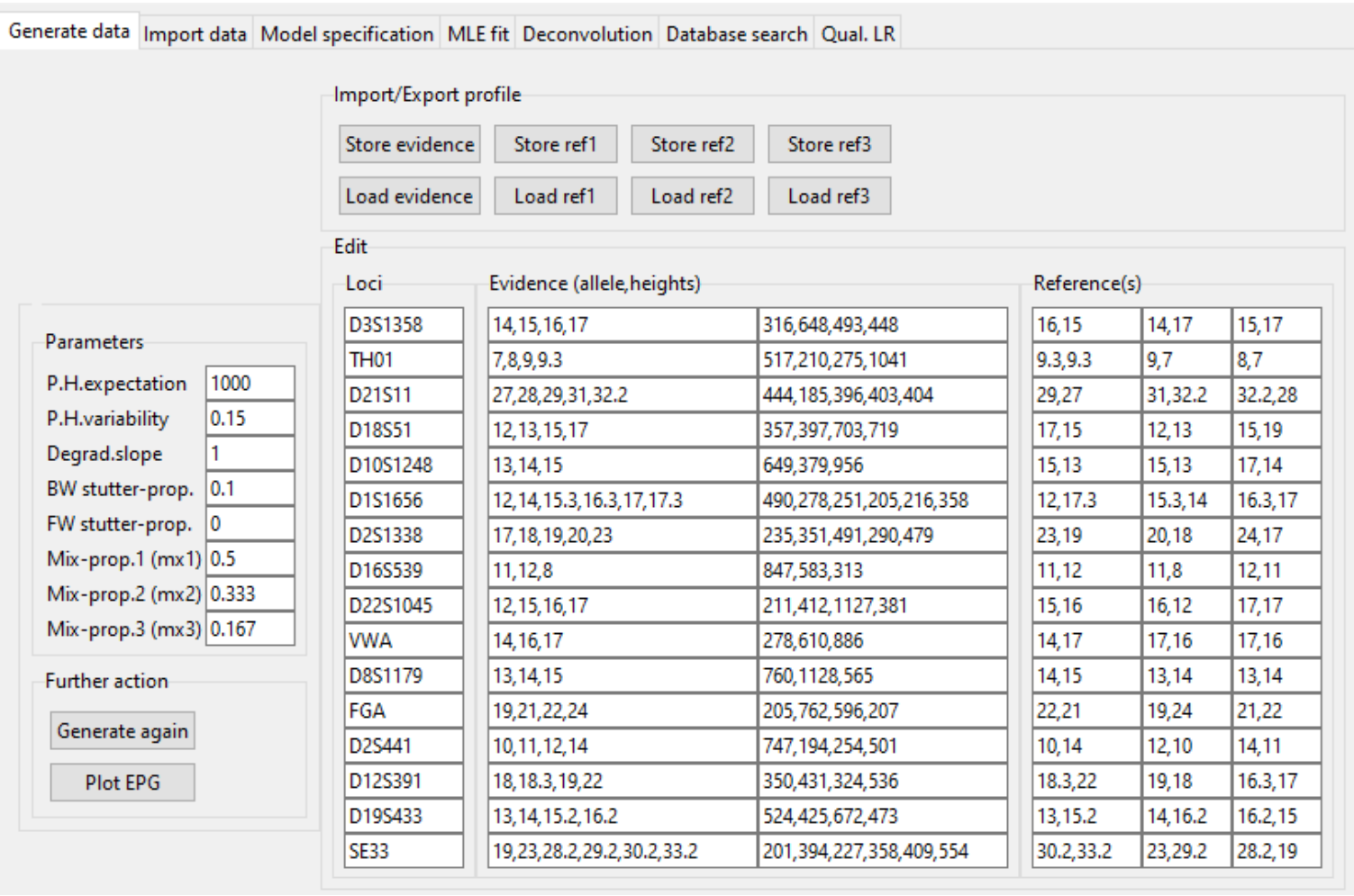
- To generate data, the user must first specify the assumptions (hypothesis and known parameters) in the quantitative model (**Figure 41**). **NB: The relationship module will be ignored!**
- The module will generate alleles using the population frequencies and simulates peak heights for a specified hypothesis using the quantitative model**. NB: Theta/Fst corrections are not used.**
- The generation may simulate allele-dropout, drop-in (with a peak height model), degradation, backward and forward stutters (**Figure 42**).
- Allele-dropout is indirectly simulated if the peak height is below the defined threshold.
- The marker specific settings (analytical thresholds, drop-in model) are used if specified.
- The generate data panel is also very useful as an editor for editing profiles (by loading existing profiles into the panel)

8.1 Parameters

- **P.H.expectation**: Expectation of the peak height for a single heterozygote (Mix-prop=1) allele without degradation
- **P.H.variability**: Coefficient of variation of the peak height for a single heterozygote (Mix-prop=1) allele without degradation
- **Degrad.slope**: A global parameter related to the degree of degradation as a function of fragment length (**kit must be selected**). Value 1 is no degradation, and lower values as for instance 0.6 is much degradation. Default is 1.
- **Backward stutter-prop:** A global parameter related to backward stutter proportion. The expected fraction of peak height that are stuttered from 'n+1' parent allele.
- **Forward stutter-prop:** A global parameter related to forward stutter proportion. The expected fraction of peak height that are stuttered from 'n-1' parent allele.
- Mix-prop. C**k** (mx**k**): Mixture-proportion for contributor '**k**'. Default is $(K:1)/K$, with *K* number of contributors.
    - Note: **mx** will be normalized if it's not already.

8.2 Edit

- **Loci**: Loci name of the population frequency used to generate the dataset.
- **Evidence**: The allele information is given in the left column while the peak height information is given in the right column. Each element **needs to be** separated with ",".
- **Reference**: The alleles of the true contributors to the generate evidence is sequentially shown in each column.
- All the loci names, evidence-allele and heights and reference-alleles may be edited before storing (**Figure 42**).



Loci	Evidence (allele,heights)		Reference(s)		
D3S1358	14,15,16,17	316,648,493,448	16,15	14,17	15,17
TH01	7,8,9,9.3	517,210,275,1041	9.3,9.3	9,7	8,7
D21S11	27,28,29,31,32.2	444,185,396,403,404	29,27	31,32.2	32.2,28
D18S51	12,13,15,17	357,397,703,719	17,15	12,13	15,19
D10S1248	13,14,15	649,379,956	15,13	15,13	17,14
D1S1656	12,14,15.3,16.3,17,17.3	490,278,251,205,216,358	12,17.3	15.3,14	16.3,17
D2S1338	17,18,19,20,23	235,351,491,290,479	23,19	20,18	24,17
D16S539	11,12,8	847,583,313	11,12	11,8	12,11
D22S1045	12,15,16,17	211,412,1127,381	15,16	16,12	17,17
VWA	14,16,17	278,610,886	14,17	17,16	17,16
D8S1179	13,14,15	760,1128,565	14,15	13,14	13,14
FGA	19,21,22,24	205,762,596,207	22,21	19,24	21,22
D2S441	10,11,12,14	747,194,254,501	10,14	12,10	14,11
D12S391	18,18.3,19,22	350,431,324,536	18.3,22	19,18	16.3,17
D19S433	13,14,15.2,16.2	524,425,672,473	13,15.2	14,16.2	16.2,15
SE33	19,23,28.2,29.2,30.2,33.2	201,394,227,358,409,554	30.2,33.2	23,29.2	28.2,19

***Figure 42**: The figure shows the Generate data page which shows the generated alleles and corresponding peak heights (under **Evidence**) for the given selected set of parameters under **Parameters**. The true contributors are given under **Reference(s)**, where first reference is 'P1' and the second is randomly generated based on the allele frequencies.*

8.3 Import/Export

- **Save data**:
  - Stores the generated (and possible edited) evidence- or reference-profile to a file.
  - Extension .csv added automatically.

- **Load data:**
  - Loads profiles from file into the selected entries (evidence or reference).
  - This is useful for generating random evidence samples where loaded references are conditioned on.
  - If any locus is missing from the loaded evidence or reference file, the edit-cell will be empty.
  - The order of the loci in the file does not matter.

8.4 Further action

- **Generate again:** Make a new simulation of the evidence sample using the selected values of the parameters under *Parameters*.
- **Plot EPG**: Plots the generated evidence together with References in an EPG-plot.
  - It will use the “kit” selected under “Import Data”-page.

9 Special for MPS data

In this section we highlight the format of the alleles for MPS based methods.

- Kit selection: Should not be done (NONE can be selected), except for when an MPS kit (e.g. *ForenSeq*) can be selected and the data is in RU/LUS/LUS+ formats (see sections 9.2.2-9.2.4).
    - Degradation model will be deactivated when no kit is selected.
- Visualization: Bar plots are shown for MPS data (View evidence / 'Show P.H. expectation'). The visualization of replicates will be shown in separate plots.

9.1 SNP format

The alleles are named 'A', 'T', 'C', 'G' letters (strings).

Stutters and degradation models are deactivated.

9.2 STR formats

The alleles are strings (not necessary for the RU format).

Stutters are still possible for the RU, LUS or LUS+ formats (see sections 9.2.2-9.2.4). The degradation model is also possible if an MPS kit is selected (e.g. *ForenSeq*).

9.2.1 Full sequence

The 'full sequence' alleles are represented as strings (example: **ATCGATCGATCGATCGATCG**).

The symbol ':' can be used to group alleles in visualization (View evidence / 'Show P.H. expectation'). This has no impact on the calculations.

Example:

The alleles **5:ATCGATCGATCGATCGATCG** and **5:TTGATTGATTGATTGATTGA** will be grouped together in the bar plots.

9.2.2 Repeat Unit format (RU)

The alleles are named as for CE based data which can be converted to numbers.

9.2.3 Longest uninterrupted sequence (LUS)

A separator "_" (underscore) is used to separate RU and LUS information.

Example: '5_5' and '5_1' has same RU, but different variant.

Notes:

- The RU information is used to group the alleles.
- The RU information is used to obtain base pair information in getKit (if kit is specified)
- The same number of underscores must be used for all alleles within a marker.

9.2.4 Longest uninterrupted sequence (LUS+) extended

Same as for LUS but may contain (several) additional "_" separators.

Example: '5_5_1' and '5_5_0' has same RU and LUS, but a third sequence repetition variant is included.