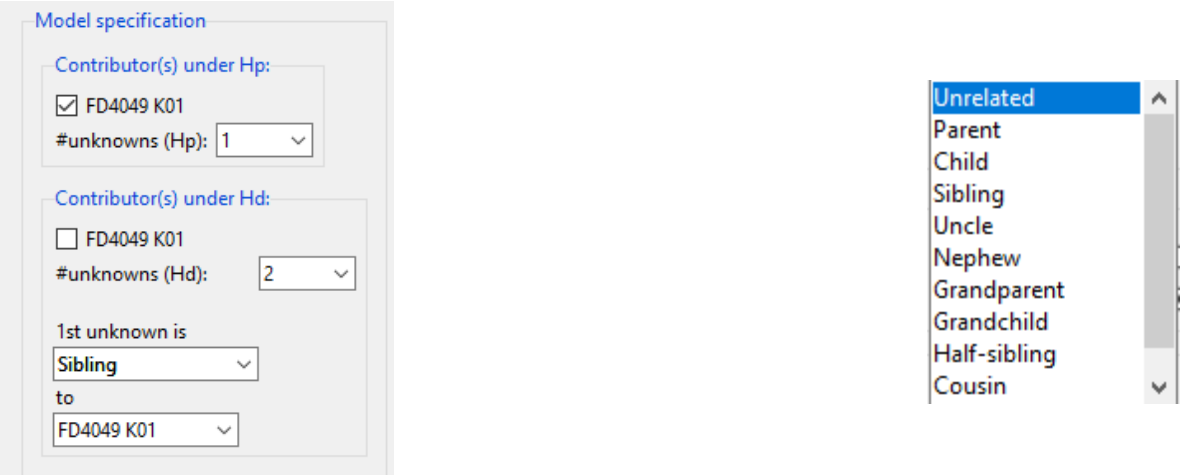
Important changes to EuroForMix (EFM) v2.0 update

3 important changes are introduced:
1) Introduction of the kinship module under Hd.
2) Improved speed when turning on the stutter-model
3) The posterior probability of unknown genotype(s) in deconvolution is generalized

Users must indicate which version of EuroForMix they use when referring to the software (e.g. publications or reports).

1) Kinship module
The user can now specify relationships between an unknown contributor (under Hd) to a typed reference:



The figure shows how the user can specify the relationship between the 1st unknown and a typed individual under Hd. The most typical relationships are defined.

Defined relationships are now taken into account for the following features:
- ML based LR calculations (also database searching)
- INT based LR calculations (also database searching)
- Non-contributor tests: the model simulates random *1st unknown* individuals specified under the Hd hypothesis. This means that the simulated individual has the relationship as specified under Hd.
- LR sensitivity
- Report
- MCMC simulations
- Deconvolution
- Model validation
- Model fitted peak heights (uses deconvolution)

Notice:
- Theta/fst-correction is taken into account for the kinship module. This means that the model will enable the possibility to both specify a relationship and a fst value greater than zero.
- A new function called **calcGjoint** has been added to the euroformix R-package to calculate the probability of the jointly genotype outcome of general X unknowns, where the 1st unknown can have a specified relationship to a reference (through IBD vector). The user can also specify the alleles of the known non-contributors and the fst value. This function is important for numerical validation.

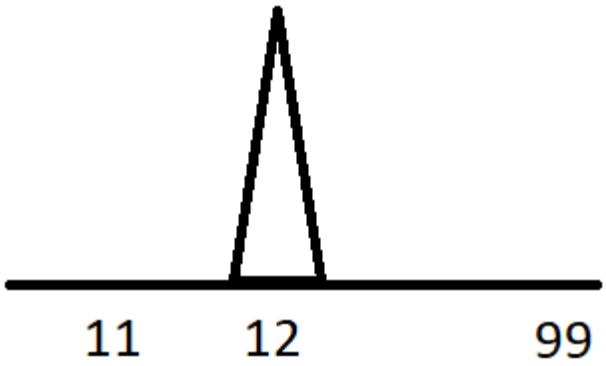
2) Improved speed of stutter model.
A small adjustment is done when the stutter model is turned on:
This difference between this and earlier versions concerns whether unknown (and possibly known) contributors can have **unobserved** potential (backward) stutters in their genotype outcome or not. The alternative is that the Q-designated allele “99” is used instead.

In earlier EFM versions the unobserved potential stutter alleles was part of the genotype outcome of the contributors. It was given a minimimum allele freq if allele not observed in the allele frequency.

In new versions (from v2.0.0), EFM does not include unobserved potential stutter alleles as part of the genotype outcome of the contributors. However these are still part of the allele outcome to take into account that each contributing allele (which is observed) can give a potential stutter.

Example:

Observed allele is 12 (see figure below). Allele 11 is then an unobserved potential stutter allele of 12. The version change concerns whether a contributor in a specified hypothesis (known or unknown) can have allele 11 or not (i.e. the contributor has allele “99”). Both versions take into account that a contribution at allele 12 leads to stutter effect to allele 11, and that this contribution can also lead to a dropout. However the dropout probability may be slightly different for some situations. Both allele 11 and 99 are concerned with allele dropout. There will be a difference in the situation where there is a significant amount of stutter-contribution to allele 11 (from allele 12) and the person of interest (or other known contributors) has allele 11 i.e. it is a minor. Version 1 will give a smaller allele dropout probability, compared to version 2 because of the added stutter-contribution (since the gamma-distribution will be have more mass to higher RFUs).



Assume that we calculate the likelihood of observations when assuming one unknown. The genotype outcome of the unknown is different for the two versions:
v1: Genotype outcome = [11/11 , 11/12, 11/99, 12/12 , 12/99, 99/99]
v2: Genotype outcome = [12/12 , 12/99, 99/99]

Version v2 reduces the set of the genotype outcome by considering allele 11 as part of allele 99. To obtain identical results between v1/v2 the dropout probability of 11 and 99 must be the same.

Additional factors that may create small differences:

- When the degradation model is considered, the (bp) position of allele 11 and allele 99 is different (the base pair of allele 99 is defined as the maximal base pair of the corresponding marker). The greater the degradation the bigger difference in the dropout probability.
- The potential frequency normalization after including allele 11 to freq-database with minimum frequency (this is only done in v1).

3) The posterior probability of unknown genotype(s) in deconvolution is generalized

Theta-correction and relationship is now also taken into account.

The deconvolution formula, for predicting the joint genotype G of the unknown(s) is given as

P(G|Evid) = P(Evid|G)*P(G)*c , where c is a constant such that the sum of all P(G|Evid)=1.

Here P(G) is the prior genotype probability based on the allele frequency and the hypothesis given.

EFM (before v2.0.0) assumed HW for the probability of unknown genotype part p(G), whereas the updated version (v2) generalizes P(G) so that fst/theta-correction/relationship are taken into account.