Version changes:

v2: Details about the Negative Binomial model

## Mathematical formulation of the MPSproto model

### An extension of the EuroForMix model

We let the model for coverage reads to be an extension of the EuroForMix model [1]. We assume that the observed sample, which contains markers in total, is composited by *K* individuals, each contributing with a mixture proportion parameter , . The read for allele at marker is distributed as a gamma distribution with following shape and scale arguments:

such that the probability density function of read , given genotypes and model parameters , is defined as

Here is the {0,1,2} genotype contribution for individual at allele in marker , which is decided by the assumed genotype for the contributors, . The parameters and are the expectation and coefficient-of-variation of a full heterozygous contribution read, whereas is the marker amplification efficiency parameter at marker . We restricted the parameters such that and .

### The dropout model

The drop-out model is modelled as in EuroForMix, needed when the shape argument for allele is positive and the corresponding read is not observed (i.e., :

where is the analytical threshold at marker We used the compound Q-allele (“99” in program) as the drop-out allele, representing all the non-observed alleles from the allele outcome (defined from the allele frequency information).

### The degradation model

The degradation is modelled as in EuroForMix where the shape argument is scaled with ,where is the degradation slope parameter and is the fragment length of allele at marker .

### The stutter model

The EuroForMix model was extended to include all kinds of stutters (of type ) by modifying the “un-stuttered” shape argument such that the modified shape argument for allele can be expressed as

where is the expected stutter proportion of stutter type for allele . For a particular stutter type , ” is the allele which allele provides **stutter to**, whereas is the allele which allele obtains **stutter from**. An example of this is when and a “n-1” stutter type is applied: Then =4 is the **stutter to** allele and =6 is the **stutter from** allele.

The expected stutter proportion of stutter type was modelled with the logit-link function providing

where the linear predictor contains the regression coefficients , and is the block length of missing motif (BLMM) for stutter and is the BLMM for a different motif also for stutter . This describes n0 stutters where two parts of the sequence have simultaneously stuttered. A coefficient is set to zero if the corresponding part is not modelled.

### The noise model

We also defined a model for alleles not originating from any of the assumed contributors or stutters (i.e., when 0 and ). Instead of defining a drop-in model we instead defined a noise model as two parts: A) The number of noise alleles and B) The read size of the noise alleles.

1. The number of noise alleles, , follows a geometric distribution with parameter :
2. The read count of the noise allele, is proportional to a pareto distribution with parameter :

Where T is the analytical threshold. We included an upper limit cap of reads for noise, such that the normalizing constant is calculated as

Importantly, the noise parameters, and , are marker specific and must be decided based on a “noise dataset”, i.e., alleles belonging to any of the contributors or being a stutter.

## Calibration of the MPSproto model

The following model parts in the MPSproto tool must be inferred before using it on mixtures data.

1. The stutter model ( all stutter types at marker)
2. The noise model ( and for markers )
3. The marker efficiency parameters ( for markers )

Inferring these parts must be based on single source profiles where the alleles of the donor are known/indicated. The stutter model (part 1) was inferred based on beta regression models as part of paper [2], so we only describe part 2-3 here:

### Part 2: Calibration of the noise model

After that different stutter types have been identified as described in part 1, it is useful to get an overview of remaining alleles that were not indicated as stutter (these are considered as “noise”).

A) For each marker we obtained the number of noise per sample as . We always appended an additional observation to make the model more robust for new observations. The parameter of the geometric distribution was estimated as , where is the mean read count of noise.

B) For each marker we obtained read count of the noise and these were vectorized as . We always appended an additional observation , the analytical threshold to make the model more robust for new observations. If (no observations) we append two more observations being , enabling the model to be fitted.

### Part 3: Calibration of marker efficiency

For each sample , the sum of reads per marker is calculated as . From the model (assuming no allele dropout), it follows that sample specific parameters for and

If , then this is a marker dropout; the following drop-out model was used instead:

We estimated the marker efficiency parameters based on maximizing the likelihood

From this the maximum likelihood estimates of were obtained and inserted into the MPSproto model.

## Inference of mixtures with MPSproto model

We follow a maximum likelihood framework to estimate the unknown parameters in the mixture model. After calibration, we only need to estimate the parameters , and these are calculated by optimizing the likelihood function for the mixture data under a specified hypothesis H defined as

where for a given marker , is the mixture information (alleles and reads), and is the joint genotype combination of contributors, traversed through the joint genotype outcome defined from hypothesis .

A likelihood ratio can be calculated by comparing two hypotheses, for instance:

* H1: “Ref1+Ref2 are contributors”
* H2: “Ref1+1 unknown are contributors”

such that the likelihood ratio is calculated as:

### Using the Negative Binomial distribution

We implemented the Negative Binomial model as described by Vilsen et al. (2018) [3], by modifying the shape argument to instead be the expectation:

Additionally, we let the size parameter be parameterized from (expectation) and (coefficient-of-variation):

such that the probability density function of read , given genotypes and model parameters , is defined as

where the first and second argument of Negative Binomial (NB) is mean and size.

The stutter and noise models are the same as previously defined, but where is replaced by in the stutter product expression.

NB is defined on a discrete outcome whereas the gamma distribution is defined on a continuous outcome. Hence the drop-out model for NB (evaluated when ) is based on sums instead of integrals such that

## References

[1] O. 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 (2016) 35-44.

[2] A comprehensive characterization of STR-MPS stutter artefacts (submitted paper).

[3] S. Vilsen, T. Tvedebrink, P.S. Eriksen, C. Hussing, C. Børsting, N. Morling; Modelling allelic drop-outs in STR sequencing data generated by MPS; Forensic Sci Int Genet 37 (2018) 6-12.