Forensic Science International: Genetics 42 (2019) 31–38

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

An assessment of the performance of the probabilistic genotyping software EuroForMix: Trends in likelihood ratios and analysis of Type I & II errors

Corina C.G. Benschop*, Alwart Nijveld, Francisca E. Duijs, Titia Sijen

Netherlands Forensic Institute, Division of Biological Traces, Laan van Ypenburg 6, 2497GB The Hague, the Netherlands

ARTICLE INFO

Keywords:
Mixed STR profiles
Low-template DNA
EuroForMix
LRmix Studio
Likelihood Ratio
PowerPlex® Fusion 6C

ABSTRACT

Continuous probabilistic genotyping software enables the interpretation of highly complex DNA profiles that are prone to stochastic effects and/or consist of multiple contributions. The process of introducing probabilistic genotyping into an accredited forensic laboratory requires testing, validation, documentation and training. Documents that include guidelines and/or requirements have been published in order to guide forensic laboratories through this extensive process and there has been encouragements to share the results obtained from internal laboratory studies. To this end, we present the results obtained from the quantitative probabilistic genotyping system EuroForMix applied to mixed DNA profiles with known contributions mixed in known proportions, levels of allele sharing and levels of allelic drop-out. The mixtures were profiled using the PowerPlex® Fusion 6C (PPF6C) kit. Using these mixtures, 427 Hp-true tests and 408 Hd-true tests were performed. In the Hd-true tests, non-contributors were selected deliberately to a have large overlap with the alleles within the mixture and worst-case scenarios were examined in which a simulated relative of one of the true donors was considered as the person of interest under the prosecution hypothesis. The effects of selecting different EuroForMix modelling options, the use of PCR replicates, allelic drop-out, and varying the assigned number of contributors were examined. Instances of Type I and Type II errors are discussed. In addition 330 likelihood ratio results from EuroForMix are compared to the semi-continuous model LRmix Studio. Results demonstrate the performance and trends of EuroForMix when applied to PPF6C profiles.

1. Introduction

It has been more than a decade since the International Society for Forensic Genetics advised to use the likelihood ratio (LR) approach for the statistical evaluation of mixed forensic DNA samples [1]. Over the years, evaluation of the weight of evidence of mixed autosomal DNA profiles has progressed from using binary models, to semi-continuous (or qualitative) models and to continuous (or quantitative) models. Probabilistic methods that take account of stochastic effects such as allelic drop-out and allelic drop-in enable LRs to be computed for complex and low template DNA profiling results. Currently, several software packages implementing either a semi-continuous [2–6] or continuous [7–16] probabilistic LR model are available and in use by many forensic DNA laboratories. The process of introducing probabilistic genotyping software into an accredited forensic laboratory primarily requires, testing, validation, documentation and training. Documents that include guidelines and/or requirements have been published in order to assist forensic laboratories in this extensive process [17–22]. Presenting data from internal validation studies (like those in, for example [23,24],) helps to gain insight into ranges of LRs that can be expected in prosecution hypothesis (Hp)-true and defence hypothesis (Hd)-true tests and accordingly, to assess the risk of Type I errors (LR $<$ 1 for a true contributor,) or Type II errors (LR $>$ 1 for a non-contributor). This is useful to define guidelines for application in forensic casework and can be helpful for court going purposes.

To this end we present the results of a study examining the performance of the open source, continuous software package EuroForMix applied to two- (2p), three- (3p) and four-person (4p) mixtures amplified using the PowerPlex® Fusion 6C (PPF6C) kit. EuroForMix was first introduced in 2015 and has evolved over time: user friendliness was improved, bugs were fixed, calculation time was decreased and new functionalities were added. The modelling within EuroForMix accommodates peak height, allelic drop-in, allelic drop-out, degradation and stutter [7]. In addition, the LR calculation includes the allowance for population substructure.

* Corresponding author.
E-mail addresses: c.benschop@nfi.nl (C.C.G. Benschop), a.nijveld@nfi.nl (A. Nijveld), f.duijs@nfi.nl (F.E. Duijs), t.sijen@nfi.nl (T. Sijen).

https://doi.org/10.1016/j.fsigen.2019.06.005
Received 11 April 2019; Received in revised form 7 June 2019; Accepted 7 June 2019
Available online 10 June 2019
1872-4973/

Table 1
Overview of the six donor combinations used for mixture preparation.

Dataset number	Type of dataset	Number of contributors			
		2	3	4	5
		Donor combinations per dataset			
1	High allele sharing	a:b	a:b:c	a:b:c:d	a:b:c:d:e
2	Low allele sharing	f:g	f:g:h	f:g:h:i	f:g:h:i:j
3	Random	k:l	k:l:k	k:l:k:n	k:l:m:n:o
4	Random	p:q	p:q:r	p:q:r:s	p:q:r:s:t
5	Random	u:v	u:v:w	u:v:w:x	u:v:w:x:y
6	Random	z:aa	z:aa:ab	z:aa:ab:ac	z:aa:ab:ac:ad

2. Materials and methods

2.1. Sample selection and DNA mixtures creation

Six sets of five DNA extracts were selected from a large sample set of extracted DNA from 2085 Dutch males [25]. Two of the sets (dataset 1 and 2) were selected based on their genotypes whilst the remaining datasets were selected essentially at random. Dataset 1 was chosen to have low numbers of alleles by selecting multiple homozygous loci and/or by selecting combinations of donors with many shared alleles between them. In contrast, dataset 2 was chosen to maximise the number of alleles by selecting combinations of donors with many unshared alleles between them (see Table 1).

Prior to mixture preparation, all DNA extracts were quantified using an ALU assay that has a lower detection limit of 0.5 pg/μL human DNA [26,27]. For each of the six datasets (*i.e.* donor combinations), 20 different mixtures were prepared according to Table 2.

2.2. STR profiling

The 120 mixed DNA extracts (six datasets (Table 1), each consisting of 20 mixtures (Table 2)) were amplified in triplicate using the PowerPlex® Fusion 6C (Promega, PPF6C) short tandem repeat (STR) typing system. PCR amplification was performed on an TAdvanced PCR apparatus (Biometra) following the manufacturer's recommendations, except for a reduced total PCR volume of 12.5 μL. This reduced PCR volume is standard in our laboratory and showed, during an in-house validation study, a larger percentage of detected alleles and larger peak heights with slightly larger variation but similar heterozygote balance when compared to 25 μL PCR volume. In each PCR plate, two positive control samples (2800 M Control DNA, Promega) and two blanks (water) were included. The total DNA input into the PCR varied between 180 and 900 pg, depending on mixture type (see Table 2).

Capillary electrophoresis (CE) was performed on an ABI3500xL (Thermo Fisher) using per run 9.6 μL HiDi formamide, 0.4 μL WEN ILS 500 and 1 μL of PCR product or allelic ladder. Injection parameters were 1.2 kV for 24 s and run parameters were set at 13 kV for 1500 s.

In total, 360 PPF6C profiles were obtained and analysed using GeneMarker® HID v2.9.0 software. The following settings were utilised: Auto Range Raw data analysis; Smooth + Superior smoothening; Cubic Spline; Auto Range Allele Call; 2% heterozygote imbalance), using dye specific detection thresholds (FL = 45, JOE = 50, TMR = 45, CXR = 80, TOM = 40 relative fluorescent units (RFU)) and in-house determined locus specific stutter ratios.

Sample nomenclature was as follows. In a mixture denoted, for example,1A2.1, the first number represents the dataset (1–6), the letter corresponds to the mixture type (A–E), the second number denotes the number of contributors (2–5) and the number after the dot represents each PCR replicate number (.1, .2 or .3). During analysis, a total of six samples (1E4.1, 1E5.1, 3E5.1, 4E3.1, 5E3.2, 6E5.1) that did not meet the required quality criteria (*i.e.* poor/broadened peak shapes) were removed from the sample set leaving 354 mixed profiles.

2.3. Likelihood Ratio calculations

The software package EuroForMix was used for model selections, MLE LR calculations and sensitivity analyses. Various software versions were used (*i.e.* v1.9.1 up to v1.11.4) as updated versions showed improved user friendliness and decreased calculation times (further details on version changes can be found at [28]). These version changes do not alter the outcomes obtained in this study.

Model selection was performed for all 2p and 3p mixtures (n = 60) in Tables 1 and 2. We followed the same model selection protocol as outlined in [29]. This is an iterative process. First likelihoods are computed using a standard model only. Second, the task is repeated with the degradation (DG) model turned on. Third, the DG model is turned off and the stutter (ST) model turned on instead. Fourth, both the degradation and stutter (DG&ST) models are jointly turned on. Next, the model selection was based on the Akaike information criterion (AIC), where the computed log likelihoods under Hd were penalized with respect to model complexity: The model that was selected is the one yielding, under Hd, the largest 'log-likelihood minus the relative number of parameters in the model' (this is equivalent to AIC). Here the relative number of parameters was zero for the standard model and then it is added by one for DG or ST and by two for DG&ST. For examples and further details on model selection in EuroForMix see [30].

LR calculations were performed for 2-4p mixtures using the MLE function in EuroForMix. These calculations were performed using the DG model turned on and the ST model turned off. The parameters used in the software were default, except for a 40rfu detection threshold, drop-in rate of 0.0043, lambda of 0.018, and an Fst value of 0.01. In all LR calculations, the allele frequencies were taken from a Dutch population database [25]. Subsequent to each LR calculation, a model validation was performed. In this, the cumulative probabilities for the expected peak height were plotted against the observed peak heights, resulting in a PP (probability-probability) plot [7]. The default significance level of 0.01 was used in EuroForMix and when at least three values were outside the 0.01-line, the model validation was scored as

Table 2
Mixture proportions and amounts of DNA used per donor to create a total of 20 different mixtures per dataset.

Mixture Type	Number of contributors			
	2	3	4	5
	Picograms DNA per contributor			
A: major 2x more than any minor	300:150	300:150:150	300:150:150:150	300:150:150:150:150
B: major 10x more than any minor	300:30	300:30:30	300:30:30:30	300:30:30:30:30
C: 2 majors with equal amount	150:150	150:150:60	150:150:60:60	150:150:60:60:60
D: major 5 to 2.5x more than minors	150:30	150:30:60	150:30:60:30	150:30:60:30:30
E: major 20 to 10x more than minors	600:30	600:30:60	600:30:60:30	600:30:60:30:30
Number of mixtures	5	5	5	5

'failed'. If the model validation failed due to deviating peak heights between replicates, the simultaneous analysis of the three replicates in EuroForMix was disregarded and omitted from the dataset.

Initially, 378 Hp-true tests and 180 Hd-true tests were performed using EuroForMix. Supplementary Table 1 presents the samples and donors selected for the Hp-true tests ($n = 56$, $n = 164$ and $n = 158$ using 2p, 3p and 4p mixtures, respectively). Hd-true tests were performed using mixture types B, C and E as for these it is expected that it will be most difficult to exclude non-contributors. Specifically, mixtures type B and E exhibit most allelic drop-out (due to the 30 pg contributions) whereas type C mixtures contain major contributions in similar amounts meaning that it is difficult to differentiate between contributors. True-genotype non-contributors were selected from a dataset of 2085 males [25] by comparing their genotypes to the B, C and E mixtures. Non-contributors were selected from this dataset of 2085 males [25] by comparing their genotypes with the mixtures. In particular, genotypes exhibiting high numbers of shared alleles were chosen in an effort to explore the relationship between the LR and the number of unseen alleles. An overview of the number of non-contributors per number of unseen alleles and mixture type is presented in Supplementary Table 2A ($n = 180$).

In addition, Hp-true tests were performed using an under-assigned number of contributors (NOC) for those mixtures that would yield an underestimation based on the maximum allele count method (*i.e.* the maximum number of alleles at a locus divided by two and rounded up to the nearest whole number). These were six 4p, and 17 5p mixtures, for which once a major and once a 30/60 pg minor contributors was regarded as POI ($n = 44$ LRs as two samples did not have a 30/60 pg contributor). Furthermore, a subset of the Hd-true analyses using 2p or 3p mixtures was re-analysed using an extra contributor under the hypotheses. These comprised 44 analyses with 2p mixtures regarded as 3p mixtures and 57 3p mixtures analysed as 4p mixtures (Supplementary Table 2B).

As well as using non-contributors with low numbers of unseen alleles, we simulated the genotypes of brothers and fathers/sons of true contributors (both $n = 100$ simulations per true contributor). Profiles of relatives were simulated as follows. For a given type of relatedness, let kappa_i be the probability that individuals related in this way share i alleles identical by descent (ibd) on an autosomal locus. To obtain a simulated profile of a relative of a reference profile, these coefficients were used to generate, on each locus independently, a number i of ibd alleles with the reference profile and then we chose i alleles of the reference profile to include in the profile of the simulated relative. The remaining 2-i alleles were obtained by independent sampling according to the population frequencies [25]. In other words, sampling was performed as if the population is in Hardy-Weinberg equilibrium and all loci are independent. The genotypes of those simulated relatives with the highest degree of overlap with the target mixture were selected as these were regarded as being most challenging ($n = 108$, see Supplementary Table 2CD).

Hp-true tests were performed using the individual replicates as well as simultaneous analysis of the three replicates (when replicates were available). Hd-true tests that resulted in Type II errors with an LR $>$ 10 for a non-relative or an LR $>$ 100 for a relative as POI under Hp were recalculated using three replicates ($n = 19$). These include analyses with the true NOC under the hypotheses.

In EuroForMix the likelihoods are maximised separately under both hypotheses using the parameters that best explain the data (single points in the parameter space). As uncertainty regarding this parameter estimation is not taken into account in the MLE calculation, EuroForMix contains an option to perform sensitivity analysis that considers the LR as a function of the parameters involved [29]. Sensitivity analysis was performed for 42 sets of Hp-true hypotheses with 2p mixtures and 84 sets of Hp-true hypotheses with 3p mixtures. In addition, Hd-true tests (non-relatives) that resulted in an LR $>$ 10 were used in the sensitivity analysis. The lowest 5% LR values were recorded and compared to the LRs obtained using the MLE calculation.

In parallel, LRmix Studio v2.1.3 was applied to all Hp-true tests using one replicate ($n = 272$) and for all Hd-true tests where EuroForMix yielded an LR $>$ 10 with a non-relative as POI ($n = 2$) or an LR $>$ 1 with a relative as a POI ($n = 30$). In addition, a selection of 26 Hd-true tests that yielded an LR $<$ 10 using EuroForMix was subjected to LRmix Studio calculations. This selection varied for the NOC, mixture type and number of unseen alleles for the POI. A drop-out estimation was performed and the drop-out value yielding the lowest LR within the most plausible range of drop-out was recorded, with a minimum drop-out value of 0.01. As with EuroForMix, the probability of drop-in was set to 0.0043 and an Fst value of 0.01 was applied.

3. Results and discussion

3.1. Model selection in EuroForMix

When EuroForMix is used, it is advised to perform model selection in order to infer which model best explains the data [30]. As model selection can be a time-consuming process and may not always be required, the effects of using the degradation and/or stutter model were examined. The time usage with EuroForMix heavily depends on *a.o.* the number of unknown contributors, the number of alleles in the evidence profile, the number of replicates and the computation power. Presented time usage varies from seconds to several days [30]. In our experience, the use of the stutter model more than doubles the time and with the most complex samples (four unknowns and three replicates) this can take up to several weeks. The DNA-profiles used in this study were pre-filtered for stutter using functionality insight the GeneMarker® HID software. Hence, the stutter modelling was likely to be redundant. In-house validation of the PPF6C kit showed that the stutter filter algorithm in GeneMarker® HID performed adequately in removing stutter but that elevated stutters could be present when PCR template levels were low (data not shown). For the 2p and 3p mixtures model selection was performed and for each sample it was recorded which model was regarded to be optimal. In all instances, analyses using the degradation model gave the largest log likelihood under Hd, and therefore provided a better explanation for the data compared to when the degradation model was not applied (Table 3A).

The stutter thresholds that were applied in GeneMarker® HID are locus-specific and take into account -0.5, -1, -2, +0.5 and +1 repeat unit stutter positions. The stutter modelling in EuroForMix only evaluates peaks in the -1 repeat unit position and so cannot model peaks in other stutter positions. We determined how often elevated stutters occurred at the -1 repeat position, including those that are on a position in between two true alleles, and recorded the preferred model (Table 3B). As expected, when more elevated stutter occurred, the DG&ST models were selected more frequently as optimal (Table 3). With three replicates, elevated stutters occurred more frequently considering the three replicates and accordingly, the DG&ST model was selected more frequently than with individual profiles (Table 3).

3.2. Relationship between model selection and LRs

Those 2p and 3p mixtures that yielded the DG&ST as the optimal model ($n = 27$ for individual replicates, $n = 34$ for three replicates) were compared with the DG model alone by computing LRs. The results are presented in Fig. 1 which shows that applying just the DG model in Hp-true tests produced LRs that were either equal to, or lower than, the corresponding LRs calculated with the joint DG&ST models. About two thirds of the LRs were within one factor of 10 (ban1 *i.e.* one unit on log10 scale). The largest difference was for a 3p mixture that yielded ban + 8.7 and ban + 4.8 for three replicates and a single replicate, respectively. Applying the ST model was in favour of the prosecution's hypothesis or had no effect. Most benefit from the ST model was observed for 3p mixtures with three replicates (Fig. 1), consistent with the

Table 3
Results for EuroForMix model selection performed for 2p and 3p mixtures using individual profiles or three replicates. (A) Percentages of analyses that yielded the largest log likelihood under Hd per model and (B) average number of stutters per profile that exceeded the minus one repeat unit stutter filters in GeneMarker® HID (*i.e.* elevated stutters).

A		Percentage optimal model (largest corrected log likelihood Hd)			
Applied model		Individual profiles		Three replicates combined	
Stutter model	Degradation model	2p mixtures (n = 90)	3p mixtures (n = 88)	2p mixtures (n = 30)	3p mixtures (n = 28)
No	No	0%	0 (0%)	0 (0%)	0 (0%)
Yes	No	0%	0 (0%)	0 (0%)	0 (0%)
No	Yes	60 (67%)	70 (79%)	8 (27%)	4 (15%)
Yes	Yes	30 (33%)	19 (21%)	22 (73%)	23 (85%)

B		Average number of elevated stutter peaks per profile			
Applied model		Individual profiles		Three replicates combined	
No	Yes	0.47	0.91	0.00	0.75
Yes	Yes	1.67	2.21	3.23	3.65

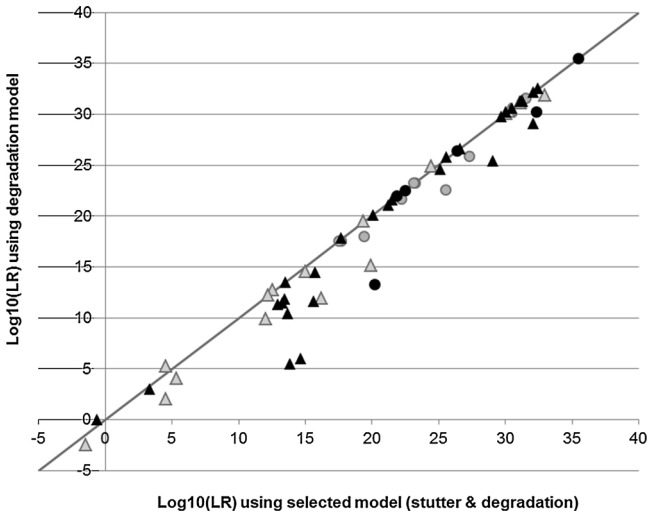

Fig. 1. LRs (log10) for 2p (circles) and 3p mixtures (triangles) with optimal model DG&ST (X-axis) plotted on the X-axis against LRs with DG model only (Y-axis). Grey and black figures represent the results when using one replicate or three replicates, respectively. X = Y is shown by the diagonal line.

finding that the highest number of elevated stutters was observed for these profiles (Table 3). If the ST model is not applied, then some of these elevated stutters would have to be explained by (larger than expected) allelic drop-in events. This will occur more often under Hp that conditions on the POI, and will strongly penalise the probability of the evidence, yielding lower likelihoods and thus a lower LR compared to the DG&ST model.

3.3. LR as a function of allelic drop-out/ unseen alleles

Mixture types B, D and E have contributions as low as 30 pg and, as a consequence, exhibited allelic drop-out of the minor contributor(s). Hp-true tests were carried out using POI's that had 0–16 allelic drop-outs. Similarly, Hd-true tests were carried using POI's for which the numbers of unseen alleles were within this range. The number of Type I and Type II errors were studied and results are shown in Fig. 2. Inspection of Fig. 2 reveals that there was no Type I or Type II error with 2p mixtures. Type I errors were observed for 3p mixtures (with five or more unseen alleles), and for 4p mixtures (with 11 or more unseen alleles) (blue symbols in Fig. 2DF). Type II errors were observed for 3p (with 12 or more unseen alleles) and 4p mixtures (with five or more unseen alleles) (red symbols in Fig. 2DF). The largest LR for a non-contributor in a Hd-true test was 65 for a 3p mixture and 120 for a 4p mixture.

Convergence of results in Hp-true and Hd-true tests is expected with large numbers of allelic drop-outs and unseen alleles respectively for the POI. This was observed for the 3p and 4p mixtures, but not for the 2p mixtures in this study (Fig. 2). Bleka and co-workers presented similar plots [29], though they did obtain convergence of results with 2p mixtures. Differences between the two studies are likely to be the result of the different STR typing kits used by the respective studies. Bleka used the NGM kit with a total of 15 autosomal STR loci [29] whereas we used the PPF6C kit (23 autosomal STR loci). The larger number of loci and the higher discriminatory power of some of the loci (SE33) might be the cause of fewer Type I and II errors for the 2p mixtures in this study.

When comparing the true-positive rate to the false-positive rate as a function of LR threshold it can be observed that false-positives were not obtained with an LR threshold of 1000 that corresponded to a true-positive rate of 96%. An LR threshold of 100 corresponded to a false-positive rate of 0.6% and a true-positive rate of 99% (Supplementary Fig. 1).

3.4. The effect of using a relative of a true donor as POI under Hp

The LRs that were calculated by EuroForMix assumed that, under Hd, the unknown individuals were unrelated to the POI and to this extent relatedness was not taken into account. In casework it is generally unknown whether a relative of the POI is a possible alternative contributor to the mixture. Therefore, we examined the effect on the LR when simulated brothers or father/sons of a true contributor were compared. Worst-case scenarios were examined by using only those simulated relatives that showed a large overlap with the mixture profiles (0–14 unseen alleles). Type II errors were obtained for 0/36 2p mixtures, 11/36 3p mixtures and 23/36 4p mixtures. These Type II errors occurred more often with a simulated brother (n = 21) than with a simulated father/son (n = 13) (Table 4). Nineteen analyses yielded an LR > 100 and these are shown in more detail in Fig. 3. This figure shows that the Type II errors were obtained with a varying number of unseen alleles (0–11). Furthermore, these Type II errors were obtained at least once for each of the datasets (donor combinations) and for each of the mixture types that was used in this study (data not shown). EuroForMix v1.11.4 and lower as used in this study was not designed to take account of relatedness, though in EuroForMix v2.0.2 and higher it is possible to replace an unknown contributor under Hd by a relative of the POI. If in casework relatedness is suspected, this option could be applied to examine how much more likely it is that the POI (and a number of unknowns) rather than a relative of the POI (and a number

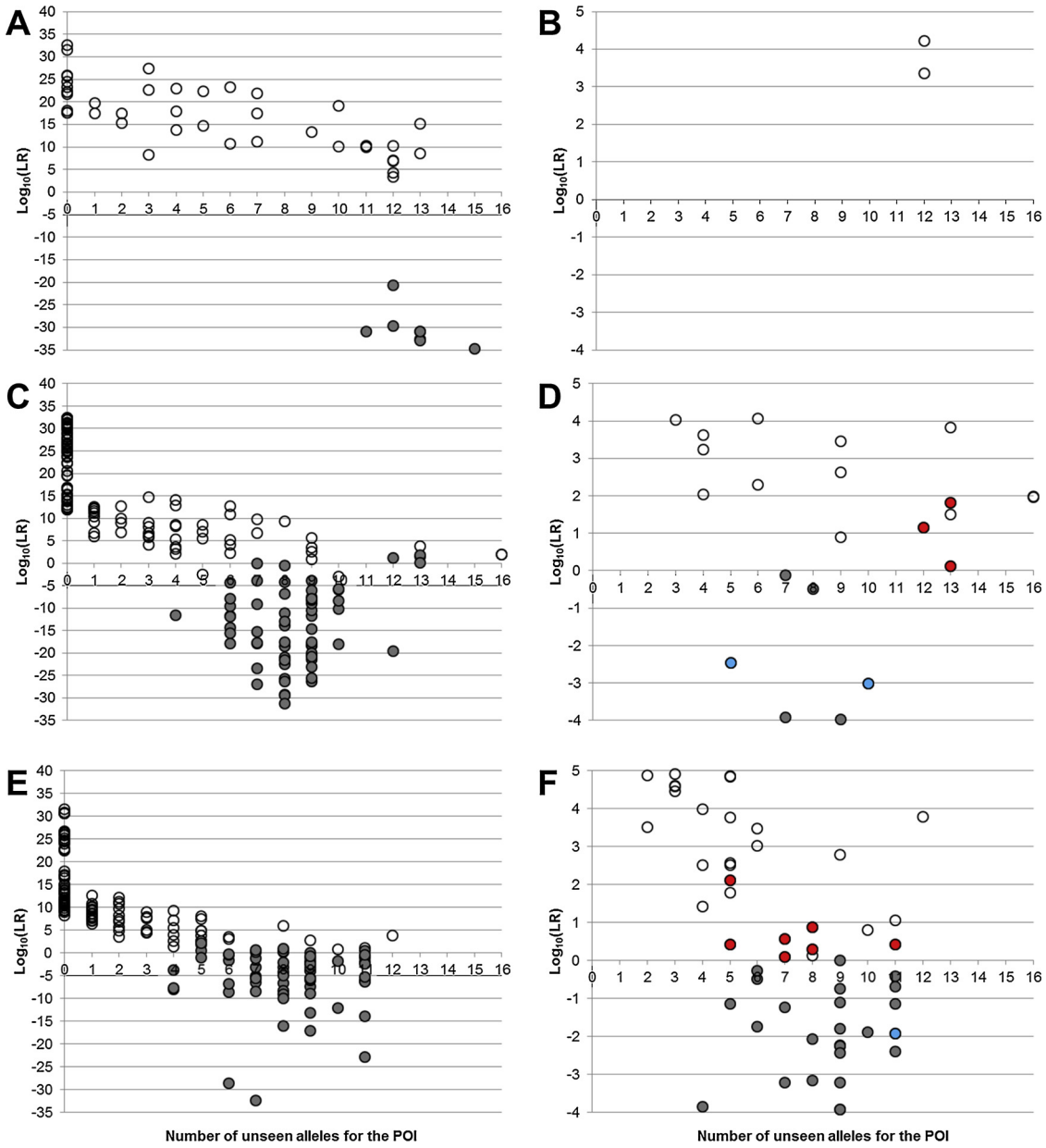

Fig. 2. Log10 LRs for 2p (panels A&B), 3p (panels C&D) and 4p mixtures (panels E&F) computed using Hp-true tests (open circles) or Hd-true tests (grey circles) plotted against the number of allelic drop-outs or unseen alleles for the POI. Plots B, D and F are zoomed on the Y-axis to allow inspection of Type I and II errors (denoted by blue and red circles, respectively) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Table 4
EuroForMix LR results for analyses using a simulated relative of a true contributor as POI under Hp.

NOC	Brother				Father/son			
	LR < 1	LR 1–100	LR > 100	Max. LR observed (rounded values)	LR < 1	LR 1-100	LR > 100	Max. LR observed (rounded values)
2	18	0	0	0.04	18	0	0	0.0007
3	10	2	6	2,800	15	3	0	8
4	5	3	10	64,000,000	8	7	3	220,000

of unknowns) contributed to the DNA sample.

3.5. Variation between replicates and the effect of using multiple replicates for Hp-true tests

In case of low template amounts of DNA one may consider using replicate analyses in order to obtain more information from an evidentiary trace [31–34]. We examined the effect of using three replicates rather than a single replicate of the same DNA extract in Hp-true tests.

Peak heights and the number of detected alleles can vary between replicates of the same DNA extract due to *e.g.* stochastic variation, pipetting variation or differences in the condition of the DNA extract when replicates are not generated at the same time [35]. In this study, the latter is not the case as replicates were generated at the same time, in the same PCR plate and in the same PCR apparatus. EuroForMix uses an optimizer to search for best explaining parameters for mixture

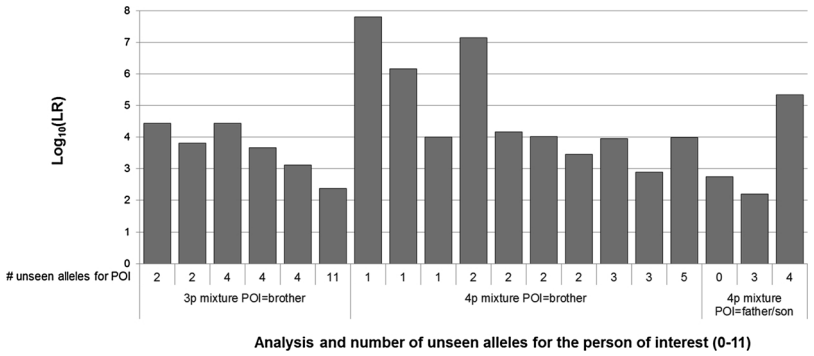

Fig. 3. Overview of the EuroForMix analyses that yielded an LR above 100 with a simulated relative of a true donor as POI under Hp.

proportion, peak height expectation, peak height variance and degradation slope. When using replicates, the model searches for parameters that best explain the replicates together. With differences in peak heights between replicates the model validation may fail (*i.e.* the model's expected peak heights may not follow the observed peak heights), as the optimal parameters are determined on the combination (average) of all replicates and therefore may not fit each individual replicate. This occurred for 9/38 analyses with three replicates of 3p mixtures and for 9/36 analyses with three replicates of 4p mixtures and were omitted from the dataset.

In our data, a maximum difference of six allelic drop-outs between the three replicates were observed for the 2p, 3p and 4p PPF6C profiles. The maximum difference in LRs between replicates of the same DNA extract was ban + 11.5 (Supplementary Table 3). This occurred for a 2p mixture; the replicate with the lowest LR (ban + 8.2) had three drop-outs for the POI and the replicate with the highest LR (ban + 19.7) had one drop-out for the POI. In addition, the replicate with the lowest LR showed an elevated +/-1 stutter peak that was 11% of the peak height of the parent allele. The LR at this particular locus was very low as the peak could not be assigned to the major contributor (the stutter model was not applied) and the height was too high to be adequately explained as drop-in. Without this locus a maximum difference of ban + 5.6 was observed between the three replicates (ban + 13.1 *vs* + 18.7). Although this difference was still very large, using an upper reporting threshold of *e.g.* a million, this would not result in a different statement in the casework report. The variation and the effect of one or three replicates are more interesting in the region where the LR is below such a threshold.

Fig. 4 shows that, in 86.7% (182/210) of the analyses the use of three replicates in Hp-true tests resulted in a larger LR when compared to the use of one replicate. The largest increase was ban + 13.1. In 29.5% of the analyses the LR was within one ban and in 6.7% of the analyses the LR was more than one ban lower with the use of three instead of one replicate. The largest decrease was ban-4.1.

When using two replicates (which is common in some laboratories) instead of three replicates a similar trend is expected. Indeed LR values were most often larger (11/14 analyses using 2p mixtures) with the two replicates jointly than when using one replicate (data not shown).

3.6. The effect of replicates on the LR for Hd-true tests

The Hd-true tests from Supplementary Table 2A that yielded an LR > 10 when using one replicate ($n = 3$) were subsequently analysed in EuroForMix using three replicates, if available ($n = 2$). These represent one 3p and one 4p mixture. Also, the analyses using a relative as POI that yielded an LR > 100 ($n = 20$) with one replicate were re-examined using three replicates if available ($n = 19$). In all but one case the LRs using three replicates were lower than those with one replicate (Fig. 5). This trend is anticipated and has previously been reported using other probabilistic genotyping software [36]. More true negatives were obtained with three replicates, although four of the 19 Type II errors remained when using a brother or father/son of a true contributor as POI under Hp in a 4p mixture. Their corresponding LRs were ban + 9.2, + 5.7, + 3.5 and + 1.7. The LR of ban + 9.2 with three replicates was larger than with one replicate (ban + 7.1). This concerned a 4p mixture with four unseen alleles for the POI who was a simulated brother of one of the true contributors. Peak height information supported Hp and this simulated brother could be well explained as one of the minor contributors of the mixed profile. The use of replicates can thus reduce the number of Type II errors, although few Type II errors were still observed with a simulated relative of the true contributor (seen as triangles above the zero line in Fig. 5). This reduction in LR is expected to be observed also when using two instead of one replicate, though the effect is likely to be smaller than when using three replicates.

3.7. The effects of assigning an incorrect number of contributors to a DNA mixture

With continuous models, it is expected that increasing the number of assigned contributors beyond that necessary to reasonably explain the observed mixture will have little effect on the LRs for Hp-true tests (at least for major contributors). However, for non-contributors in Hd-true tests, the effects are more difficult to predict and could produce

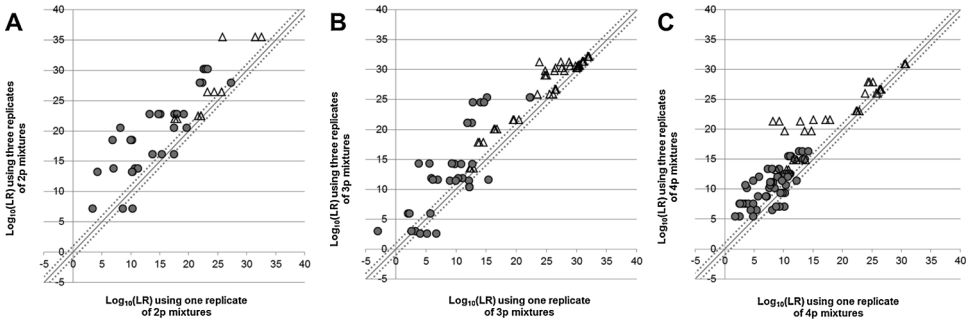

Fig. 4. Log10 LRs obtained in Hp-true tests using either a single replicate (X-axes) or three replicates (Y-axes) for 2p (panel A), 3p (panel B) and 4p mixtures (panel C). Filled circles are minor contributors (30 or 60 pg) and open triangles are major contributors (150, 300 or 600 pg). The diagonal line represents X = Y and the dotted lines + /- ban 1.

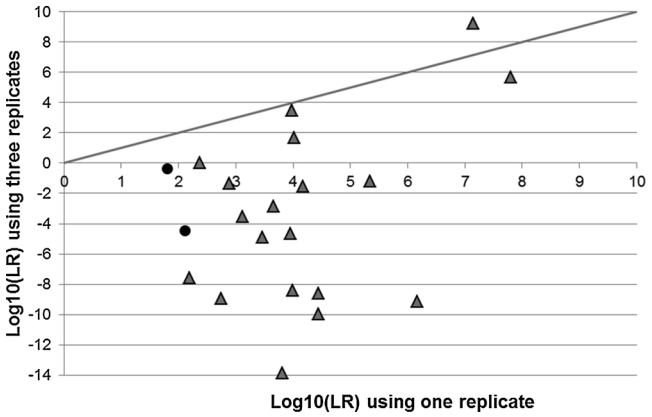

Fig. 5. Log10 LRs obtained in Hd-true tests using either a single replicate (X-axis) or three replicates of the same DNA extract (Y-axis). Black circles indicate unrelated individuals that yielded an LR > 1 with a single replicate whereas grey triangles are from simulated relatives of a true contributor that produced an LR > 100 with a single replicate. The diagonal line represents X = Y.

Type II errors. We examined this issue using our data. In total, 101 Hd-true tests were performed with an over-assigned NOC under the hypotheses. Forty-four 2p mixtures were analysed as having a total of three contributors and 57 3p mixtures were analysed as having a total of four contributors (Supplementary Table 2A). Results were compared with LRs obtained using the actual NOC under the hypotheses. The LR increased when an extra contributor was added for all of the 2p mixtures and for 91% of the 3p mixtures (Fig. 6). This increase resulted in LRs that tended towards one (neutral evidence). With 2p and 3p mixtures 84% and 72%, respectively, yielded an LR of ban-1 and higher. This trend was expected as a non-contributor can be more easily accommodated if it is in a low mixing proportion. Using the true NOC for these analyses, 2p mixtures did not yield Type II errors. When invoking an additional contributor, the largest LR observed was 120. For 3p mixtures the largest LR observed with the true NOC was 65 and this number did not alter after analyses using an extra contributor under the hypotheses. LRs larger than one were seen for each of the mixture types B, C and E (Fig. 6). The B and C mixtures resulted most often in LRs around one when overestimating the NOC, whilst the results with type E mixtures showed more variability. The type E mixtures comprise a clear majority contribution together with low-template minor contributors. As all of the donors contributed a different amount of DNA (600:30(:60)pg) it is difficult to find a non-donor that fits the observed peak heights when the true NOC are compared. When an extra contributor is added, however, it is possible to obtain Type II errors with these types of mixtures (Fig. 6).

Besides the effect of over-assigning the NOC in Hd-true tests, the effect of under-assigning was examined for Hp-true tests. These were performed for 23 DNA-profiles that could reasonably be under-assigned based on the maximum number of alleles per locus. For 3/23 mixtures the specified model could not explain the data with an under-assigned NOC. The remaining 20 analyses all produced an LR larger than ban + 9 with the major contributor as POI. Eighteen LRs could be computed with a minor contributor as POI of which the largest LR was ban + 8.9. Type I errors were obtained for 4/18 analyses; *i.e.* two 5p and two 4p mixtures (Supplementary Table 4). The under-assigned 4p mixtures were also analysed using the true NOC and the results revealed hardly effect on the LR for major contributors. Nevertheless, for 3/6 of these analyses a lower likelihood was obtained under the hypotheses when using the under-assigned NOC, indicating that the data could be better explained with the true NOC. For 2/5 4p mixtures that were analysed as 3p mixtures, a strong exclusionary LR was obtained; *i,e.* ban-15.5 and -13.2. Interestingly, these analyses did not yield a Type I error when regarding the true NOC (ban + 3.8 and + 5.4, respectively) (Supplementary Fig. 2).

In summary, over-assigning the NOC in Hd-true tests appeared to produce more Type II errors, though the LRs obtained in this study were relatively low (LR < 150). Furthermore, under-assigning the NOC in Hp-true tests with a minor low level contributor as POI can produce Type I errors.

3.8. Effect of sensitivity analysis on the LR

In EuroForMix the likelihoods are maximised under both hypotheses using the parameters that best explain the data. In addition, EuroForMix contains an option to perform sensitivity analysis that considers the LR as a function of the parameters involved [29]. As a so-called 'conservative' value (*i.e.* favouring Hd) one can use the lower 5 percentiles of these analyses. We performed sensitivity analyses, took the 5% conservative LR and compared these with the MLE LR. In Hp-true tests, the conservative approach resulted in LRs that were 0.6%–10.7% lower than the MLE value. The average decrease in LR was 3% of the MLE LR. This trend appeared to be independent of the NOC and of whether a true (n = 126) or a non-contributor (n = 9) was used as POI (Supplementary Fig. 3). The reduction in the value of the LR is relatively small and therefore would only have a marginal effect on the number of Type II errors.

3.9. Comparison to LRmix Studio

If the peak heights in a DNA profile are informative for the contribution per donor, it is expected that Hp-true tests result in larger LRs when using a continuous model than when using a semi-continuous model. Conversely, the use of peak height information in the LR calculation may lead to lower LRs in Hd-true tests. We compared the LRs from Hp- and Hd-true tests when obtained using the continuous model EuroForMix and the semi-continuous model LRmix Studio.

As expected, most benefit from using peak height information in the LR calculations using PPF6C profiles was observed for Hp-true tests in which a minor contributor having allelic drop-out is the POI under Hp as these resulted in larger LRs compared to models without peak height.

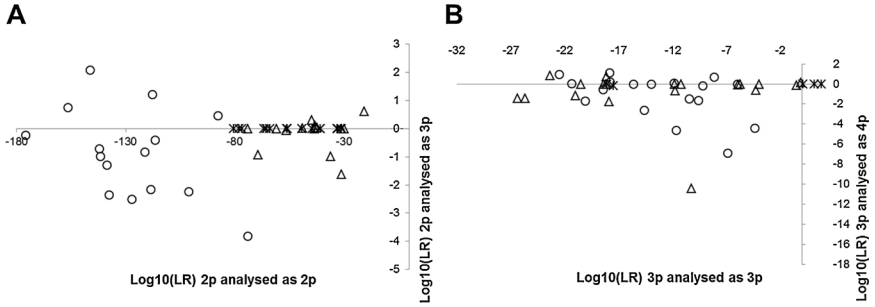

Fig. 6. The effect of over-assigning the NOC by one in Hd-true tests for 2p (A) and 3p mixtures (B). Log10 LRs are shown for type B mixtures (triangles), type C mixtures (crosses) and type E mixtures (circles).

Using peak height information in the LR calculations had a variable effect on Hd-true tests. With the data used in this study a lower number of Type II errors was obtained with LRmix Studio than with EuroForMix. When applying an LR threshold (also presented as 'uninformative LR zone') for reporting of, for example, > 100 the numbers of Type II errors are quite similar though (18 and 16 with EuroForMix and LRmix Studio, respectively). Further details and results of these comparisons are presented in Supplementary Material 1.

4. Concluding remarks

In this study we examined the ranges of LRs that can be expected from EuroForMix in Hp-true and Hd-true tests carried out on 2-4p mixtures produced with the PPF6C STR typing kit. The incidence of Type I and II errors was investigated.

The 2p mixtures used in this study did not produce Type I or II errors. The minor contributors to these mixtures had as low as 30 pg of DNA present. However, if lower levels of DNA than this are present, Type I errors are to be expected. For 3p and 4p mixtures, in Hp-true tests, Type I errors were observed for minor contributors, that had at least five or 11 allelic drop-out events, respectively. The number of Type I errors decreased if multiple replicates were utilised in the analysis.

For Hd-true tests, non-contributors were deliberately selected to have a large number of shared alleles with the mixtures that were created. Nevertheless, relatively few Type II errors were observed and the largest LR obtained was 120. To reduce the chance of obtaining type II errors in a casework setting, one could consider creating a lower reporting threshold that is larger than this value. The use of multiple replicates also decreased the number of Type II errors.

The system was pushed further to its limits by using simulated relatives of true contributors that showed a large overlap with the mixture profiles. These analyses showed that a brother or father/son of a true contributor can yield a large LR, and that an LR threshold or sensitivity analysis cannot eliminate all of the Type II errors.

Acknowledgements

We are thankful to Anouk Backx for mixture creation, to Kristiaan van der Gaag for technical assistance, to Klaas Slooten for providing simulated reference profiles, to Øyvind Bleka and Peter Gill for useful discussions and to Peter Gill for critically reading the manuscript. We are grateful to Peter de Knijff from Leiden University Medical Center, The Netherlands, for providing the DNA extracts used in this study.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.fsigen.2019.06.005.

References

[1] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, N. Morling, M. Prinz, P.M. Schneider, B.S. Weir, DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2–3) (2006) 90–101.

[2] P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, Forensic Sci. Int. Genet. 7 (2) (2013) 251–263 http://lrmixstudio.org/.

[3] K. Inman, N. Rudin, K. Cheng, C. Robinson, A. Kirschner, L. Inman-Semerau, K.E. Lohmueller, Lab Retriever: a software tool for calculating likelihood ratios incorporating a probability of drop-out for forensic DNA profiles, BMC Bioinformatics 16 (2015) 298 http://www.scieg.org.

[4] A.A. Mitchell, J. Tamariz, K. O'Connell, N. Ducasse, Z. Budimlija, M. Prinz, T. Caragine, Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in, Forensic Sci. Int. Genet. 6 (6) (2012) 749–761.

[5] R. Puch-Solis, T. Clayton, Evidential evaluation of DNA profiles using a discrete statistical model implemented in the DNA LiRa software, Forensic Sci. Int. Genet. 11 (2014) 220–228 https://d1jj3zdoktl3jd.cloudfront.net/european-west/media/1418957/lgc_lira_fact_sheet_en_0815_90.pdf.

[6] https://nichevision.com/armedxpert/.

[7] 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 (2016) 35–44 http://euroformix.com.

[8] C.D. Steele, M. Greenhalgh, D.J. Balding, Verifying likelihoods for low template DNA profiles using multiple replicates, Forensic Sci. Int. Genet. 13 (2014) 82–89 https://sites.google.com/site/baldingstatisticalgenetics/software/likeltd-r-forensic-dna-r-code.

[9] M.W. Perlin, A. Sinelnikov, An information gap in DNA evidence interpretation, PLoS One 4 (12) (2009) e8327 https://www.cybgen.com.

[10] https://www.qualitype.de/en/solutions/products/evaluation-software/genoproof-mixture/.

[11] D.A. Taylor, J.A. Bright, J.S. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int. Genet. 7 (2013) 516–528 https://www.strmix.com.

[12] S. Manabe, C. Morimoto, Y. Hamano, S. Fujimoto, K. Tamaki, Development and validation of open-source software for DNA mixture interpretation based on a quantitative continuous model, PLoS One 12 (11) (2017) e0188183 https://github.com/manabe0322/Kongoh/releases.

[13] https://www.softgenetics.com/MaSTR.php.

[14] http://dna-view.com/downloads/Mixture%20Solution%20poster.pdf.

[15] R.G. Cowell, T. Graversen, S. Lauritzen, J. Mortera, Analysis of forensic DNA mixtures with artefacts, J. R. Stat. Soc. Ser. C 64 (1) (2015) 1–48 http://dnamixtures.r-forge.r-project.org.

[16] H. Swaminathan, A. Garg, C.M. Grgicak, M. Medard, D.S. Lun, CEESIt: a computational tool for the interpretation of STR mixtures, Forensic Sci. Int. Genet. 22 (2016) 149–160.

[17] UK Forensic Science Regulator, 31 July 2018, Software validation for DNA mixture interpretation, FSR-G-223, Issue 1, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/730994/G223_Mixture_software_validation_Issue1.pdf.

[18] J.A. Bright, I.W. Evett, D. Taylor, J.M. Curran, J. Buckleton, A series of recommended tests when validating probabilistic DNA profile interpretation software, Forensic Sci. Int. Genet. 14 (2015) 125–131.

[19] SWGDAM, Guidelines for the Validation of Probabilistic Genotyping Systems, (2015).

[20] H. Haned, P. Gill, K. Lohmueller, K. Inman, N. Rudin, Validation of probabilistic genotyping software for use in forensic DNA casework: definitions and illustrations, Sci. Justice 56 (2016) 104–108.

[21] M.D. Coble, J. Buckleton, J.M. Butler, T. Egeland, R. Fimmers, P. Gill, L. Gusmão, B. Guttman, M. Krawczak, N. Morling, W. Parson, N. Pinto, P.M. Schneider, S.T. Sherry, S. Willuweit, M. Prinz, DNA Commission of the International Society for Forensic Genetics: recommendations on the validation of software programs performing biostatistical calculations for forensic genetics applications, Forensic Sci. Int. Genet. 25 (2016) 191–197.

[22] ENFSI (2017) Best Practice Manual for the Internal Validation of Probabilistic Software to Undertake DNA Mixture Interpretation, (2019) QCC-BPM-003 issue 001.

[23] T.R. Moretti, R.S. Just, S.C. Kehl, L.E. Willis, J.S. Buckleton, J.A. Bright, D. Taylor, A.J. Onorato, Internal validation of STRmix™ for the interpretation of single source and mixed DNA profiles, Forensic Sci. Int. Genet. 29 (2017) 126–144.

[24] J.A. Bright, R. Richards, M. Kruijver, H. Kelly, C. McGovern, A. Magee, A. McWhorter, A. Ciecko, B. Peck, C. Baumgartner, C. Buettner, S. McWilliams, C. McKenna, C. Gallacher, B. Mallinder, D. Wright, D. Johnson, D. Catella, E. Lien, C. O'Connor, G. Duncan, J. Bundy, J. Echard, J. Lowe, J. Stewart, K. Corrado, S. Gentile, M. Kaplan, M. Hassler, N. McDonald, P. Hulme, R.H. Oefelein, S. Montpetit, M. Strong, S. No¨el, S. Malsom, S. Myers, S. Welti, T. Moretti, T. McMahon, T. Grill, T. Kalafut, M.M. Greer-Ritzheimer, V. Beamer, D.A. Taylor, J.S. Buckleton, Internal validation of STRmix™—a multi laboratory response to PCAST, Forensic Sci. Int. Genet. 34 (2018) 11–24.

[25] A.A. Westen, T. Kraaijenbrink, E.A. Robles de Medina, J. Harteveld, P. Willemse, S.B. Zuniga, K.J. van der Gaag, N.E.C. Weiler, J. Warnaar, M. Kayser, T. Sijen, P. de Knijff, Comparing six commercial autosomal STR kits in a large Dutch population sample, Forensic Sci. Int. Genet. 10 (2014) 55–63.

[26] J.A. Nicklas, E. Buel, Development of an Alu-based, real-time PCR method for quantitation of human DNA in forensic samples, J. Forensic Sci. 48 (2003) 936–944.

[27] J.A. Nicklas, E. Buel, Simultaneous determination of total human and male DNA using a duplex real-time PCR assay, J. Forensic Sci. 51 (2006) 1005–1015.

[28] http://www.euroformix.com/?q=changes, Accessed May 2019.

[29] Ø. Bleka, C.C.G. 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 (2016) 85–96.

[30] http://euroformix.com/sites/default/files/EuroForMixTheory_ISFG17.pdf Accessed May 2019.

[31] C.C.G. Benschop, C.P. van der Beek, H.C. Meiland, A.G.M. van Gorp, A.A. Westen, T. Sijen, Low template STR typing: effect of replicate number and consensus method on genotyping reliability and DNA database search results, Forensic Sci. Int. Genet. 5 (2011) 316–328.

[32] C.D. Steele, M. Greenhalgh, D.J. Balding, Verifying likelihoods for low template DNA profiles using multiple replicates, Forensic Sci. Int. Genet. 13 (2014) 82–89.

[33] P. Gill, J. Whitaker, C. Flaxman, N. Brown, J. Buckleton, An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA, Forensic Sci. Int. 112 (2000) 17–40.

[34] C.C.G. Benschop, H. Haned, S.Y. Yoo, T. Sijen, Evaluation of samples comprising minute amounts of DNA, Sci. Justice 55 (2015) 316–322.

[35] K.R. Duffy, N. Gurram, K.C. Peters, G. Wellner, C.M. Grgicak, Exploring STR signal in the single- and multicopy number regimes: deductions from an in silico model of the entire DNA laboratory process, Electrophoresis 38 (2017) 855–868.

[36] S. Noel, J. Noel, D. Granger, J.F. Lefebre, D. Seguin, STRmix put to the test: 300.000 non-contributor profiles compared to four-contributor DNA mixtures and the impact of replicates, Forensic Sci. Int. Genet. 41 (2019) 24–31.