Journal of Forensic Sciences 2021; 66 (6), pp. 2138-2155 (DOI: 10.1111/1556-4029.14886)

A comparison of likelihood ratios obtained from EuroForMix and STRmix™

Kevin Cheng M.Sc.[1,2], Øyvind Bleka Ph.D.[3], Peter Gill Ph.D.[3,4], James Curran Ph.D.[2], Jo-Anne Bright Ph.D.[1], Duncan Taylor Ph.D.[5,6], John Buckleton D.Sc.[1,2]

1. Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland, 1142 New Zealand
2. Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
3. Forensic Genetics Research Group, Oslo University Hospital, Oslo, Norway
4. Department of Clinical Medicine, University of Oslo, Oslo, Norway
5. Forensic Science SA, GPO box 2790, Adelaide, South Australia 5001
6. School of Biological Sciences, Flinders University, GPO Box 2100 Adelaide SA, Australia 5001

Acknowledgements

The authors would like to thank Drs Hannah Kelly and Maarten Kruijver for their helpful contributions and critical review of the manuscript.

This work was supported in part by grant NIJ 2020-DQ-BX-0022 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of their organizations.

Conflict of Interest

Two of the authors were the original developers of EuroForMix, and three of the authors were the original developers of STRmix™.

ABSTRACT

Likelihood ratios (*LR*) differences between the probabilistic genotyping software EuroForMix and STRmix™ are examined. After considering differences in the allele probabilities, the *LRs* from both software for an unambiguous single-source profile were identical (four significant figures). *LRs* from both software for an unambiguous single-source profile with alleles previously unseen in the allele frequency database (rare alleles) were the same (three significant figures) for θ=0.01. Due to differences in the minimum allele frequencies, the *LR*s differed by three orders of magnitude when θ=0.

For both software, the *LR*s for a single-source dilution series decreased as the input amount decreased. The *LR*s from both software were within an order of magnitude for known contributors. The largest difference was where the target input amount was 0.0156 ng: The $LR_{\text{EuroForMix}}$ was 2.1×10^{25} and the LR_{STRmix} was 8.0×10^{24}.

Both software show similar *LR* behaviour with respect to mixture ratio. For two person mixtures the *LR* increases for both the major and the minor as the ratio moves away from 1:1. The *LR* for the major stabilises at about 3:1 whereas the *LR* for the minor reaches its maximum at about 3:1 and then declines.

Greater differences in *LR* were observed between EuroForMix and STRmix™ for mixtures. One-hundred and twenty-nine (129) mixtures from the PROVEDIt dataset were compared. *LR*s for 84% of the comparisons for known contributors without rare alleles were within two orders of magnitude. Five divergent results were investigated, and a manual intervention approach was applied where appropriate.

KEYWORDS

Probabilistic genotyping, forensic DNA analysis, mixtures, EuroForMix, STRmix, STRs

HIGHLIGHTS

- A comparison of likelihood ratios (*LR*) between two probabilistic genotyping software – EuroForMix and STRmix™.
- Similarities and differences between software were assessed with single-source profiles and 129 mixtures.
- Results demonstrate that even though there are differences, both software can be useful in assigning an *LR*.

Like other disciplines, the forensic interpretation of DNA mixtures is becoming increasingly automated, by the application of statistical models using computer-based methods. The interpretation of forensic DNA profiles using continuous models and computer software is collectively termed probabilistic genotyping (PG) and all modern PG software are able to assign likelihood ratios (*LR*) (1-9).

The British statistician, George Box, has been famously quoted as saying "Essentially, all models are wrong, but some are useful." (10). By saying that "all models are wrong" is to say that every model makes some fundamental assumptions about reality, no model can ever hope to cover all the intricacies of a real-world system. This is applicable in all PG software, where there are many modelling assumptions made about the interpretation of forensic DNA profiles. With a good understanding of each software, the differences arising from these assumptions can be predicted; and in some cases, software options or workarounds allow these differences to be minimised.

Although, making these assumptions, or simplifications of reality, means that the models are "wrong"; they can be very useful for better understanding what is being modelled and predicting the outcome given certain inputs. The use of models within PG software allows forensic practitioners to evaluate DNA profiles and assign *LR*s to a pair of propositions. The question is then whether the *LR* from different PG software are "equally reliable" or "equally useful".

As an example, consider the probability assigned for an allele that has never been seen before in the population sample, but is observed in the evidence in this case. We can say for certain that the "true" probability of observing this allele in a randomly selected person is not zero, but we are uncertain exactly what it is. Whenever something is unknown and uncertain it is best to model the uncertainty with a probability density function. A workable option may be to insert a reasonable point estimate. Further, in forensic science, some aspects of utility are usually confounded into the probability assignment by deliberately biassing the assignment in a direction thought to be conservative. However, in mixture evaluation the conservative direction is very uncertain. For example, it is typically conservative to increase the allele frequency for the alleles that correspond with the person

of interest (POI) in the *LR* calculation, but for any other alleles the effect may be neutral or may vary either way. The use of a point estimate biased upwards (for example $5/2N$ or $3/2N$ where N is the number of individuals in the population database) is plausibly conservative on average although we are unaware of any systematic investigation of this assumption. The use of a probability distribution and resampling may enable the choice of a conservative quantile but requires assignment of a distribution. It would be very difficult, and be a matter of subjective judgement, to choose which of these methods is appropriately conservative.

Earlier within the same text, Box states, "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful." (10). In the context of PG software, where two software may implement two different models for the same process if we can assess how well the models describe the empirical data and we can ensure the veracity of the inferential process, then we can have confidence in the result. This can be readily supplemented by varying the model within reasonable limits dictated by the data and thus creating a range of plausibly "correct" or "useful" outcomes. We are left with the uncertainty that small modelling and inferential errors accrue, or that the training data for the models are inappropriate.

In this work we compare two PG software: EuroForMix and STRmix™ (1, 11). The Maximum Likelihood based approach was used in EuroForMix. Both software attempt to give some sensibly conservative lower-bound to the *LR*. Hence the number should not be considered "the *LR*", but something more like: a number assigned from the lower tail of the plausible range. We accept that this is vastly too much of a mouthful for any actual usage and needs some considerable truncation for court. We also use the word "assigned" rather than "estimated" although both are appropriate. We were taught to use the word "assigned" by Evett who, correctly, felt that it indicated the subjective nature of certain underlying assumptions, this is because Bayesian estimation is subjective by definition, thus rendering this distinction unnecessary. We add to this complex mix of thoughts the fact that in some countries such as the United Kingdom and Australia set a limit on the reported *LR* (UK at 10^9 and Australia at 10^{11}). This means that any assignment given that it is above these numbers, however different, would be reported the same.

There are strong drivers for carrying out comparisons between different probabilistic genotype models. It is well known that different models, implemented in different software products, can produce divergent results. Studies, such as that published by Alladio et al. (12) have shown that similar models (i.e. both qualitative, or both quantitative) will produce mostly consistent results. However, there are published examples of differences (13) between software in ways that may affect the court outcome. As a consequence of this we have been asked by members of the legal community whether it would be best to run each profile through multiple systems before reporting a result. While this would represent one possible option for investigating whether the *LR* obtained in any one system is robust it is unlikely to be a viable option due to the overnight increase in workload. However, the best parts of that ethos can be taken and pursued. The most important aspect of analysing a profile using multiple models is to guard against the situation where they give divergent results. Previous work has shown that divergence between the models will mostly not occur, however the 'risk areas' can be identified and investigated from studies comparing software (12, 14-16). In doing so, the aim is to identify the aspects of modelling that fundamentally leads to the divergent results and determine whether there is any scope to improve the modelling.

This thinking is also reflected in the report given in the President's Council of Advisors on Science and technology, PCAST (17). In their report from 2016, in the discipline of biology the authors called for (amongst other things) an investigation into "*Under what circumstances – and why – does the method produce results (random inclusion probabilities) that differ substantially from those produced by other methods?*"

PCAST advocated that this comparison should be carried out by independent groups (i.e. not the developers of the software. An independent comparison of EuroForMix (version 2.1) and STRmix™ (version 2.6) was recently published out by Riman et al. (18). We believe that our concurrent study reinforces the findings from Riman et al. Additionally, the inclusion of two sets of developers as collaborators and developers within this study should alleviate the concern that the work will be biased towards a single model and provide in-depth understanding of the two software.

This suggests that a sensible goal for this work might be to identify those factors driving any difference in the assigned *LR* without any of the "amendments," for example a lower or upper bound. We will call this "the *LR*" but remind the reader that it should probably be called something like "a plausible *LR*." Once identified, the driving factors may be assessed, models altered, and the differences potentially ameliorated.

Where it was possible, we have removed the differences between these two software, including most differences in allele probability assignment and all in the population genetic model.

2. Method

2.1. Analysis and Interpretation

All *LR*s were assigned using the NIST 1036 Caucasian allele frequencies (19). In STRmix™ the allele frequencies are normalised if the sum of the allele frequencies at each locus does not equal one. EuroForMix has the user-defined option of enabling or disabling allele frequency normalisation.

Additionally, EuroForMix has the option of setting the size of the frequency database, *N*; where *N* is the number of individuals sampled. This value is used in the minimum allele probability calculation, which is set at $\frac{5}{2N}$ and remains unchanged if normalisation is disabled. If normalisation is enabled in EuroForMix, the frequencies of all the alleles are normalised (including those which are assigned with a minimum value).

In STRmix™, *N* has a similar definition and this value is also used in the posterior mean allele frequency calculation, $f_i^{'} = \frac{x_i + \frac{1}{k}}{2N+1}$; where:

- x_i is the observed allele count in the database; and,
- k is the number of observed allele classes for a particular locus.

Note that in STRmix™, *N* is technically the number of alleles sampled rather than the number of individuals sampled for the allele frequency database. The posterior mean formula is therefore, $f_i' = \frac{x_i + \frac{1}{k}}{N+1}$. To make the definition of *N* equivalent in both software, we multiple the STRmix™ *N* by 2; hence $f_i' = \frac{x_i + \frac{1}{k}}{2N+1}$.

For rare alleles or previously unobserved alleles in the allele frequency database, the posterior mean allele frequency is effectively a minimum allele frequency. Consider the previously unobserved 6 allele at CSF1PO in the NIST 1036 Caucasian allele frequency database. x_i would equal 0, k for CSF1PO is 7, and *N* equals 361 for the NIST 1036 Caucasian allele frequencies. The posterior mean allele frequency for the 6 allele at CSF1PO is 0.0002. Comparatively, using the minimum allele frequency implemented in EuroForMix, the frequency of the same allele is 0.0069 (also after normalisation).

When *N* is sufficiently large, it should mitigate the differences between the minimum allele frequency used in EuroForMix and the posterior mean allele probabilities used in STRmix™. Consider an *N* of 1,000,000; the posterior mean allele frequency for the 6 allele at CSF1PO calculated in STRmix™ is 7.1×10^{-8} and the minimum allele frequency for the same allele calculated in EuroForMix is 2.5×10^{-6}. Unless otherwise stated, in this study we have set *N* to 1,000,000 in both software.

Given that the NIST 1036 allele frequencies sum to one at each locus, normalisation in EuroForMix was disabled in order to retain the $\frac{5}{2N}$ calculation.

GlobalFiler® profiles were selected from the PROVEDIt dataset and analysed by an experienced analyst without reference to the ground-truth known genotypes in GeneMapper ID-X with an analytical threshold of 75 rfu (20). Allele, back stutter, and forward stutter peaks were retained for the interpretation in EuroForMix (version 3.0.3). A few selected profiles were reinterpreted in

EuroForMix version 3.3.0, discussed further below. Allele, back stutter, forward stutter, and double back stutter were also retained at all loci for the interpretation of profiles in STRmix™ (version 2.7.0). Two base pair back stutter peaks at SE33 and D1S1656 were also retained for STRmix™ interpretation.

A summary of STRmix™ settings that were previously determined using a calibration dataset is given in the supplementary material (Table S1). In the interpretation of the mixtures in this study, there were six observations of exclusions of known donors to the mixture using STRmix™. Following normal casework protocol, we carefully scrutinized the results by first assessing the primary diagnostics (21). We would have also further scrutinized the secondary diagnostics should it have been required (21). Examining the per-locus *LR*s for these seven observations, we noted that these were all a result of single-locus exclusions. These can be broken down into two categories,

1. Unresolved peak due to poor one base-pair separation,
2. Dropout was not proposed and accepted under default MCMC run parameters.

The usual casework interventions were applied where applicable (see the Supplementary Materials for a detailed disclosure of the subjective interventions).

2.2. Single-source profiles

We interpreted four single-source profiles in order to better understand the similarities and differences between EuroForMix and STRmix™. These profiles included a fully-resolvable single-source profile, a fully resolvable single-source profile with an allele that had not been previously observed in the allele frequency database, a fully-resolvable single-source profile with an artificial drop-in peak added to the profile; and a partial single-source profile where two alleles at different loci have dropped out of the profile. A single-source dilution series was also interpreted in both PG software.

Single-source profiles were interpreted in both software and the following propositions were considered.

148 H_1: The DNA profile originates from the POI.

149 H_2: The DNA profile originates from one unknown, unrelated individual.

150

151 2.2.1. Unambiguous single-source profile

152 An unambiguous single-source profile, B01_RD14-0003-15d2a-0.5GF-Q0.9_02.15sec, was
153 interpreted in both software. When *N* is 361, a difference in the *LR* is expected, due to the posterior
154 mean allele frequencies. When *N* is set to 1,000,000, we expect the *LR*s to be similar; if not the same.
155 *LR*s were assigned to the comparison using N=361 and N=1,000,000. We also assigned *LR*s using
156 θ=0 and θ=0.01. We replicated the *LR*s in MS (Microsoft) Excel™.

157

158 2.2.2. Unambiguous single-source profile, rare alleles

159 When assigning an *LR*, the two software treat not previously observed alleles differently. Unless
160 otherwise specified, EuroForMix will apply the minimum allele frequency calculation as the
161 frequency of an allele not previously observed in the allele frequency database, whereas STRmix™
162 will use the posterior mean allele frequency.

163 We interpreted another unambiguous single-source profile, F05_RD14-0003-50d2a-0.5GF-
164 Q0.8_06.15sec. The same propositions above were considered with N=1,000,000; θ=0 and θ=0.01.
165 We replicated the *LR*s in MS Excel™.

166

167 2.2.3. Drop-in

168 The two software have different models for drop-in. STRmix™ uses a user defined gamma or
169 uniform distribution to model drop-in, with a cap on the allowable drop-in peak height. Any peak that

is below this drop-in cap can be considered as drop-in. EuroForMix uses the drop-in hyper-parameter (λ) and an exponential distribution to model drop-in.

As an example, the same profile in section 2.2.1 was reinterpreted with an artificial drop-in artefact (TH01, 9.3) added to the evidence file, with a peak height of 99 rfu. Within STRmix™ the drop-in rate parameter was used (uniform model, 0.0001), and the EuroForMix drop-in hyper-parameter was set to the default value of 0.01.

2.2.4. Dropout

The concept of modelling dropped out alleles in the two software is similar. They consider the probability of observing an allele with a peak height between 0 and the analytical threshold. However, because of the differences in how each software models allelic peak heights, as well as the implementation of the dropout model, differences in the results are to be expected.

As an example, we interpret a partial single-source profile from the PROVEDIt dataset, F01_RD14-0003-01d3a-0.0313GF-Q0.7_06.15sec, in both software. This sample was chosen because there are two alleles that have dropped out of the profile at two different loci; the 12 allele at CSF1PO and the 6 allele at TH01.

2.2.5. Single-source dilution series

Each sample from a dilution series with target template amounts ranging between 0.0078-0.5 ng was interpreted in both software. In each case, the same propositions were considered.

The samples from the PROVEDIt dataset are:

- F05_RD14-0003-50d2a-0.5GF-Q0.8_06.15sec.hid_SS
- G05_RD14-0003-50d2a-0.25GF-Q0.8_07.15sec.hid_SS

- H05_RD14-0003-50d3a-0.125GF-Q0.9_08.15sec.hid_SS
- A06_RD14-0003-50d4a-0.0625GF-Q0.7_01.15sec.hid_SS
- B06_RD14-0003-50d4a-0.03125GF-Q0.7_02.15sec.hid_SS
- C06_RD14-0003-50d4a-0.0156GF-Q0.7_03.15sec.hid_SS
- D06_RD14-0003-50d4a-0.0078GF-Q0.7_04.15sec.hid_SS

2.3. Mixtures

2.3.1. Two-person mixtures

Five two-person mixtures comprised of individual A and individual B were simulated *in silico* to mimic a 1:1, 2:1, 3:1, 5:1, and 10:1 mixture proportion. Mixtures were generated *in silico*, because at the time of writing, two-person mixtures meeting the experimental design were not present in the PROVEDIt dataset. The mixtures were interpreted in both PG software and *LR*s were assigned considering the following propositions:

H_1: The DNA originated from the person of interest (known major or minor) and one unknown unrelated individual

H_2: The DNA originated from two unknown unrelated individuals

and

H_1: The DNA originated from the two known contributors

H_2: The DNA originated from two unknown unrelated individuals

The purpose of this experiment was to test the observations described by Bille et al. (22), where the *LR* for a contributor to a 1:1 mixture decreases compared to when they are a major contributor to another mixture. This is because the information content associated with height is less useful at a ratio of 1:1, as the two donors' allele heights are similar, resulting in ambiguity in the interpretation.

When the mixture proportions begin to deviate from 1:1, the major contributor's alleles are more readily distinguishable with more template amount resulting in an increased *LR*. For the minor contributor, the *LR* is expected to initially rise compared with the 1:1 mixture and then reduce as the amount of DNA template the minor is contributing decreases.

2.3.2. Sensitivity and specificity

Sensitivity is the ability of the software to reliably resolve the DNA profile of true contributors within a mixed DNA profile. It is typically tested over a range of starting DNA templates and mixture proportions. Specificity is the ability of the software to reliably exclude non-contributors within a mixed DNA profile.

To demonstrate sensitivity and specificity for EuroForMix and STRmix™, a range of PROVEDIt mixtures was interpreted following Taylor et al. [1], with the exception of using average peak height (*APH*) in place of the experimentally designed DNA template. This was done because *APH* can be more readily estimated from the PROVEDIt mixture electropherograms than the amount of DNA template input to the PCR per contributor.

One-hundred and twenty-nine (129) GlobalFiler® profiles, comprising 74 two-person, 30 three-person, and 25 four-person mixtures, were selected from the PROVEDIt dataset. The profiles included varying mixture proportions and template amounts. A full summary of the profiles used in this sensitivity and specificity study is available in the Supplementary Materials.

Each profile was interpreted using each software, and the results were compared to a database containing 250 individuals. This included the 50 PROVEDIt known reference profiles and 200 non-contributors that were simulated *in silico* using the NIST 1036 Caucasian allele frequency database.

Using the NIST 1036 Caucasian allele frequencies, θ=0, and N=1,000,000, the point estimate sub-source *LR* was assigned where the propositions considered were:

H_1: The DNA originated from the database individual and *NoC*-1 unknown unrelated individuals

H_2: The DNA originated from *NoC* unknown unrelated individuals

where *NoC* is the experimentally designed number of contributors to the profile.

The *APH,* was calculated using unmasked, unshared, alleles not in a stutter position for each contributor in the profile. Where the contributor's alleles were all masked, or had dropped out of the mixture, an *APH* of half the analytical threshold was used to represent the *APH*. For the non-contributors, the lowest *APH* of the known contributors was used.

We also considered the effect of allele sharing between mixture donors on the *LR*. We plot the $\log_{10}LR$ versus the fraction of allele sharing. In this study we define allele sharing for the known contributors to the mixture as the fraction of alleles shared between at least two donors. For mixtures with more than two contributors, we consider the maximum number of alleles shared between the donor of interest and all other donors. For example, consider a three-person mixture, comprised of donors A, B, and C. Each contributor's reference profile contains 20 alleles (10 loci). Donor A shares 3 alleles with donor B, and shares 2 alleles identical by state (IBS) with donor C. The maximum fraction of alleles shared for donor A is 3/20. If donor B shares 5 alleles IBS with donor C, then the max fraction of alleles shared for donor B is 5/20.

For non-donors to the mixture, we consider the fraction of alleles shared between the non-donor and the observed DNA profile, not the individual donor references. Consider the non-donor's reference profile of 20 alleles. If the observed profile has 45 peaks and 14 of the 20 non-donor's alleles are labelled in the observed profile, the fraction of alleles shared between the non-donor and the observed DNA profile is 14/45.

3. Results

3.1. Single-source profiles

We present a summary of the experiments run on single-source profiles. More details of the results, including the MS Excel™ results, are available in the Supplementary Materials Tables S2 through S5.

3.1.1. Unambiguous single-source profile

As shown in Table 1 (the per locus results appear in Tables S2 and S3) the *LR* for an unambiguous single-source profile calculated in EuroForMix and STRmix™ when using the same value of θ and when N was set to 1,000,000 agreed to four significant figures As expected, the *LR*s are slightly different when N is set to 361, because the minimum allele frequency that is used for rare alleles in EuroForMix is different to posterior mean allele frequency that is used in STRmix™.

The per locus *LR*s and the total *LR* calculated in STRmix™ can be replicated in MS Excel™. The *LR*s calculated in EuroForMix cannot be replicated beyond four significant figures in MS Excel™, since some of the per-marker *LR*s differs (see Supplementary Materials) – however the *LR*s are in the same order of magnitude. The reason for this divergence is because EuroForMix version 3 has increased the 'convergence tolerance' for the optimizer (by default). This allows for faster optimization or convergence with minimal impact on the parameter estimates and the *LR*.

Table 1: *LR*s when $\theta \in \{0, 0.01\}$ for an unambiguous single-source profile, B01_RD14-0003-15d2a-0.5GF-Q0.9_02.15sec, in EuroForMix and STRmix™. The *LR*s were also replicated using MS Excel™. Values rounded to 6 significant figures. *N* is the number of individuals sampled for the allele frequency database.

			N=1,000,000		*N*=361	
Theta	LR_{Excel}	$LR_{\text{EuroForMix}}$	LR_{Excel}	LR_{STRmix}	LR_{Excel}	LR_{STRmix}
0	5.30834×10^{33}	5.30840×10^{33}	5.30865×10^{33}	5.30865×10^{33}	4.88379×10^{33}	4.88379×10^{33}
0.01	7.97537×10^{30}	7.97547×10^{30}	7.97564×10^{30}	7.97564×10^{30}	7.90595×10^{30}	7.90595×10^{30}

3.1.2. Unambiguous single-source profile, rare alleles

The *LR*s calculated in EuroForMix and STRmix™ for an unambiguous single-source profile containing two rare alleles are presented in **Error! Reference source not found.** (the per locus results appear in Tables S4 and S5). Similar to the results in 3.1.1, the *LR*s calculated in the two PG software are the same to three significant figures when θ=0.01 and *N* is set to 1,000,000. However, the *LR*s are three orders of magnitude different when θ=0. The per locus *LR*s and the total *LR* calculated in STRmix™ can be replicated in MS Excel™, whereas the *LR*s calculated in EuroForMix cannot be replicated in MS Excel™ to the same level of precision, but are of the same order of magnitude.

Table 2: *LR*s when $\theta \in \{0, 0.01\}$ for an unambiguous single-source profile with two rare alleles, F05_RD14-0003-50d2a-0.5GF-Q0.8_06.15sec, in EuroForMix and STRmix™. The *LR*s were replicated using MS Excel™. Values rounded to 6 significant figures. Where *N*, the number of individuals sampled in the database, has been increased to 1,000,000.

	Minimum Allele Frequency		Posterior Mean Allele Frequency	
Theta	LR_{Excel}	$LR_{EuroForMix}$	LR_{Excel}	LR_{STRmix}
0	3.38796×10^{44}	3.38796×10^{44}	5.42105×10^{47}	5.42105×10^{47}
0.01	4.98121×10^{34}	4.98121×10^{34}	4.98372×10^{34}	4.98372×10^{34}

3.1.3. Drop-in

The profile used in section 2.2.1 was reinterpreted with an artificial drop-in artefact (TH01, 9.3) added to the evidence file, with a peak height of 99 rfu. The *LR* calculated in the two PG software is presented in Table 3. Because of the differences in the drop-in models, a difference in the *LR* is expected. However, the two *LR*s are the same to two significant figures.

Table 3: Summary of drop-in settings and the *LR* assigned by both software to sample B01_RD14-0003-15d2a-0.5GF-Q0.9_02.15sec when an artificial drop-in peak was added to the evidence file.

	EuroForMix	STRmix™
Drop-in Probability	0.0001	0.0001
Drop-in Cap	N/A	150 rfu
Drop-in Hyperparam (lambda)	0.01	N/A
LR	7.99853×10^{30}	7.97564×10^{30}

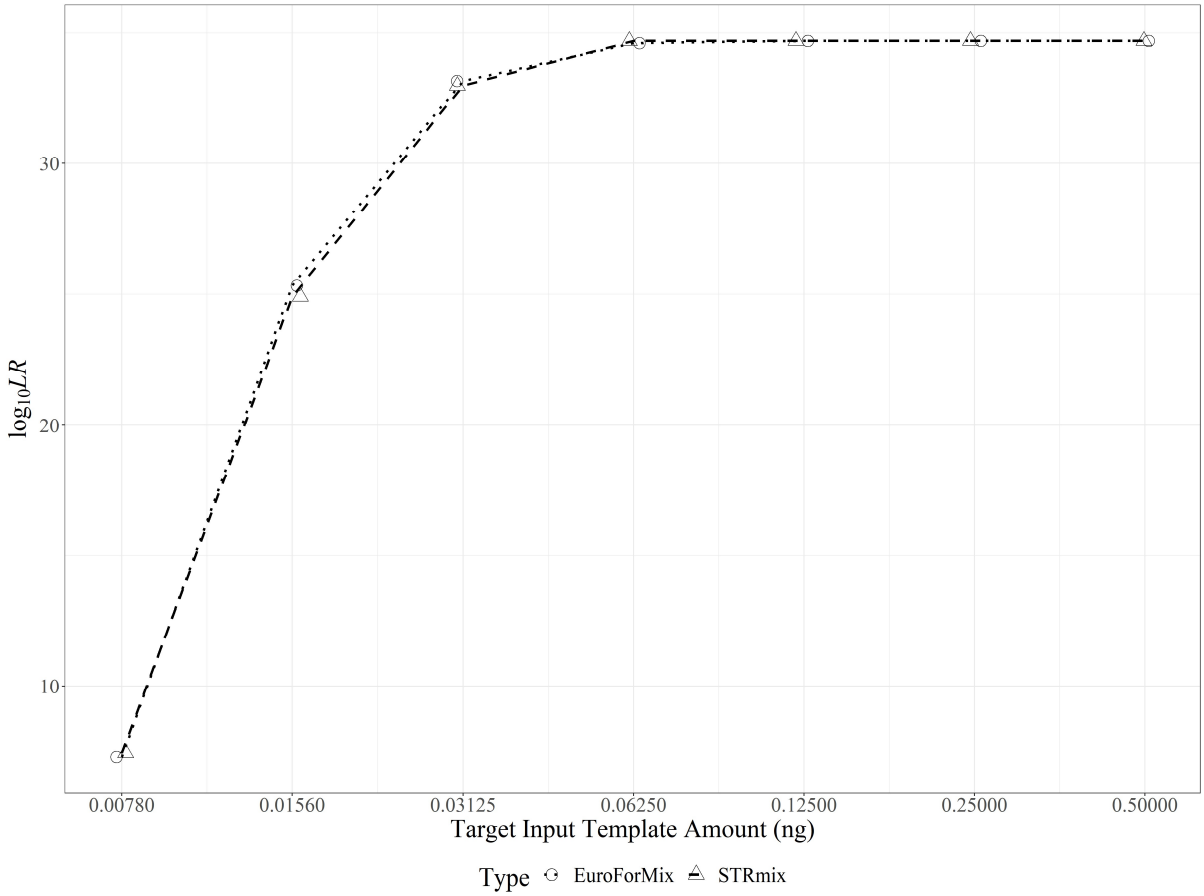
3.1.4. Dropout

The $LR_{EuroForMix}$ and LR_{STRmix} calculated in the interpretation of a partial single-source profile is 1.97×10^{25} and 1.69×10^{25}, respectively. A summary of the per-locus *LR*s and the input profile is provided in the Supplementary Material (Table S6). Differences in the results are to be expected, as there are differences in how each software models allelic peaks, as well as how each software treats potential dropout. During the developmental validation of STRmix™ the application of the models within STRmix™ has been verified separately. Dropout in EuroForMix has been compared against

empirical data (single-source dilution series) as part of a validation study (supplementary of Bleka et al. (23)).

3.1.5. Single-source Dilution Series

Figure 1 shows the *LR* assigned by both PG software to the known contributor for a single source dilution series (0.0078 – 0.5 ng). As expected, the *LR* calculated for a single-source profile decreases towards 1 as the target input amount decreases. The *LR*s between the two software are also similar, all within one order of magnitude. The largest difference was where the target input amount was 0.0156 ng where the EuroForMix *LR* was 2.1×10^{25} and the STRmix™ *LR* was 8.0×10^{24}.



322

323 Figure 1: $Log_{10}(LR)$ vs target template amount assigned by both software, EuroForMix (EFM)
324 (dashed line / circles) and STRmix™ (dotted line / triangles), to the known contributor for a dilution
325 series (0.0078 – 0.5 ng).

326

327 *3.2. Mixtures*

328 3.2.1. Two-person mixtures

329 As shown in Figure S1, in both software the *LR* for the 1:1 mixture decreases in comparison with the
330 *LR* for the major contributor when the mixtures deviate from a 1:1 ratio. This is because the
331 information content associated with height is lower in these profiles, as the two donor's allele heights
332 are similar; which results in ambiguity in the genotype (22). When the mixture proportions begin to
333 deviate from 1:1, the *LR* for the major contributor increases with the increasing template for this
334 contributor. Initially the *LR* also increases for the minor contributor, the *LR* also increases despite the

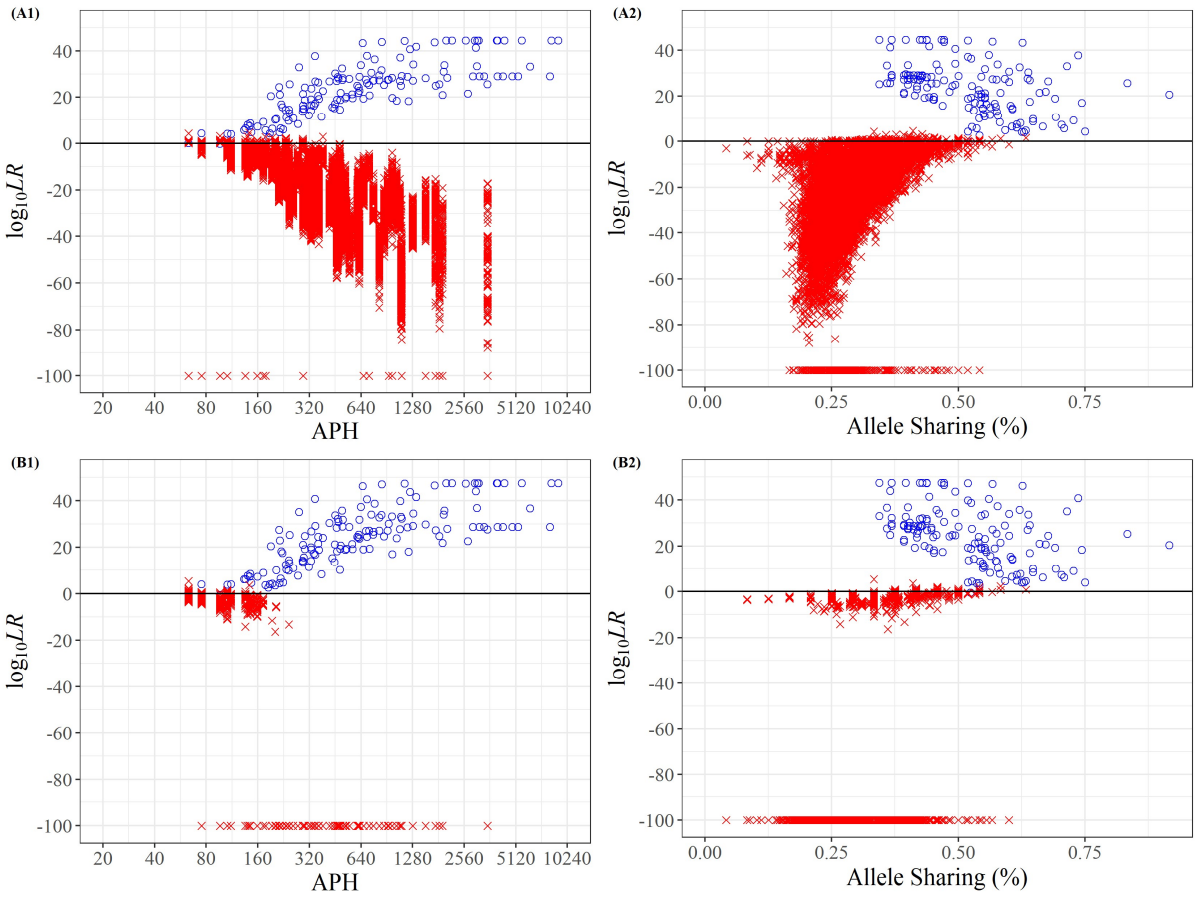
335 template for this contributor decreasing. The *LR* assigned to the minor contributor begins to decrease
336 after 3:1, as the amount of DNA template for this contributor decreases such that dropout of the minor
337 contributor's alleles is now observed. The rise and then fall of the *LR* for the minor contributor is
338 explained by competing effects. First, the minor's alleles having distinguishable peak height across
339 the profiles (with enough of an effect on the major that peak imbalances in masked peaks are
340 identifiable) which increases resolution for unbalanced profiles. Second, the effect of dropout,
341 increased peak height variability, and masking (which at high ratios the effect on major peak heights
342 can be subsumed into the expected peak height variability of the major) reduce resolution for
343 unbalanced profiles. When we consider both individuals under H_1 (and two unknown unrelated
344 individuals under H_2) we see a large increase in the *LR* because we are considering the combination of
345 two known donors to the mixture. More importantly, a similar trend in the *LR*s as highlighted above
346 would be observed when the mixture proportion changes. The rise and fall of the *LR* is also explained
347 by the same competing effects. This is the expected result in both software.

348

349 3.2.2. Sensitivity and specificity

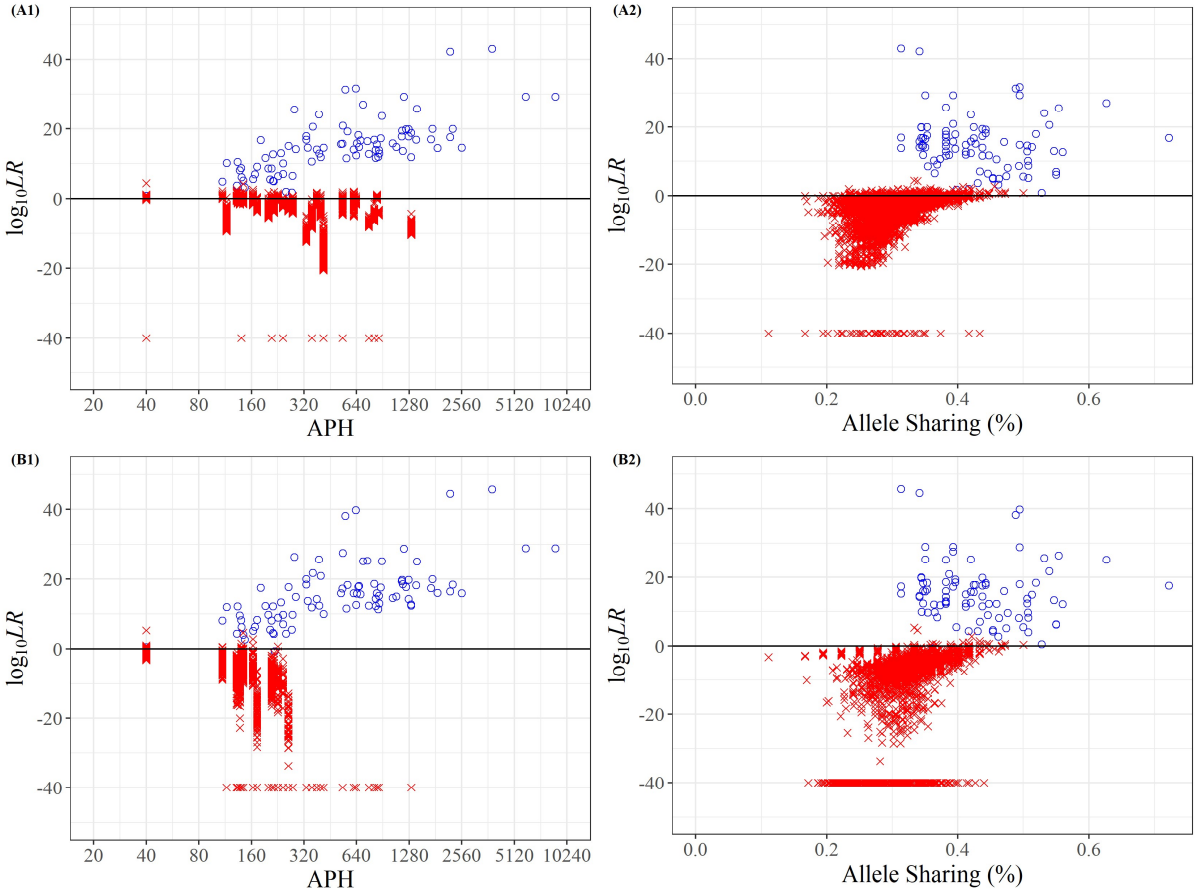
350 Plots of the log(*LR*) versus the average peak height (*APH*) and the log(*LR*) versus the fraction of allele
351 sharing for the two-person mixtures (74 profiles), three-person mixtures (30 profiles), and four-person
352 mixtures (25 profiles) are presented in Figure 2 through Figure 4. Known donors are shown as
353 circles, and non-donors as crosses.

354



355

356 Figure 2: Scatter plot of the $\log_{10}LR$ versus the average peak height from the interpretation of two-
357 person mixtures using EuroForMix (A1) and STRmix™ (B1). Scatter plot of the $\log_{10}LR$ versus the
358 % allele sharing from the interpretation of two-person mixtures using EuroForMix (A2) and
359 STRmix™ (B2). Panels A are the results for EuroForMix. Panels B are results for STRmix™. *LR*s
360 assigned to known contributors (circles) and known non-contributors (crosses) are shown. *LR*s of 0
361 are presented as -100 for known non-contributors.



362

363 Figure 3: Scatter plot of the $\log_{10}LR$ versus the average peak height from the interpretation of three-
364 person mixtures using EuroForMix (A1) and STRmix™ (B1). Scatter plot of the $\log_{10}LR$ versus the
365 % allele sharing from the interpretation of three-person mixtures using EuroForMix (A2) and
366 STRmix™ (B2). Panels A are the results for EuroForMix. Panels B are results for STRmix™. *LR*s
367 assigned to known contributors (circles) and known non-contributors (crosses) are shown. *LR*s of 0
368 are presented as -40.

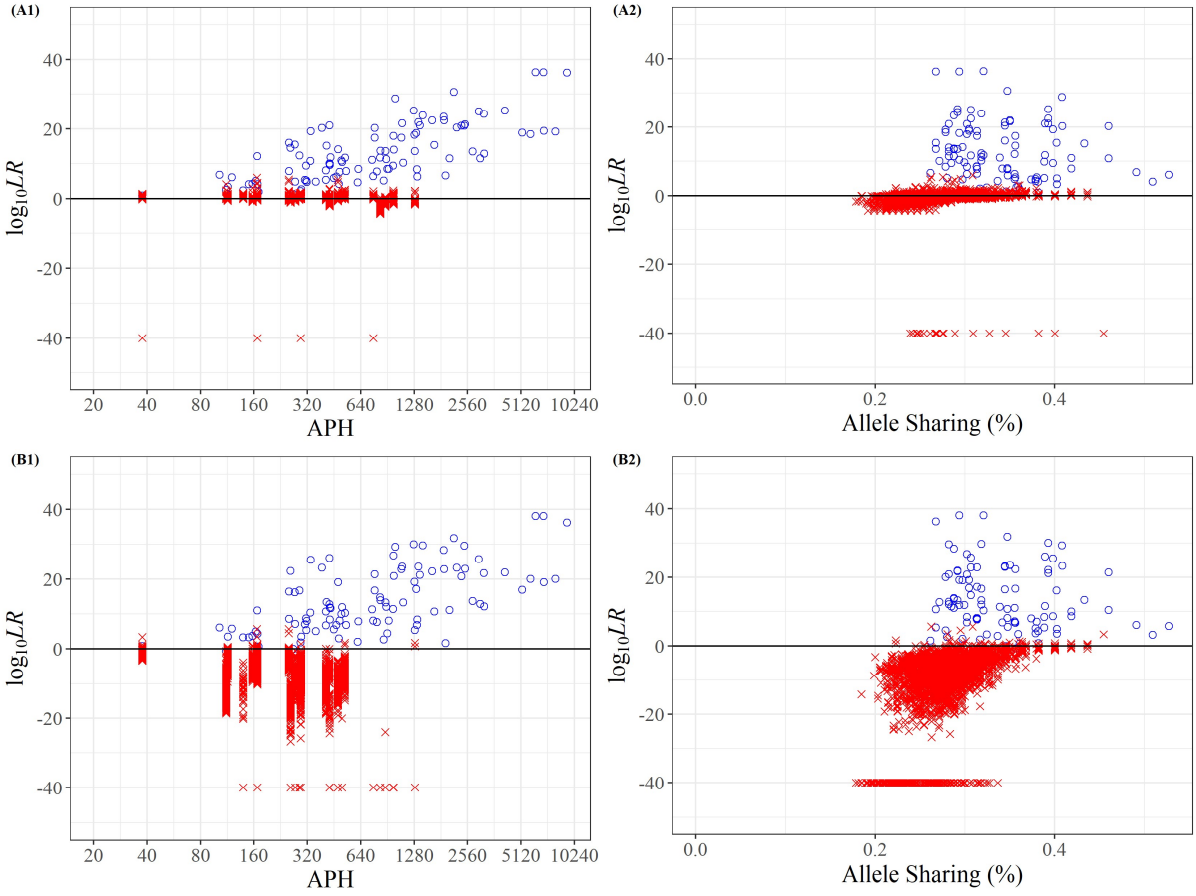


Figure 4: Scatter plot of the $\log_{10}LR$ versus the average peak height from the interpretation of four-person mixtures using EuroForMix (A1) and STRmix™ (B1). Scatter plot of the $\log_{10}LR$ versus the % allele sharing from the interpretation of four-person mixtures using EuroForMix (A2) and STRmix™ (B2). Panels A are the results for EuroForMix. Panels B are results for STRmix™. *LR*s assigned to known contributors (circles) and known non-contributors (crosses) are shown. *LR*s of 0 are presented as -40.

We plot $\log_{10}LR$ for STRmix™ vs $\log_{10}LR$ for EuroForMix for the known donors to the two-, three-, and four-person mixtures in Figure 5. Five divergent results between the two software, marked with black arrows in Figure 5, were further investigated:

- B10_RD14-0003-39_40-1;2-M3a-0.045GF-Q0.8_02.25sec; Contributor K39
- E03_RD14-0003-33_34-1;2-M3a-0.045GF-Q0.8_05.25sec; Contributor K33
- E10_RD14-0003-39_40-1;2-M3c-0.093GF-Q1.0_05.25sec; Contributor K39

- H01_RD14-0003-31_32-1;1-M2c-0.062GF-Q2.0_08.25sec; Contributor K32
- H09_RD14-0003-30_31_32-1;4;4-M2d-0.75GF-Q0.6_08.25sec; Contributor K30

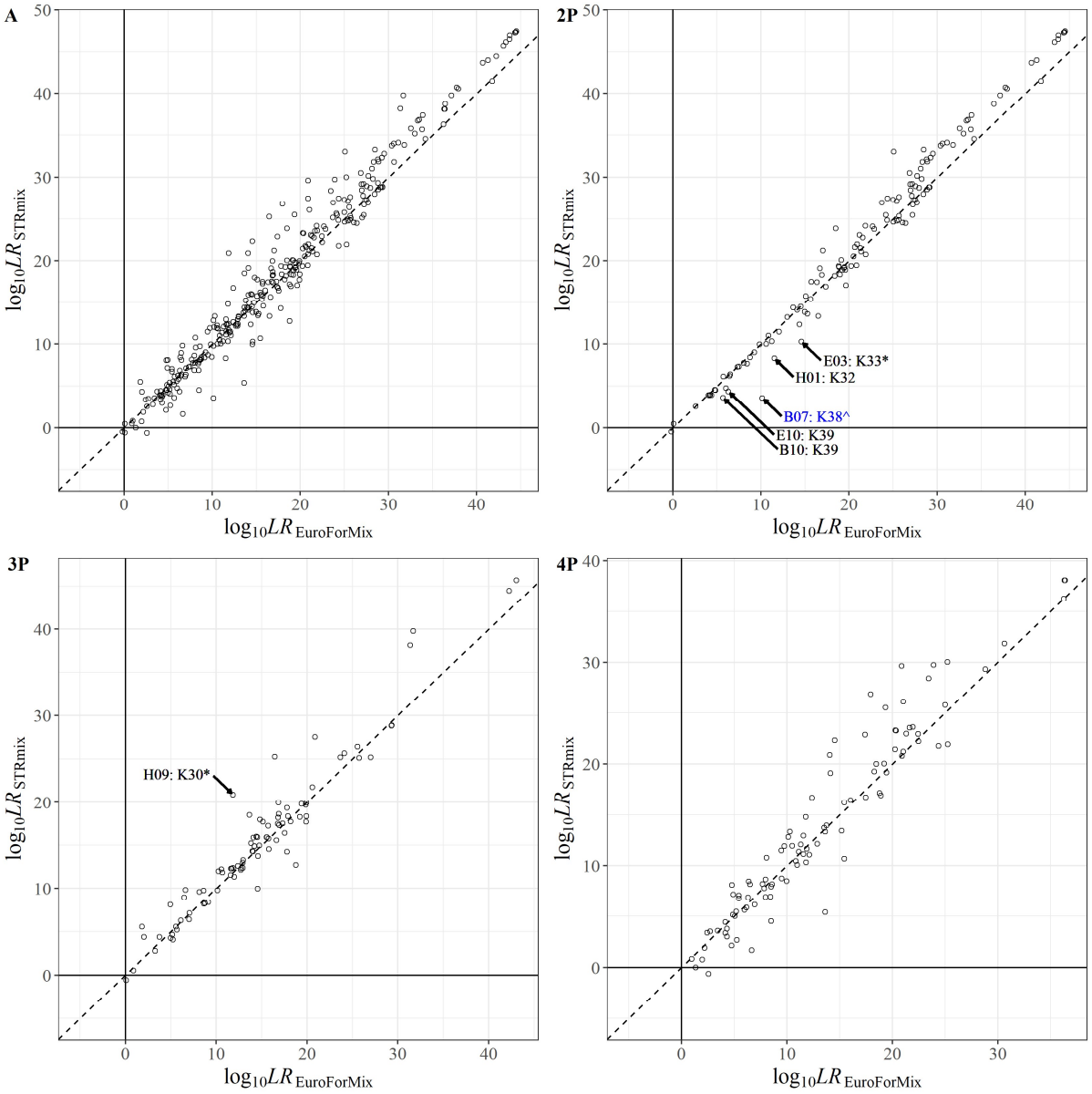
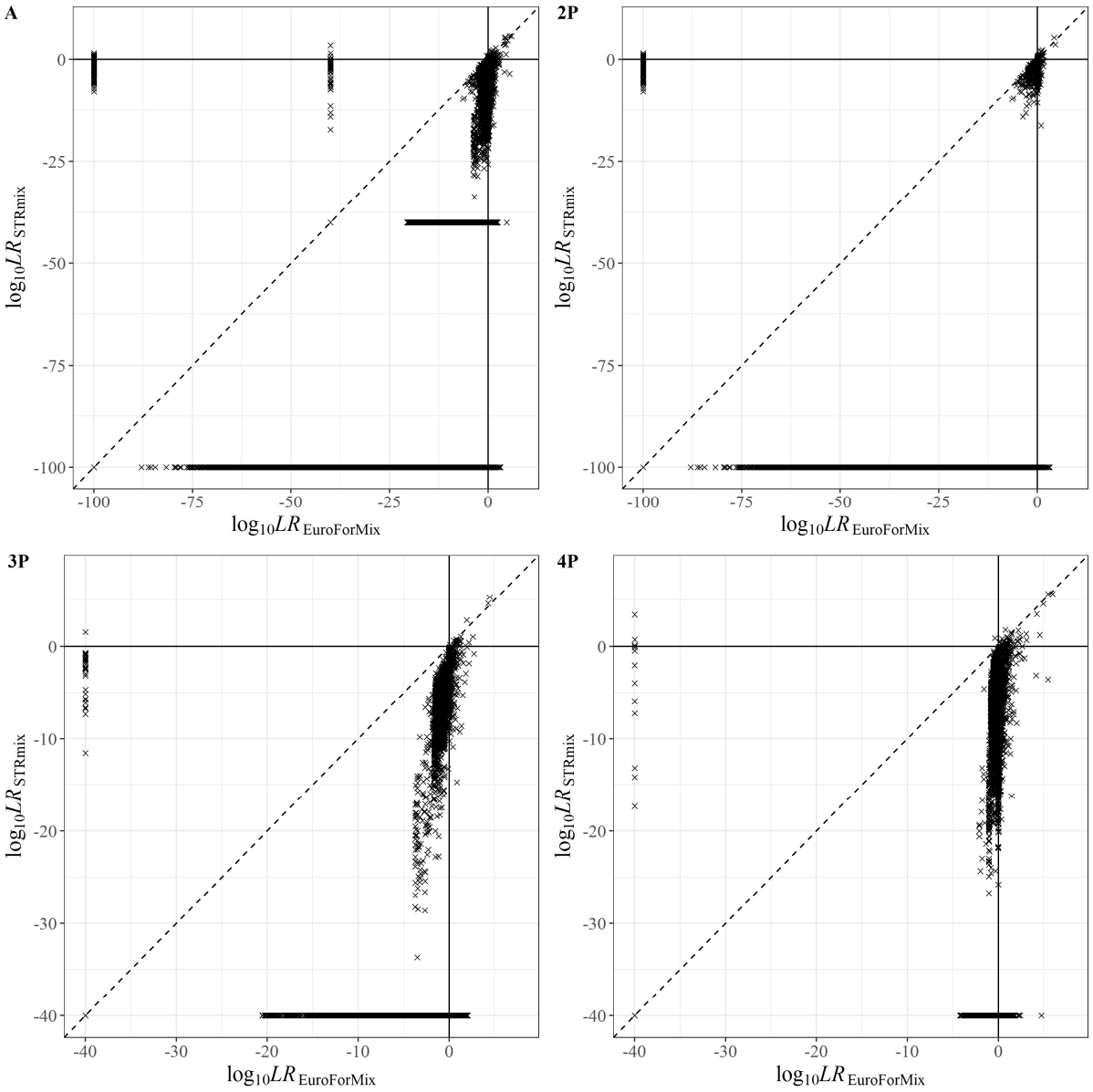


Figure 5: Scatter plot of the STRmix™ $\log_{10}LR$ and EuroForMix $\log_{10}LR$ for known contributors (circles) from the interpretation of all (panel A), two- (panel 2P), three- (panel 3P), and four-person (panel 4P) mixtures. The black arrows mark the five divergent results that were further investigated. The label shows the sample identifier followed by the donor identifier. The samples E03:K33 and H09:K30 marked with an asterisk (*) are described further in detail. A sample of interest (B07:K48) marked with a blue arrow and the ^ symbol is explained in the Supplementary Materials as Sample E.



391

392 Figure 6: Scatter plot of the STRmix™ $\log_{10}LR$ versus EuroForMix $\log_{10}LR$ for non-contributors from
393 the interpretation of all (panel A), two- (panel 2P), three- (panel 3P), and four-person (panel 4P)
394 mixtures. *LR*s of 0 are presented as -100 for two-person mixtures, and -40 for three- and four-person
395 mixtures.

396 The $\log_{10}LR$ for STRmix™ vs $\log_{10}LR$ for EuroForMix for the non-donors to the two-, three-, and
397 four-person mixtures are given in Figure 6. A plot the $\log_{10}LR$ for STRmix-$\log_{10}LR$ for EuroForMix
398 versus the average peak height (*APH*) per contributor is given in the supplementary material, Figure

S2. The results show that there is no general trend between the difference between $\log_{10}LR_{\text{STRmix}}$ and $\log_{10}LR_{\text{EuroForMix}}$ and *APH*.

In Figure 6, there were several observations of *LR*s assigned to known non-donors that exhibited an inclusionary *LR* in both software. Two of the highest inclusionary *LR*s to the known non-donors were investigated. These were *LR*s assigned to K2 to a four-person mixture (E08_RD14-0003-37_38_39_40-1;9;9;1-M2c-0.5GF-Q0.6_05.25sec) and K43 to a four-person mixture (F08_RD14-0003-37_38_39_40-1;9;9;1-M2c-0.3GF-Q0.6_06.25sec). The EuroForMix $\log_{10}LR$ to K2 is 5.4496 and the STRmix™ $\log_{10}LR$ is 5.6829. Thirty-three (33) of the 42 autosomal alleles for K2 were present in the mixture. The EuroForMix $\log_{10}LR$ to K43 is 5.9712 and the STRmix™ $\log_{10}LR$ is 5.6641. Thirty-six (36) of the 42 autosomal alleles for K43 were present in the mixture.

3.2.3. Analysis of divergent results

E03_RD14-0003-33_34-1;2-M3a-0.045GF-Q0.8_05.25sec

We selected the *LR* assigned to the minor contributor, K33, to the sample (E03_RD14-0003-33_34-1;2-M3a-0.045GF-Q0.8_05.25sec) from the two-person mixtures that showed a difference in the $\log_{10}LR$ between EuroForMix (14.59) and STRmix™ (10.59) (these are point estimates and for STRmix™ sub-source, $\theta = 0$). A per locus comparison of $\log_{10}LR$ is given in Table S7. The most divergent locus is D1S1656 (EuroForMix 1.37 versus STRmix™ 0.05). A snapshot of the electropherogram (epg) for this locus is shown in Figure 7. In the absence of the known genotypes, the combination of a 13,13 major and a 17.3,18.3 would be the most supported.

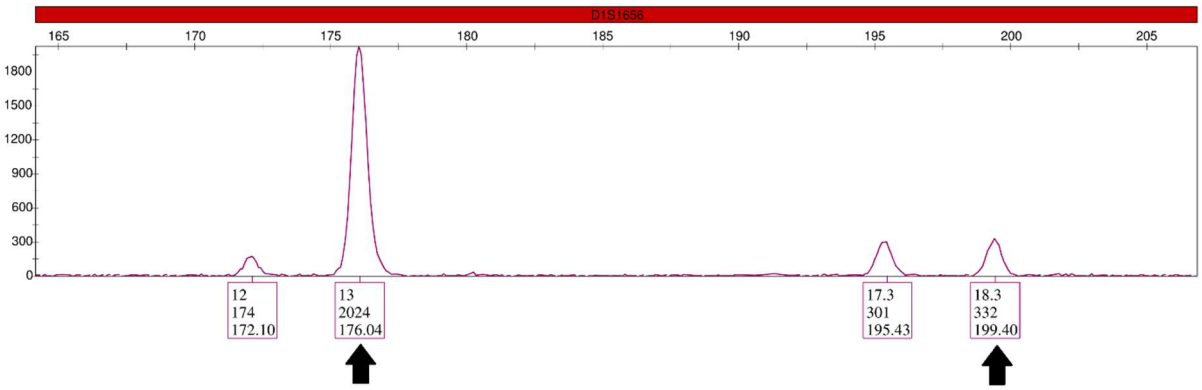


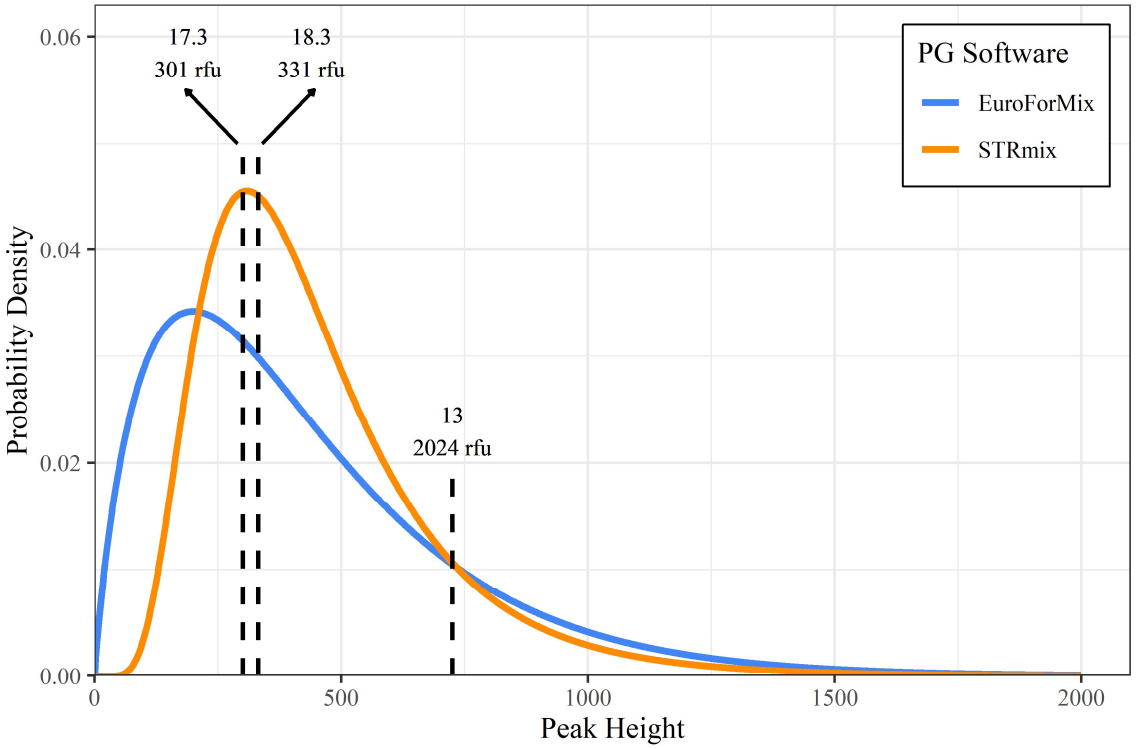
Figure 7: The D1S1656 locus of two-person mixed sample E03. This is targeted as a 2:1 mixture. The ground truth for the minor is 13,18.3 (indicated by the black arrows).

This mixture is made from references 34 (D1S1656 13,17.3) and 33 (D1S1656 13,18.3) targeted in a 2:1 ratio. STRmix™ gives estimated mixture proportions 77% (posterior mean template of 848 rfu) and 23% (posterior mean template of 251 rfu). The estimated mixture proportions under H_1 is 64% and 36% (0.66:0.33 under H_2). The combinations 13,17.3 and 13,18.3 for the minor would both be poorly supported and would have been excluded by some of the binary rules previously in use (24). Using the deconvolution function within EuroForMix, we are able to retrospectively collate weights for plausible genotype combinations under H_2 (see Table 4). Weights for STRmix™ are output natively in the interpretation process. Using these values, we give the relative probability densities of the evidence given the proposed genotype of the minor (Gm) and any genotype suggested for the major, termed support hereafter. On the other side, the weights from the maximum likelihood based deconvolution function within EuroForMix are proportional to the inner-sum terms which are evaluated in the *LR* calculation for the corresponding hypothesis (these weights are equivalent to the posterior genotype probabilities).

Table 4: The sum of the relative probability densities of the evidence given any major and the minor genotype (Gm) for EuroForMix (under H_2) and STRmix™ for the D1S1656 locus, termed support for two-person mixture E03.

Genotype of the minor (Gm)		Support/Weight	
		EuroForMix	STRmix™
17.3	18.3	32.6%	98.2%
13	17.3	14.3%	1.0%
13	18.3	12.1%	0.4%
Plus many other genotypes			

Consistent with the observed profile (but not the experimental design of the mixture), EuroForMix (under H_2) and STRmix™ both give the highest support to the 17.3,18.3 minor. However, EuroForMix is vastly more tolerant of the 13,17.3 or 13,18.3 minor. This can be tracked back to the peak height variability parameter in EuroForMix (P.H.variability) having adjusted to the relatively high value of 0.43. Some conversion is required to place the gamma distribution used by EuroForMix and the lognormal distribution used by STRmix™ on a comparable scale. This comparison is shown in Figure 8.



447

448 Figure 8: A comparison of probability density for peaks at their given heights (marked by the vertical
449 dashed lines) using the EuroForMix (EFM) ($\Gamma(1.88, 227.14)$) and STRmix™ (logarithmic transform
450 of $N\left(287.22, \frac{c^2}{287.22}\right)$) variance models for two-person mixed sample E03. The 13 allele height is
451 plotted at 2024×0.36=728 rfu, the expected height of a minor 13 allele proposed under EuroForMix
452 H_1. The observed height of the 13 allele is 2024 rfu. The value of 0.36 is the estimated mixture
453 proportion for the minor contributor under H_1.

454 STRmix™ supports the 17.3,18.3 minor and penalised the 13,17.3 or 13,18.3 minor more relative to
455 EuroForMix. The effect of a more tolerant peak variance parameter is a lower false exclusion rate
456 and a higher rate of false support for non-donors.

457 EuroForMix uses Maximum Likelihood Estimation (MLE) to set its peak height variance parameter
458 (phv). This is analogous to using a uniform [0,1] prior for the peak height variability parameter in
459 STRmix™. In contrast, STRmix™ uses a prior distribution based on implementation data, in this
460 case $\Gamma(8.45, 1.746)$. The posterior mean allele variance using this prior was 14.55. This means that

461 the phv in EuroForMix is free to move to any position that maximises the likelihood of the data
462 summed across all genotype sets, i.e. the marginalized sum. The STRmix™ equivalent parameter is
463 partially constrained. It can move, but the further it gets from the mode of the distribution, the larger
464 the penalty. This has the effect of pulling the proposed variance estimate back from extreme values.
465 We believe that it is this modelling difference that drives the difference in the $\log_{10}LR$ between
466 EuroForMix and STRmix™ for this sample.

467 We tested this hypothesis by using a nearly flat prior for STRmix™, $\Gamma(1, 100)$. The resulting locus
468 by locus $\log_{10}LR$s and the $\log_{10}LR$ across all loci are given in Table S7. The majority of locus
469 $\log_{10}LR$s and the overall $\log_{10}LR$ for EuroForMix (14.6) and STRmix™ (14.8) are closer using the
470 nearly flat prior in STRmix™. The posterior mean allele variance using a flat prior in STRmix™ was
471 33.43. The difference, then, is seen as not a property of the software but a judgement about whether
472 we should expect future casework samples to be similar to validation samples or to have no
473 relationship with the validation samples and further to take on any value at all.

474 Motivated by this observation, we were interested to see if EuroForMix would show increased rates of
475 adventitious support for non-donors that are a poor fit to the observed peak heights. To study this, we
476 created a high-risk database of 200 non-donors by sampling with replacement from the alleles of the
477 two true donors. For example, at vWA the known donors' alleles are [16,17] and [17,18]. When a
478 non-donor's genotype was generated for this locus, two alleles were sampled with replacement from
479 the alleles 16, 17, and 18. This sampling was undertaken at each locus independently to create a full
480 GlobalFiler profile for the non-donor. This is a different database to the one described in section
481 2.3.2. The 200 non-donors from this high-risk database were compared with the mixture E03 using
482 EuroForMix and STRmix™ (using the informed prior). This would rarely impact casework since the
483 probability that any individual would have two genotypes with this level of overlap to this mixture is
484 about 1.6×10^{-13} (unrelated individuals) and 9×10^{-6} (siblings). We are therefore looking at a tail of the
485 H_2 true distribution. The results are shown in Figure 9. The LRs assigned to the high-risk database
486 were all higher for EuroForMix than STRmix™. Also see (25) for further non-donor and sibling

comparisons for EuroForMix and (26, 27) for a description of further non-donor tests using STRmix™.

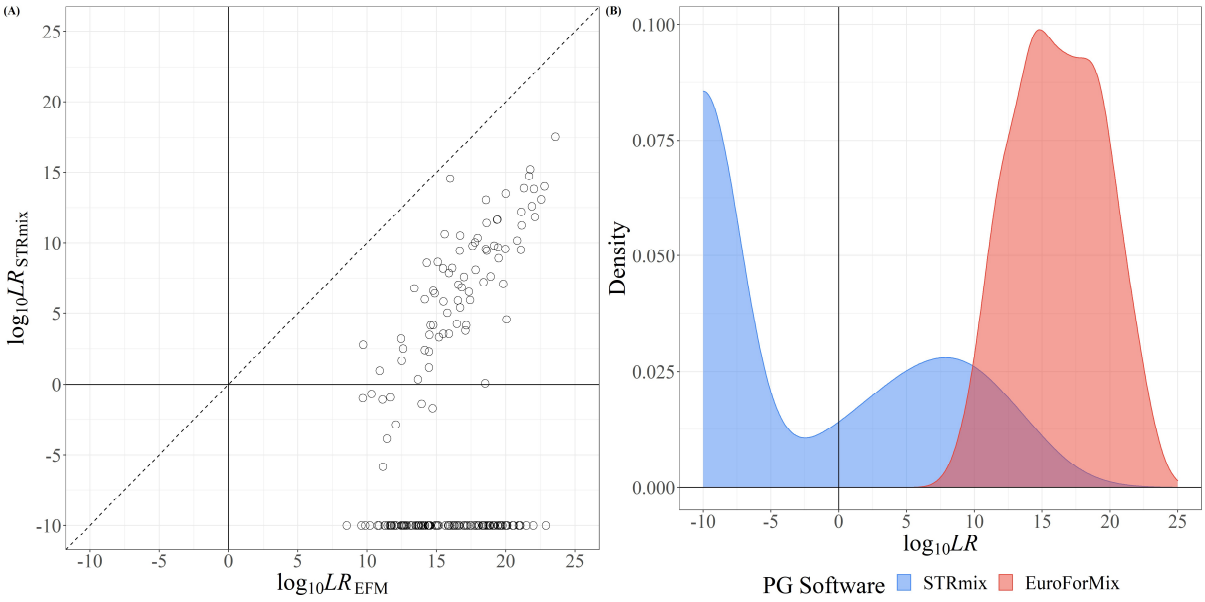


Figure 9: (A) A plot of the $\log_{10}LR$s produced for the 200 simulated false donors using STRmix™ (implemented prior) and EuroForMix for two-person mixed sample E03. (B) A density plot of $\log_{10}LR$ for the 200 simulated false donors using STRmix™ (implemented prior) and EuroForMix. Where $LR = 0$ were plotted as $\log_{10}LR = -10$. STRmix™ had 73 LRs > 1, EuroForMix had 200 LRs > 1.

The remaining differences in the three additional two person mixtures examined (B10:K39, E10:K39, and H01:K32 marked with a black arrow) could all be attributed to the greater tolerance of peak height differences in EuroForMix, similar to that observed in this example.

Furthermore, the models within STRmix™ considers locus specific amplification efficiencies and expected stutter ratios are determined using empirical data. Whereas, EuroForMix does not consider these locus specific amplifications efficiencies and has a blanket expected stutter rate. This may be one of the contributing reasons why the peak height variance in EuroForMix needs to be more tolerant. An example of this modelling difference is explored in sample H09_RD14-0003-30_31_32-1;4;4-M2d-0.75GF-Q0.6_08.25sec below.

H09_RD14-0003-30_31_32-1;4;4-M2d-0.75GF-Q0.6_08.25sec

We also selected for review one sample (H09_RD14-0003-30_31_32-1;4;4-M2d-0.75GF-Q0.6_08.25sec) from the three-person mixtures that showed a substantial difference in the $\log_{10}LR$ assigned to the minor contributor K30 between EuroForMix (11.85) and STRmix™ (21.36); these are point estimates and for STRmix™ sub-source, $\theta = 0$. These values are both greater than the $\log_{10}LR$=9 threshold implemented by some laboratories. H09 is targeted as a 4:4:1 mixture and the STRmix™ posterior mean mixture proportions from the MCMC sampling process are 49:39:11 (approx. 5:4:1). EuroForMix gives mixture proportions 40:40:20 (2:2:1) under H_1, but near 1:1:1 under H_2.

A per locus comparison of $\log_{10}LR$ and the $\log_{10}LR$ across all loci are given in Table S8. The most divergent locus is FGA (EuroForMix 0.07 versus STRmix™ 1.68). The epg for this locus appears in Figure 10.

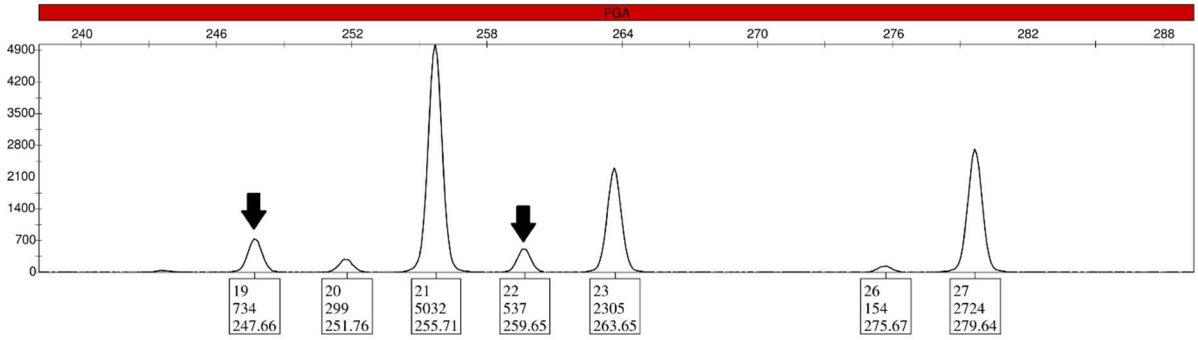


Figure 10: The epg for the FGA locus for three-person mixture H09. The ground truth for the minor is 19,22 (indicated by black arrows). The peak at 22 is 23% of the height of the 23 allele. The average stutter ratio for FGA 23 is 7.3%.

After inspection of Figure 10 and Table 5, STRmix's support or weight of genotypes at locus FGA appear intuitive whereas EuroForMix has treated the 22 peak as potentially an allele belonging to one of the other contributors. The peak at 22 is 23% of the height of the 23 peak. Using the STRmix™ kit settings, the expected stutter ratio for FGA 23 for the PROVEDIt single source GlobalFiler data is

524 7.3%. EuroForMix has spread the support for the second minor allele more broadly than STRmix™.
525 Given the size of the 22 peak it seems reasonable to expect the majority of the support to be on the
526 19,22 genotype for the minor. A possible reason for this discrepancy is that EuroForMix is unable to
527 resolve the differences in the mixture proportion of the minor contributor under H_2.

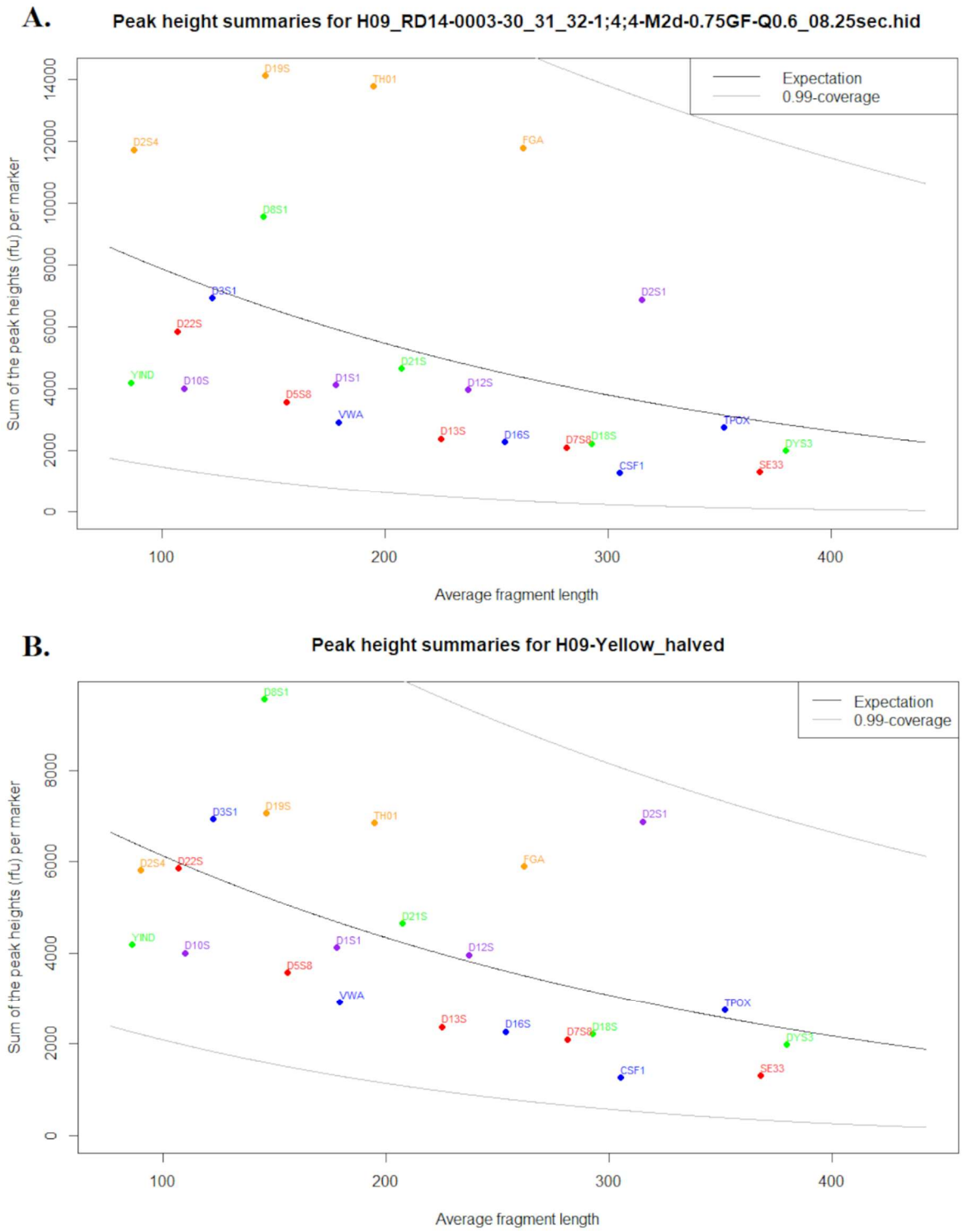
528 An analysis of the peak heights at each locus (see Figure 11) suggests that the loci with yellow dye are
529 much higher than the trend line for the other loci. STRmix™ models locus specific amplification
530 efficiencies, to allow for differential amplification between loci. EuroForMix does not model locus
531 specific amplification efficiencies. A large difference in total peak heights for one or a few loci
532 would increase the phv in EuroForMix and the increased peak height variation leads to the non-
533 resolved mixture under H_2 that is observed above. We test this hypothesis by artificially reducing the
534 heights of the four yellow dye loci. The results appear in Table 5 and Table S8.

535 Table 5: The support for the minor genotype (Gm) for EuroForMix (under H_2) and STRmix™ for
536 three-person mixture H09, locus FGA

	Support/Weight			
	Original evidence		Yellow dye peak heights halved	
Genotype of the minor Gm	EuroForMix	STRmix™	EuroForMix	STRmix™
19,22	2.72%	88.45%	13.85%	99.67%
19,27	3.33%	7.67%	0.82%	0.16%
19,21	5.75%	2.15%	10.47%	0.12%
19,23	4.63%	1.31%	8.34%	0.04%
19,19	0.09%	0.41%	1.14%	0.01%

537 Table 6: The allele variability parameters for STRmix™ and EuroForMix for three-person mixture
538 H09. These values are not directly comparable and need translation to the same scale for comparison
539 between STRmix™ and EuroForMix. However, the comparison between the yellow dye peak heights
540 halved and not halved is valid.

	Original evidence	Yellow dye peak heights halved
EuroForMix peak height variability	0.46839	0.33101
Posterior mean of STRmix™ allele variance parameter c^2	9.291	8.254



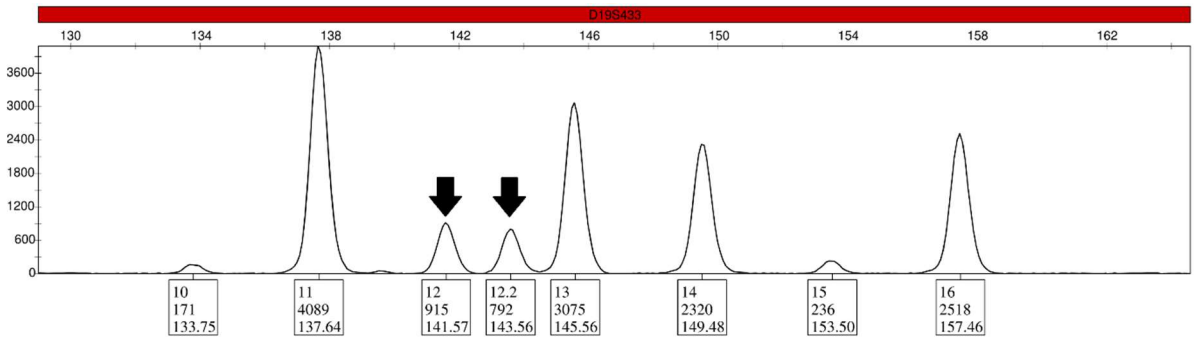
541

542 Figure 11: A plot of the sum of the peak heights per marker versus average fragment length (bp) for
543 three-person mixture H09 (Panel A) and the same sample where the peak heights in the yellow dye
544 channel are halved (Panel B). Panel A shows the high total peak height for the yellow dye loci.

The reduction of the height of the yellow dye loci has, as hoped, taken the pressure off the allele variability parameter (phv). This is shown in Table 6. The mixture proportions reported by the two software with the yellow dye loci halved are 49:41:10 (approx. 5:4:1) for STRmix™, and 43:43:15 (approx. 4:4:1) under H_1 and 44:44:12 (approx. 4:4:1) under H_2 for EuroForMix. This is a marked improvement for EuroForMix from its initial values of 1:1:1 under H_2. Target ratios are 4:4:1.

Halving the yellow dye peak heights has moved the *LR*s for STRmix™ and EuroForMix closer (Table S8) but they still differ by nearly seven-orders of magnitude (log*LR* EuroForMix 15.5 and STRmix™ 22.1). The peak height variance for EuroForMix is larger than the equivalent for STRmix™ (see Figure S6). We hypothesise that this is the reason that EuroForMix has spread its support across more genotypes for the minor than STRmix™. This is still driving the difference between EuroForMix and STRmix™ in the resulting *LR*s. Furthermore, upon inspection of the model validation PP-plots in Figure S7, we can see that under H_1 (or H_1 in the figure) for both the original evidence and the evidence with half the yellow dye peak heights, several observations deviate far from the identity line. This indicates that the EuroForMix model did not fit well.

We also examine D19S433 for the three-person mixture H09 which has a high *LR* (EuroForMix 2.61 and STRmix™ 3.71). In Figure 12 we show the epg for locus D19S433. The genotype of the known minor is 12,12.2. Looking at the epg of this locus subjectively one might assign the minor as including the 12.2 allele unambiguously and the 12 with very high confidence since the peak at 12 is 30% of the height of the 13 allele. The average observed back stutter ratio for D19S433 13 for the single-source PROVEDIt GlobalFiler profiles is 5.2%: The estimated stutter proportion in EuroForMix under H_2 was 12%. STRmix™ gives all its support to the genotype 12,12.2 for the minor whereas EuroForMix distributes its support over various options largely containing the 12.2 but not necessarily the 12 (see Table 7); the allele is explained as an elevated stutter from allele 13 which is most likely to originate from two donors (shared). The mixture is targeted as 4:4:1 and this is obtained by STRmix™. EuroForMix obtains 1:1:1 under H_2 with the original inputs, but close to 4:4:1 under H_2 with the yellow dye loci halved.



571

572 Figure 12: The epg for the D19S433 locus for three-person mixture H09. The ground truth for the

573 minor is 12,12.2 (indicated by black arrows). The peak at 12 is 30% of the height of the 13 peak.

574

Table 7: The support for various minor genotypes at locus D19S433 of the three-person mixed DNA profile H09 using STRmix™ and EuroForMix for the original epg and with the yellow dye peak heights halved.

	EuroForMix		STRmix™	
Genotype	Original evidence	Yellow dye peak heights halved	Original evidence	Yellow peak heights halved
12.2,14	8.35%	27.58%	0%	0%
12.2,13	8.43%	25.85%	0%	0%
12,12.2	1.93%	16.91%	100.00%	100.00%
12.2,15	1.05%	10.99%	0%	0%
12.2,16	6.82%	6.17%	0%	0%
14,14	1.90%	1.85%	0%	0%

4. Discussion

After taking into account the differences in allele probability models, the *LRs* from EuroForMix and STRmix™ for single-source profiles were the same to at least two significant figures. For a fully-resolvable single-source profile they were the same to four significant figures for $\theta \in \{0.00, 0.01\}$. As shown in **Error! Reference source not found.**, *LRs* from EuroForMix and STRmix™ for a fully-resolved single-source profile with two previously unobserved alleles were the same to three significant figures for θ=0.01 and differed by three orders of magnitude when θ=0. This difference was due to the different models for assigning the minimum allele probability within the two software. The *LR*s were the same to two significant figures for a single source profile with drop-in. For the partial profile with dropout the *LR*s differed in the second significant figure.

For both software, the *LR* assigned for a single-source dilution series decreased towards 1 as the target input amount decreased. The *LR*s for EuroForMix and STRmix™ were all within one order of

591 magnitude of each other. The largest difference was where the target input amount was 0.0156 ng.
592 The EuroForMix *LR* was 2.1×10^{25} and 8.0×10^{24} for the STRmix™ *LR*.

593 The results from the experiments involving single-source profiles are reassuring. It demonstrates that
594 even though the models implemented in each of the PG software are different, both software give
595 similar answers when $\theta>0$. Additionally, because the *LR*s for the unambiguous single-source profiles
596 can be replicated in MS Excel™ for both software to the fourth significant figure, this shows that the
597 *LR* calculation is performing as expected.

598 Similarly, the observations from the single-source dilution series, and the experiment involving two-
599 person mixtures of varying mixture proportions demonstrate that both PG software are performing as
600 expected. The *LR* increases with increasing template information, although if the peak heights of the
601 donors are similar, this can create ambiguity resulting in a decreased *LR* in comparison to a clear
602 major:minor mixture. The results presented in the two PG software show similar and intuitive trends.

603 Within these experiments we were also able to detail some of the key differences between the two
604 software. In section 3.2.2 within the sensitivity and specificity experiments, we demonstrate the
605 sensitivity and specificity for a range of PROVEDIt two-, three-, and four-person mixtures using both
606 PG software. Similar to the experiments above, Figure 2 to Figure 4 show similar trends in the *LR* for
607 both PG software.

608 Gill et al. (28) described the use of receiver operating characteristic (ROC) plots to compare the
609 performance of different models. We have chosen not to plot ROC plots, as our plots show that as the
610 *APH* increases, the *LR*s assigned to known donors to the mixture and *LR*s assigned to the non-donors
611 become reasonably well separated for this set of mixtures. As the number of contributors increased,
612 and the *APH* lowered, the distributions of *LR*s for known donors and non-donors begin to converge on
613 $LR = 1$. We have also explored the behaviour of the *LR* versus the two definitions of allele sharing,
614 where we show that as the amount of allele sharing increases, the *LR*s begin to decrease.

615 Review of Figure 5 shows similar *LR*s between the software for many of the same comparisons.
616 Because the models are different, divergent results are to be expected. Ignoring profiles with rare
617 alleles where significant differences in the *LR* between the two software were observed (three orders
618 of magnitude within a full, single source profile) 84% of *LR*s were within two orders of magnitude.
619 Part of the goal for this work was to identify factors driving any difference in the assigned *LR*
620 between the two software. This was explored in six divergent results, where we identified differences
621 in the peak height variance model, locus specific amplification model, and the stutter model.

622 An observation from the specificity study was that the *LR*s assigned to the non-donors were mostly
623 lower using the STRmix™ PG software. This may be because EuroForMix has a more tolerant peak
624 height variance model in comparison to STRmix™, and EuroForMix uses an MLE approach to
625 evaluate H_1 and H_2 separately. EuroForMix maximizes the likelihood under H_1 and H_2 independently.
626 For example, H_1 could be considering POI + 2U and H_2 is considering 3U. Because the MLE is
627 maximizing the likelihood separately, different parameter values (such as mixture proportions) can be
628 observed.

629 We also found an example where EuroForMix, under H_2, explained an allele of a minor contributor in
630 back stutter position as an elevated stutter (Figure 12) – consequentially reducing the *LR* for the
631 corresponding marker. The prior distributions for the stutter parameters (global) were specified with a
632 uniform distribution (default). By specifying a non-uniform distribution instead, for example
633 assigning more weight to stutter peaks below 10%, would possibly improve modelling.

634 As part of this work an important miscode was discovered in EuroForMix (versions 3.0.0 - 3.2.0)
635 regarding the stutter models. The miscode was triggered when the observed alleles at a locus fully
636 overlapped with the alleles observed in the allele frequency database, leading to the wrong indexing
637 for the stutter-relation vectors. The miscode was discovered when carefully comparing the results for
638 the four-person mixture H09_RD14-0003-48_49_50_29-1;4;4;4-M2a-0.75GF-Q0.4_08.25sec. A
639 substantial difference in the $\log_{10}LR$ between EuroForMix (-1.75) and STRmix™ (19.79) were

obtained. In our study, 13 4-person mixtures were affected by the miscode, with 9 of the differences being greater than three orders of magnitude (Supplementary materials Figure S8).

Returning to the quote by George Box, the similarity between the two set of results demonstrate that even though there are different assumptions and models within the two software, both can be useful in assigning in *LR*. The results of sensitivity and specificity studies can help inform the limits of the PG software for a given laboratory. Additionally, analysts operating PG software tools should review the results and any diagnostic values for intuitiveness.

We have not further examined the effect of assuming the presence of the POI under H_1 in the interpretation that is used in EuroForMix or the effect of separate analysis of the parameters under H_1 and H_2. Contrastingly, STRmix™ interprets a DNA profile in the absence of knowledge of the POI's reference profile. It is only after the interpretation, when the *LR* is assigned using a set of propositions where the POI is assumed under H_1.

The divergences identified here for the first time with fully qualified operators using the two commonly used PG software are undoubtedly real. However, we are very positive about both the diagnosability of the cause of the differences and the likelihood of being able to build from this study towards more convergent and robust solutions.

References

1. Bleka Ø, Storvik G, Gill P. 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 2016 Mar 1;21:35-44. https://doi.org/10.1016/j.fsigen.2015.11.008

2. Cowell RG, Graversen T, Lauritzen SL, Mortera J. Analysis of forensic DNA mixtures with artefacts. J R Stat Soc C-Appl 2015 Jan 1;64(1):1-48. https://doi.org/10.1111/rssc.12071

3. Manabe S, Morimoto C, Hamano Y, Fujimoto S, Tamaki K. Development and validation of open-source software for DNA mixture interpretation based on a quantitative continuous model. PLoS One 2017 Nov 17;12(11):e0188183. https://doi.org/10.1371/journal.pone.0188183

4. Kelly H, Bright J-A, Buckleton JS, Curran JM. A comparison of statistical models for the analysis of complex forensic DNA profiles. Sci Justice 2014 Jan 1;54(1):66-70. https://doi.org/10.1016/j.scijus.2013.07.003

5. Haned H, Gill P, Lohmueller K, Inman K, Rudin N. Validation of probabilistic genotyping software for use in forensic DNA casework: Definitions and illustrations. Sci Justice 2016 Mar 1;56(2):104-8. https://doi.org/10.1016/j.scijus.2015.11.007

6. Steele CD, Balding DJ. Statistical Evaluation of Forensic DNA Profile Evidence. Annu Rev Stat Appl 2014;1:361-84.

7. Swaminathan H, Garg A, Grgicak CM, Medard M, Lun DS. CEESIt: A computational tool for the interpretation of STR mixtures. Forensic Sci Int Genet 2016 May 1;22:149-60. https://doi.org/10.1016/j.fsigen.2016.02.005

8. Taylor D, Bright J-A, Buckleton J. The interpretation of single source and mixed DNA profiles. Forensic Sci Int Genet 2013 Sep 1;7(5):516-28. https://doi.org/10.1016/j.fsigen.2013.05.011

9. Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, et al. Validating TrueAllele® DNA mixture interpretation. J Forensic Sci 2011 Nov 1;56:1430-47. https://doi.org/10.1111/j.1556-4029.2011.01859.x

10. Box GEP, Draper NR. Empirical model-building and response surfaces. New York: Wiley, 1987;74, 424

11. Bright J-A, Taylor D, McGovern CE, Cooper S, Russell L, Abarno D, et al. Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles. Forensic Sci Int Genet 2016 Jul 1;23:226-39. https://doi.org/10.1016/j.fsigen.2016.05.007

12. Alladio E, Omedei M, Cisana S, D'Amico G, Caneparo D, Vincenti M, et al. DNA mixtures interpretation – A proof-of-concept multi-software comparison highlighting different probabilistic methods' performances on challenging samples. Forensic Sci Int Genet 2018 Nov 1;37:143-50. https://doi.org/10.1016/j.fsigen.2018.08.002

13. The People of the State of New York versus Oral N. Hillary. State of New York County Court; 2015.

14. Bright J-A, Evett IW, Taylor D, Curran JM, Buckleton J. A series of recommended tests when validating probabilistic DNA profile interpretation software. Forensic Sci Int Genet 2015 Jan 1;14:125-31. https://doi.org/10.1016/j.fsigen.2014.09.019

15. Benschop CCG, Hoogenboom J, Hovers P, Slagter M, Kruise D, Parag R, et al. DNAxs/DNAStatistX: Development and validation of a software suite for the data management and probabilistic interpretation of DNA profiles. Forensic Sci Int Genet 2019 Sep 1;42:81-9. https://doi.org/10.1016/j.fsigen.2019.06.015

16. Swaminathan H, Qureshi MO, Grgicak CM, Duffy K, Lun DS. Four model variants within a continuous forensic DNA mixture interpretation framework: Effects on evidential inference and reporting. PLoS One 2018 Nov 20;13(11):e0207599. https://doi.org/10.1371/journal.pone.0207599

17. President's Council of Advisors on Science and Technology. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods – 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf (accessed April 27, 2017)

18. Riman S, Iyer H, Vallone PM. Examining Discrimination Performance and Likelihood Ratio Values for Two Different Likelihood Ratio Systems Using the Provedit Dataset. bioRxiv 2021 May 27:2021.05.26.445891. https://doi.org/10.1101/2021.05.26.445891

19. Hill CR, Duewer DL, Kline MC, Coble MD, Butler JM. U.S. population data for 29 autosomal STR loci. Forensic Sci Int Genet 2013 May 1;7(3):e82-e3. https://doi.org/10.1016/j.fsigen.2012.12.004

20. Alfonse LE, Garrett AD, Lun DS, Duffy KR, Grgicak CM. A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt. Forensic Sci Int Genet 2018 Jan 1;32:62-70. https://doi.org/10.1016/j.fsigen.2017.10.006

21. Russell L, Cooper S, Wivell R, Kerr Z, Taylor D, Buckleton J, et al. A guide to results and diagnostics within a STRmix™ report. WIREs Forensic Science 2019 Nov 1;1(6):e1354. https://doi.org/10.1002/wfs2.1354

22. Bille TW, Weitz SM, Coble MD, Buckleton JS, Bright J-A. Comparison of the performance of different models for the interpretation of low level mixed DNA profiles. Electrophoresis 2014 Nov 1;35(21-22):3125-33. https://doi.org/10.1002/elps.201400110

23. Bleka Ø, Benschop CCG, Storvik G, Gill P. A comparative study of qualitative and quantitative models used to interpret complex STR DNA profiles. Forensic Sci Int Genet 2016 Nov 1;25:85-96. https://doi.org/10.1016/j.fsigen.2016.07.016

24. Bill M, Gill P, Curran J, Clayton T, Pinchin R, Healy M, et al. PENDULUM - A guideline based approach to the interpretation of STR mixtures. Forensic Sci Int 2005 Mar 10;148(2-3):181-9. https://doi.org/10.1016/j.forsciint.2004.06.037

25. Benschop CCG, Nijveld A, Duijs FE, Sijen T. An assessment of the performance of the probabilistic genotyping software EuroForMix: Trends in likelihood ratios and analysis of Type I & II errors. Forensic Sci Int Genet 2019 Sep 1;42:31-8. https://doi.org/10.1016/j.fsigen.2019.06.005

26. Noël S, Noël J, Granger D, Lefebvre J-F, Séguin D. 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 2019 Jul 1;41:24-31. https://doi.org/10.1016/j.fsigen.2019.03.017

27. Bright J-A, Richards R, Kruijver M, Kelly H, McGovern C, Magee A, et al. Internal validation of STRmix™ – A multi laboratory response to PCAST. Forensic Sci Int Genet 2018 May 1;34:11-24. https://doi.org/10.1016/j.fsigen.2018.01.003

737 28. Gill P, Bleka Ø, Hansson O, Benschop C, Haned H. Chapter 9 - Validation. In: Gill P, Bleka
738 Ø, Hansson O, Benschop C, Haned H, editors. Forensic Practitioner's Guide to the Interpretation of
739 Complex DNA Profiles: Academic Press; 2020;277-308.

740