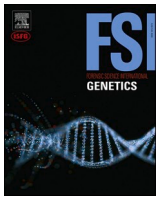
Forensic Science International: Genetics 59 (2022) 102715

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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



Short communication

Quantification of forensic genetic evidence: Comparison of results obtained by qualitative and quantitative software for real casework samples

Camila Costa [a,b,*], Carolina Figueiredo [a,b], António Amorim [a,b,c], Sandra Costa [d], Paulo Miguel Ferreira [d], Nádia Pinto [b,c,e]

[a] Faculdade de Ciências, Universidade do Porto, Portugal
[b] i3S - Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal
[c] IPATIMUP - Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Portugal
[d] Biologia, Laboratório de Polícia Científica da Polícia Judiciaria (LPC-PJ), Lisboa, Portugal
[e] CMUP, Centro de Matemática da Universidade do Porto, Portugal

ARTICLE INFO

Keywords:
Mixture interpretation
Weight-of-evidence
Genotyping software
LRmix Studio
EuroForMix
STRmix

ABSTRACT

To overcome the multifactorial complexity associated with the analysis and interpretation of the capillary electrophoresis results of forensic mixture samples, probabilistic genotyping methods have been developed and implemented as software, based on either qualitative or quantitative models. The former considers the electropherograms' qualitative information (detected alleles), whilst the latter also takes into account the associated quantitative information (height of allele peaks). Both models then quantify the genetic evidence through the computation of a likelihood ratio (LR), comparing the probabilities of the observations given two alternative and mutually exclusive hypotheses.

In this study, the results obtained through the qualitative software LRmix Studio (v.2.1.3), and the quantitative ones: STRmix™ (v.2.7) and EuroForMix (v.3.4.0), were compared considering real casework samples. A set of 156 irreversibly anonymized sample pairs (GeneMapper files), obtained under the scope of former cases of the Portuguese Scientific Police Laboratory, Judiciary Police (LPC-PJ), were independently analyzed using each software. Sample pairs were composed by (i) a mixture profile with either two or three estimated contributors, and (ii) a single contributor profile associated. In most cases, information on 21 short tandem repeat (STR) autosomal markers were considered, and the majority of the single-source samples could not be a priori excluded as belonging to a contributor to the paired mixture sample. This inter-software analysis shows the differences between the probative values obtained through different qualitative and quantitative tools, for the same input samples. LR values computed in this work by quantitative tools showed to be generally higher than those obtained by the qualitative. Although the differences between the LR values computed by both quantitative software showed to be much smaller, STRmix™ generated LRs are generally higher than those from EuroForMix. As expected, mixtures with three estimated contributors showed generally lower LR values than those obtained for mixtures with two estimated contributors.

Different software products are based on different approaches and mathematical or statistical models, which necessarily result in the computation of different LR values. The understanding by the forensic experts of the models and their differences among available software is therefore crucial. The better the expert understands the methodology, the better he/she will be able to support and/or explain the results in court or any other area of scrutiny.

1. Introduction

Whenever biological material is found during a criminal investigation, the standard practice in most countries is to perform DNA analysis by capillary electrophoresis, after PCR amplification. DNA mixtures, i.e., samples containing contributions from more than one donor (exact number unknown), are typically encountered in this context. An estimation of the number of contributors of the DNA mixture is a parameter

* Correspondence to: i3S, R. Alfredo Allen 208, 4200-135 Porto, Portugal.
E-mail address: camila75@live.com.pt (C. Costa).

https://doi.org/10.1016/j.fsigen.2022.102715
Received 6 December 2021; Received in revised form 29 March 2022; Accepted 21 April 2022
Available online 26 April 2022

required by probabilistic genotyping systems to evaluate the weight of the evidence. Although alternative methods have been proposed [1–16], the standard methodology to estimate the number of contributors is the maximum allele count approach [17–20], where the expert establishes the minimum number of contributors based on the number of observed alleles. This estimation may be a challenging procedure, especially when involving high order mixtures (three or more contributors) and/or low levels of DNA, and may not reflect the actual number of contributors.

Beyond the unknown number of contributors, mixture DNA samples may have associated several other characteristics that make it complex and difficult to interpret and analyze, such as allele sharing between contributors, contributions in different proportions, stutters which may be confounded with alleles of a minor contributor (or vice versa), amplification stochastic effects (heterozygotic imbalance, drop-in, and drop-out) due to low-template DNA, or degradation [21–32].

To minimize the impact of the abovementioned conditions and quantify the weight of the forensic genetic evidence under a robust and scientifically supported statistical approach, probabilistic genotyping informatics tools were developed. The standard approach relies on the computation of a Likelihood Ratio (LR) [24], comparing the probabilities of observing the genetic evidence, assuming two alternative and mutually exclusive hypotheses.

Depending on the model on which they are based, existing software can be divided into two major groups: qualitative (e.g., Lab Retriever [33] or LRmix Studio [34]) or quantitative (e.g., EuroForMix [35] or STRmix™ [36]). The qualitative models only consider the qualitative information of the electropherogram (observed alleles in the profile, after the screening of the expert), while the quantitative ones also consider the quantitative information (height of the detected peaks, both alleles and stutters). Therefore, the first approach has simpler algorithms and computations compared to the second, which is more complete and produces complex results that are more sensitive to artifacts and stochastic effects [37].

Several works have been developed to evaluate and compare the performance of the different existing software resorting to controlled, mocked, mixture samples [38–51]. In this case, and in the absence of contamination, the number of contributors and respective proportion of the analyzed mixture samples is, a priori, known.

Under the framework of forensic identification problems, the main goal of this paper is to compare the results obtained by widely used probabilistic genotyping programs, considering real casework samples. This type of samples was chosen due to their unknown composition, uniqueness, and inability to predict and replicate, carrying a much higher level of complexity than mock ones. The software used in this work were based on qualitative – LRmix Studio v.2.1.3 [34] – and quantitative models – EuroForMix v. 3.4.0 [35] and STRmix™ v.2.7 [36]. Both LRmix Studio (https://github.com/smartrank/lrmixstudio) and EuroForMix (http://www.euroformix.com/) are open-source and free of charge. EuroForMix works in R [52] and assumes either Maximum Likelihood Estimation (MLE) or Bayesian inference approaches. Gamma distribution is assumed for modeling peak heights, and Exponential is used for drop-in occurrence. On other hand, STRmix™ (http://strmix.esr.cri.nz/) is a commercial software based on Bayesian approach using Markov Chain Monte Carlo (MCMC). A log-normal distribution is assumed for the peak height and the drop-in is modeled using either a Gamma or Uniform distribution. Indeed, despite both being quantitative software, there are significant differences on the considered approaches to analyze the data, which necessarily results in different evidence quantifications – for a comprehensive review see [50].

2. Methods and materials

A set of 156 irreversibly anonymized mixture/single contributor sample pairs (mixtures with two and three estimated contributors equally distributed) were selected from real casework and analyzed at Portuguese Scientific Police Laboratory, Judiciary Police (LPC-PJ). For most of the pairs, the single contributor sample could not be excluded from being from one of the respective mixture contributors. The samples were previously processed in the context of the respective casework, under manufacturer protocols. The extraction was carried out in the Automate Express™ Forensic DNA Extraction System (Applied Biosystems™), using the PrepFiler Express BTA™ Forensic DNA Extraction kit (Applied Biosystems™) and QIAsymphony SP/AS Extraction System using the QIAsymphony DNA Investigator Kit, followed by the quantification in the equipment 7500 Real-Time PCR System (Applied Biosystems™), using the Quantifiler™ Trio DNA Quantification kit (Applied Biosystems™). Then, the amplification was performed in thermal cycler GeneAmp® PCR System 9700 (Applied Biosystems™) using GlobalFiler™ PCR Amplification Kit and GlobalFiler™ Express PCR Amplification Kit (Applied Biosystems™), both 24-locus STR kits; and, finally, the PCR products were detected and separated by capillary electrophoresis in 3500 Genetic Analyzer with GeneMapper™ ID-X (Applied Biosystems™), with an analytical threshold equal to 100 RFU. This work was then developed considering the genetic information obtained (GeneMapper files) for each irreversibly anonymized pair of samples (mixture/single contributor) for the set of 21 autosomal short tandem repeat markers analyzed.

For each mixture/single contributor sample pair, the quantification of the genetic evidence was computed through a global likelihood ratio (LR) assuming the following alternative hypotheses: H_1: "DNA originated from the person of interest (POI) and N-1 unknown, unrelated individual(s)", and H_2: "DNA originated from N unknown, unrelated, individuals", where N represented the estimated number of contributors. The LR value was computed for the same set of markers using LRmix Studio v.2.1.3 [34], EuroForMix v.3.4.0 [35] and STRmix™ v.2.7 [36]. In all the cases, the single contributor sample was assumed to come from the POI. In EuroForMix, the MLE approach was employed. All LR values were computed using the allelic frequencies of the National Institute of Standards and Technology (NIST) database concerning the Caucasian population [53], available at https://strbase.nist.gov//NISTpop.htm.

To compare the results obtained by the three software, as much as possible similar conditions were maintained for the parameters described in Table 1. A co-ancestry coefficient (F_{ST}) equal to 0.01 was considered in all the tools; this value was used by studies with similar populations [45,54,55] and recommended for a broad geographic group by the second National Research Council report (NRC II) [56]. The drop-in frequency was set to 0.05 in all the programs, a value considered in similar study [54], coinciding with the pre-set value in LRmix Studio. Drop-in peaks, the peak heights are modeled in both quantitative computer tools [37], EuroForMix using a Lambda distribution (λ), where the software's default value of 0.01 was used, and STRmix™ employs either a Gamma (γ) or Uniform distribution, the latter having been considered in this work. A minimum allele frequency of 0.001 (default value in LRmix Studio) was used in both LRmix Studio and EuroForMix, while in STRmix™ this value is locus-specific and a N equal to 0 was considered.

Table 1
Parameter values introduced in LRmix Studio v.2.1.3, EuroForMix v. 3.4.0, and STRmix™ v.2.7, to perform LR computations. N/A: Not Applicable. [a] Dropout is directly estimated through the peak height distribution. [b] Per locus specified by the software considering N = 0.

Parameters	Values		
	LRmix studio	EuroForMix	STRmix™
Co-ancestry Coefficient (F_{ST})	0.010	0.010	0.010
Drop-in frequency	0.05	0.05	0.05
Drop-in parameters' distribution	N/A	λ: 0.01	Uniform
Drop-in cap	N/A	N/A	100
Dropout	0.1	[a]	[a]
Minimum allele frequency	0.001	0.001	[b]
Threshold detection	N/A	100	100

In both EuroForMix and STRmix™, the allele frequencies were normalized. Regarding dropout, in LRmix Studio the pre-set value of 0.1 was used, while in both quantitative programs this parameter is directly estimated through the peak heights distribution [57,58]. A threshold detection equal to 100 RFU was considered in both quantitative software.

For the other parameters, specific to each software implementation, the default values provided by the authors were considered. Since STRmix™ software requires the inclusion of the stutters in the mixture sample files, they were also included in the EuroForMix analysis, both back and forward. LRmix Studio does not consider stutters and therefore its inclusion is not recommended. For LRmix Studio only, the selection of allelic peaks was decided by the same expert for all samples. In the quantitative software the stutter is automatically accommodated by the software, which introduce another source of variation between computed LR values. In the cases where a LR equal to zero was obtained for specific markers, the corresponding markers were removed from the analysis and the corresponding global LR value was denoted as $LR_{<21}$, as less than 21 markers were analyzed in the three tools.

LR values (log_{10} scale) for the same mixture/ single contributor sample pair and set of markers were obtained through EuroForMix, LRmix Studio, and STRmix™. Three pairwise differences were then computed for each pair mixture/single contributor sample. Statistical tests (chi-square test and test-t, respectively) were carried out to infer significant trends ($\alpha = 0.05$): (i) a specific software provided greater or smaller LR values than the others, and (ii) LR values obtained for mixtures with two estimated contributors showed to be higher/smaller than those with three.

3. Results and discussion

The global LR values obtained for each of the 156 mixture (with two or three estimated contributors)/ single contributor sample pairs analyzed were computed using the three software – LRmix Studio v.2.1.3 [34], EuroForMix v.3.4.0 [35], and STRmix™ v.2.7 [36]. Values obtained are presented in Fig. 1 and in Tables S1, S2.

STRmix™ returned a LR equal to zero for some mixture/ single contributor sample pairs, namely, in 17% and 9% of the cases with mixtures with two and three estimated contributors, respectively. In these cases, either alleles present in the single contributor sample were absent from the mixture sample, or MCMC sampling failed to identify the true genotype combination, i.e., MCMC algorithms failed to converge, as similarly observed in other studies [44,51,59]. Riman et. al found that this event mainly occurred when minor contributors were compared to mixture profiles with low templates [59]. Nevertheless, in this work, a LR equal to zero was observed indiscriminately for sample pairs whose single contributor profile considered could not be excluded from being either the major or minor contributor of the mixture.

Some approaches have been presented to circumvent this question, which include (i) ignore that locus during deconvolution, (ii) repeat the deconvolution with a random starting seed for the MCMC different than the one that resulted in a LR equal to zero, or (iii) repeat the deconvolution with an increase in number of MCMC parameters values [59,60]. In this case, for simplicity and to maintain the same conditions throughout STRmix™ computations (i.e., MCMC parameters values) these loci were ignored. Inevitably, to guarantee equal circumstances between all tools, these loci were also ignored in both EuroForMix and LRmix Studio software. Thus, as previously mentioned, these pairs were re-analyzed in the three programs, disregarding the marker(s) for which a LR equal to zero was achieved in STRmix™, and so, less than 21

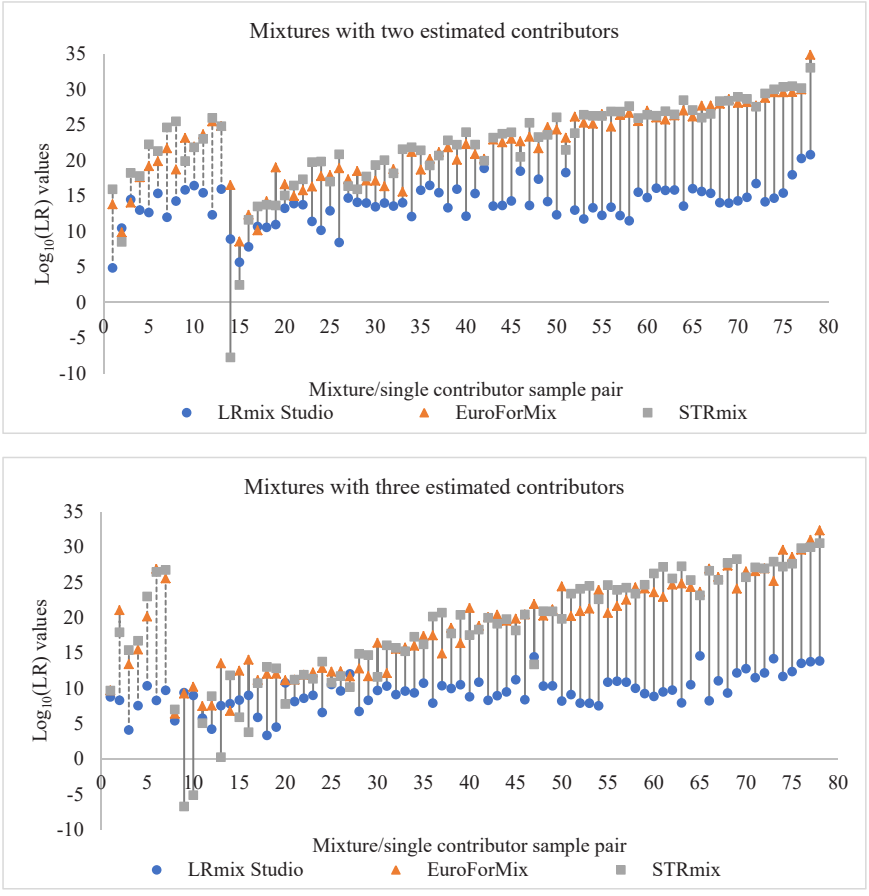


Fig. 1. LR values (log_{10} scale) computed for the same mixture/ single contributor sample pairs, with two (upper plot) and three estimated contributors (lower plot) in LRmix Studio v.2.1.3 (circle), EuroForMix v.3.4.0 (triangle) and STRmix™ v.2.7 (square). In the first 13 and 7 mixture/single contributor sample pairs illustrated in the upper and lower plot, respectively, less than 21 markers were analyzed (dashed line). Within the pairs highlighted with a dashed line in the upper plot, only 18 markers were considered in the first pair and 20 markers in the other 12 pairs. On the other hand, in the lower plot, in the first two pairs highlighted with a dashed line 19 markers were considered, and 20 markers in the remaining pairs highlighted with a dashed line.

markers (minimum 18) were considered per case. These specific situations are highlighted with a dashed line in Fig. 1.

STRmix™ LR values showed to be higher than those from EuroForMix for cases with mixtures with two estimated contributors (p = 2.35E-02), the same trend not being supported for the cases with mixtures with three estimated contributors (p = 8.21E-01). Also, LR values computed by STRmix™ and EuroForMix showed to be higher than those computed by LRmix Studio, for both cases with mixtures with two (p = 3.57E-16, for both software) and three estimated contributors (p = 1.09E-11 and p = 3.57E-16, resp.).

There are however three cases (2_2, 9_3, and 27_3) for which the LR value computed by LRmix Studio showed to be the highest. For the case 2_2, LR values retrieved by LRmix Studio, EuroForMix, and STRmix™ equated 3.04E+10, 7.43E+09, and 3.47E+08, respectively. For 9_3, LR values computed were 2.45E+09, 1.77E+09, and 1.89E-07, resp. And, finally, for case 27_3, LR values of 1.13E+12, 4.77E+11, and 1.51E+10, resp., were observed. As expected, these exceptions corresponded to mixture/single contributor sample pairs whose single contributor profile could not be excluded from being the minor contributor of the mixture under analysis. Since both EuroForMix and STRmix™ consider the quantitative information related to the height of the peaks, the computed LR values were lower than those computed by LRmix Studio, which does not consider that information.

In concordance with results previously reported [50], LR values obtained for the cases of mixtures with two estimated contributors showed to be higher than those obtained for cases with mixtures with three estimated contributors, for the three software: LRmix Studio (p = 1.00E-20), EuroForMix (p = 1.19E-03) and STRmix™ (p = 1.45E-03). The computation of lower LR values for cases with mixtures with three estimated contributors may be a consequence of the higher complexity associated, which may also be reflected in a greater dispersion of data. In both quantitative software, this trend was verified and corroborated by the standard deviation values computed for the distribution of LRs obtained by each software. The standard deviation values achieved for cases with mixtures with two and three estimated contributors were, respectively – in EuroForMix, 5.311 and 6.428, and in STRmix™, 6.397 and 8.077. On the other hand, for the qualitative software (LRmix Studio) the standard deviation values obtained, for cases with mixtures with two and three contributors showed to be 2.772 and 2.321, respectively, which do not support the trend obtained for quantitative software. The fact that qualitative software uses less information when compared to quantitative ones, may lower its sensitivity to more complex situations.

The observed discrepancies between the LR values (log_{10} scale) computed by the different software for the same mixture (with two and three estimated contributors)/single contributor sample pairs are summarized in Table 2. After calculating the difference between computed LR values (represented by D in Table 2), the number of cases with differences within the same range (zero and two units (log_{10} scale), two and four, and so on) was accounted and divided by N (total number of cases) to calculate the percentage presented. An example clarification regarding the way D was computed is presented made in Table S3.

As expected, the greatest differences were found when the results of the quantitative tool LRmix Studio were compared with those of the quantitative ones: EuroForMix and STRmix™. Indeed, it was noticed that the category with the highest frequency concerns the cases showing differences greater than ten units in a log_{10} scale – 37% and 46% in cases with mixtures with two, and 46% and 54% in cases with mixtures with three estimated contributors, respectively. Within the cases of mixtures with two estimated contributors, the greatest differences found between LRmix Studio and EuroForMix (LR = 3.20E+11 vs. LR = 5.14E+26) was smaller than the one found between LRmix Studio and STRmix™ (LR = 8.61E+08 vs. LR = 1.79E-08), the latter also supporting contrasting hypotheses. Similarly, for the cases of mixtures with three estimated contributors, the largest discrepancy observed between LRmix Studio and EuroForMix (LR = 1.73E+08 vs. LR=8.83E+26) was also smaller than the one observed between LRmix Studio and STRmix™ (LR = 8.59E+07 vs. LR = 1.88E+27).

In almost all the cases, LR values computed by the quantitative software (EuroForMix and STRmix™) were consistently higher than those computed by qualitative (LRmix Studio). Coincidentally, the single-source profile used as a reference in these cases could not be excluded from being the major contributor of the mixture. The non-consideration of the peak heights information by qualitative tools makes them less sensitive to several aspects, such as contributors in different proportions, which may result in these differences.

A different scenario was observed when comparing LR values obtained through the quantitative tools STRmix™ and EuroForMix. In both cases with mixtures with two and three estimated contributors, most of the cases (73% and 59%, respectively) did not differ by more than two units on a log_{10} scale. Regarding cases with mixtures with two estimated contributors, only one case showed a difference greater than ten units in a log_{10} scale between EuroForMix (3.48E+16) and STRmix™ (1.79E-08). It is noteworthy that this was the same sample pair for which LRmix Studio and STRmix™ showed the greatest divergence of results. For the cases with mixtures with three estimated contributors, 5% of the observed LR differences showed to be greater than ten units on a log_{10} scale. The maximum difference of 15.970 units (log_{10} scale) was observed in a case for which LR values retrieved by EuroForMix and STRmix™ equated 1.77E+ 09 and 1.89E-07, respectively.

For simplicity sake and practical reasons, only cases with differences higher than ten units in a log_{10} scale between quantitative software were further analyzed, namely, pairs 14_2, 9_3, 10_3, 13_3 and 16_3 – Fig. S1, Tables S4 and S5. None of these pairs contain alleles that have not been

Table 2
Pairwise differences between the LR values (log_{10} scale) computed by LRmix Studio v.2.1.3, EuroForMix v.3.4.0, and STRmix™ v.2.7, for cases with mixtures with two and three estimated contributors, and maximum, mean, and median values of these differences.

D = ABS [Log_{10} (LR1) – Log_{10} (LR2)]	Estimated number of contributors					
	Two			Three		
	EuroForMix and LRmix Studio	STRmix™ and LRmix Studio	EuroForMix and STRmix™	EuroForMix and LRmix Studio	STRmix™ and LRmix Studio	EuroForMix and STRmix™
	N = 78	N = 78	N = 78	N = 78	N = 78	N = 78
0 < D < 2	8%	6%	73%	13%	9%	59%
2 < D < 4	13%	15%	19%	8%	8%	24%
4 < D < 6	19%	13%	4%	8%	9%	9%
6 < D < 8	5%	6%	3%	14%	10%	1%
8 < D < 10	18%	13%	0%	12%	10%	1%
D > 10	37%	46%	1%	46%	54%	5%
Maximum	15.206	16.682	24.288	18.707	19.341	15.970
Mean	8.097	8.823	1.839	9.425	10.252	2.468
Median	8.811	9.687	1.145	9.690	10.445	1.333

previously observed in the used allele frequency database, and therefore differences on the minimum frequencies values cannot justify these discrepancies, as previously shown in other study [51]. For these pairs, the great discrepancies observed in the global LR values resulted from the accumulation of small differences in the analysis of each marker within the specific cases. At this point, it is noteworthy that the differences found and the proportion of cases in these conditions were not greater because, within the same analysis, there were markers where the LR was greater when computed with STRmix™ than when EuroForMix was used, occurring the opposite situation for others. This resulted in smoother global LR differences than those that may have occurred if one of the two software consistently attributed lower (or higher) LR values than the other. Anyway, most of cases with differences greater than ten units (log_{10} scale) resulted from cases where EuroForMix computed LR values higher than STRmix in most markers.

An extended analysis of the greatest differences found between quantitative software was performed (LRs differing in more than log_{10} scale units), for both cases with mixtures with two and three estimated contributors.

Starting with the only case of mixture with two estimated contributors (pair 14_2), a difference of 24.288 units (log_{10} scale) was observed between the two quantitative programs: LR(EuroForMix) = 3.48E+ 16 and LR(STRmix™) = 1.79E-08. For this pair, the single contributor profile was considered the minor contributor of the mixture by both EuroForMix and STRmix™ with a mixture proportion of 14.28% and 25.91%, respectively. However, in a per-locus LR analysis, 19 out of the 21 markers favored inclusion (LR>1) in EuroForMix, while in STRmix™ only 9 showed the same trend, as described in Table S4. In Table S6, the profiles of the loci showing the greatest differences, namely, D10S1248 (2.947 units, log_{10} scale), D8S1179 (2.946), D22S1045 (2.900), and D16S539 (2.068) are described. Neither of the statistic diagnoses given by both software supported a good performance for this analysis. The peak height variability parameter in EuroForMix adjusted to 0.476, a very high value – meaning a poor describing model. A more tolerant peak variance parameter in this software allows a lower false exclusion rate and a higher rate of false support of non-contributors [51]. On the other hand, the STRmix™ equivalent parameter (peak height variance parameter) shows slightly larger values than the mode of the respective prior distribution, which may indicate stochastic effects or incorrect estimation of the number of contributors [60]. Also, the remaining STRmix™ secondary diagnostics – log(likelihood) and Gelman-Rubin – reinforce the software inability to describe the observed data. A negative log(likelihood) of − 253.95 was observed, which may be due to several reasons such as, profile used as a reference being very low level [60]. In Gelman-Rubin parameter, the acronym NaN (Not a Number) is presented. So, even though the diagnostics of both software pointed out the model bad fitting, the result obtained was very distinct, probably due to different modeling assumptions and parameters settings. STRmix™ models locus specific amplification efficiencies and consider expected stutter ratios defined using empirical data. On the contrary, EuroForMix does not model locus specific amplification efficiencies and has a blanket expected stutter rate. As EuroForMix does not allow a differential amplification between loci, the need for a more tolerant peak height variance parameter is understandable [51].

Within the cases with mixtures with three estimated contributors, the greatest difference occurred in a mixture/single contributor pair (9_3) with a difference of 15.970 units (log_{10} scale) between EuroForMix (LR = 1.77E+09) and STRmix™ (LR = 1.89E-07). Like the per-marker LR previously analyzed, most markers in EuroForMix (17 out of 21) supported inclusion, while in STRmix™ only 8 out of the 21 showed the same trend – see Table S4. The STR profiles of markers D8S1179, D13S317, and D19S433 are described in Table S7 since the greatest differences were found in those loci – 2.547, 2.092, and 2.087 log_{10} scale units, respectively. For this case, both EuroForMix and STRmix™ considered the single contributor profile the second contributor of the mixture with a mixture proportion of 15.67% and 22.92%, respectively. Evaluating the statistic diagnostics of EuroForMix and STRmix™, both tools supported a good performance for this analysis. EuroForMix peak height variability parameter adjusted to a value of 0.195, revealing a good describing model. STRmix™ secondary diagnostics also revealed a good fit of the model – a log(likelihood) of 28.58 and a Gelman-Rubin value of 1.06 indicating that STRmix™ was able to well describe the data and the model converged [60]. Contrarily to the situation previously analyzed considering two estimated contributors, the diagnostics of both software reflected the model good fitting in this case. The very discrepant results observed are probably due to different modeling assumptions and parameters settings, as stated before.

A more in-depth investigation on these cases where greater differences were reported by different informatics approaches are being performed and soon presented in a dedicated research article.

4. Conclusions

Informatics tools for the analysis and interpretation of forensic samples DNA results, perform a quantification of the genetic evidence through the computation of a probative value using electropherograms' information.

Software based on quantitative models take into account a larger amount of data than the qualitative ones, as these only consider the detected alleles, disregarding the corresponding peak heights. So, while qualitative tools only include statistical models for probability of drop-out and drop-in, quantitative tools model also the peak height information (i.e., peak alleles, peak balance, and stutter peaks). Indeed, the major distinctive feature of existing software is the ability to consider the (inferred) DNA quantity (after amplification).

As expected, the greatest differences between LRs were generally found when results retrieved from qualitative and quantitative tools were compared. Notwithstanding, and despite the smaller differences, quantitative software also retrieved different LR results reflecting different mathematical modeling of various parameters. Indeed, even using similar parameter values, each software assigns a different statistical modeling to each parameter leading to the computation of different LR values. Nevertheless, some of the differences observed were unexpected and intriguing large. This prompted us to develop a deeper research to explore the characteristics of the samples that lead to such differences, which results will be presented in a dedicated research article.

The uniqueness, and inability to predict and replicate real casework mixture samples, were the main reasons for the selection of this kind of samples. The drawback is that its composition (donors and proportion) is unknown, contrary to what occurs for mock samples (noting that accidental contaminations may also occur in these cases). Thus, the inference on which software produced more accurate results couldn't be performed. We further anticipate that some differences may be enhanced if the (some of the) contributors are closely related (as is often the case in sexual violence) and or related to the POI.

Software based on probabilistic genotyping methods is essential in the interpretation and quantification of complex DNA mixtures. Different software solutions encode distinct mathematical, statistical and informatics models which lead to the return of different values for the same input samples and approximated parameters. In what seems a paradox, this software dependence for the quantification of the evidence implies a need of improvement of the education and training of the experts, when compared to a previous situation where only a binary exclusion/inclusion interpretation (without statistical calculations) was developed.

Our aim with this short communication is to show how different LR computations can be obtained when considering real casework samples, regardless of whether the informatics tools consider the quantitative information of the electropherogram or not. Quantitative software used (EuroForMix and STRmix™) have built-in diagnostic tools that allow the expert to know if the models are good fits. Nevertheless, for some cases,

despite statistic diagnostics of both tools supported a good fit, discrepant LR results were computed. So, since in the forensic routine the genetic evidence is not usually quantified resorting to more than one informatics tool, there are no confrontations between possibly different results, practitioners in the field assuming the (only) calculated value as steel plated. Thus, we intend to warn experts not to blindly trust the program they use and to be critical of the computed result, paving the way for the routine consideration of more than one tool in routine forensic casuistic. If for a specific case discrepant results are obtained, forensic specialists should investigate the samples at stake, as well as the mathematical modeling corresponding, to further act in conformity and reach a solid, informed, decision.

Indeed, forensic experts play a crucial role in the profile interpretation, from the electropherogram analysis (decisions about the presence of non-allelic peaks and estimation of the number of contributors, for example) to the proper use of the software for LR computations. Furthermore, and considering the different approaches, the expert knowledge should be extended to the understanding of how each software operates and its respective encoded models. Only this way he/she will be able to sustain the conclusions in court or any other scrutiny scenario.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowlegments

This work was partially financed by FEDER- Fundo Europeu de Desenvolvimento Regional funds through the COMPETE 2020 - Operacional Program for Competitiveness and Internationalization (POCI), Portugal 2020, and by Portuguese funds through FCT- Fundação para a Ciência e a Tecnologia/Ministério da Ciência, Tecnologia e Inovação in the framework of the projects "Institute for Research and Innovation in Health Sciences" (POCI-01–0145-FEDER-007274). NP is supported by FCT, under the program contract provided in Decree-Law no.57/2016 of August 29. CC is funded by a FCT doctoral grant 2021.05655. BD.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2022.102715.

References

[1] C.C. Benschop, C.P. van der Beek, H.C. Meiland, A.G. 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 (4) (2011) 316–328.
[2] H. Haned, L. Pène, F. Sauvage, D. Pontier, The predictive value of the maximum likelihood estimator of the number of contributors to a DNA mixture, Forensic Sci. Int. Genet. 5 (4) (2011) 281–284.
[3] D.R. Paoletti, D.E. Krane, T.E. Doom, M. Raymer, Inferring the number of contributors to mixed DNA profiles, IEEE/ACM Trans. Comput. Biol. Bioinform. 9 (1) (2011) 113–122.
[4] J. Perez, A.A. Mitchell, N. Ducasse, J. Tamariz, T. Caragine, Estimating the number of contributors to two-, three-, and four-person mixtures containing DNA in high template and low template amounts, Croat. Med. J. 52 (3) (2011) 314–326.
[5] A. Biedermann, S. Bozza, K. Konis, F. Taroni, Inference about the number of contributors to a DNA mixture: comparative analyses of a Bayesian network approach and the maximum allele count method, Forensic Sci. Int. Genet. 6 (6) (2012) 689–696.
[6] C. Benschop, H. Haned, T. Sijen, Consensus and pool profiles to assist in the analysis and interpretation of complex low template DNA mixtures, Int. J. Leg. Med. 127 (1) (2013) 11–23.
[7] C.A. Rocheleau, Organ donation intentions and behaviors: application and extension of the theory of planned behavior, J. Appl. Soc. Psychol. 43 (1) (2013) 201–213.
[8] T. Tvedebrink, On the exact distribution of the numbers of alleles in DNA mixtures, Forensic Sci. Int. Genet. Suppl. Ser. 4 (1) (2013) e278–e279.
[9] C.C. Benschop, H. Haned, L. Jeurissen, P.D. Gill, T. Sijen, The effect of varying the number of contributors on likelihood ratios for complex DNA mixtures, Forensic Sci. Int. Genet. 19 (2015) 92–99.
[10] H.M. Heyman, F. Senejoux, I. Seibert, T. Klimkait, V.J. Maharaj, J.J.M. Meyer, Identification of anti-HIV active dicaffeoylquinic-and tricaffeoylquinic acids in Helichrysum populifolium by NMR-based metabolomic guided fractionation, Fitoterapia 103 (2015) 155–164.
[11] H. Swaminathan, C.M. Grgicak, M. Medard, D.S. Lun, NOCIt: a computational method to infer the number of contributors to DNA samples analyzed by STR genotyping, Forensic Sci. Int. Genet. 16 (2015) 172–180.
[12] L.E. Alfonse, G. Tejada, H. Swaminathan, D.S. Lun, C.M. Grgicak, Inferring the number of contributors to complex DNA mixtures using three methods: exploring the limits of low-template DNA interpretation, J. Forensic Sci. 62 (2) (2017) 308–316.
[13] M.A. Marciano, J.D. Adelman, PACE: probabilistic assessment for contributor estimation – a machine learning-based assessment of the number of contributors in DNA mixtures, Forensic Science International: Genetics 27 (2017) 82–91.
[14] C. Benschop, A. Backx, T. Sijen, Automated estimation of the number of contributors in autosomal STR profiles, Forensic Sci. Int. Genet. Suppl. Ser. 7 (1) (2019) 7–8.
[15] C.C. Benschop, J. van der Linden, J. Hoogenboom, R. Ypma, H. Haned, Automated estimation of the number of contributors in autosomal short tandem repeat profiles using a machine learning approach, Forensic Sci. Int. Genet. 43 (2019), 102150.
[16] C.C. Benschop, J. Hoogenboom, F. Bargeman, P. Hovers, M. Slagter, J. van der Linden, R. Parag, D. Kruise, K. Drobnic, G. Klucevsek, Multi-laboratory validation of DNAxs including the statistical library DNAStatistX, Forensic Sci. Int. Genet. 49 (2020), 102390.
[17] T. Clayton, J. Whitaker, R. Sparkes, P. Gill, Analysis and interpretation of mixed forensic stains using DNA STR profiling, Forensic Sci. Int. 91 (1) (1998) 55–70.
[18] Methods SWGoDA: SWGDAM interpretation guidelines for autosomal STR typing by forensic DNA testing laboratories. 2010.
[19] J.M. Butler, Advanced Topics in Forensic DNA Typing: Interpretation, Academic Press, 2014.
[20] P. Gill, Forensic Practitioner's Guide to the Interpretation of Complex Dna Profiles (is), Academic Press, 2020.
[21] P.S. Walsh, N.J. Fildes, R. Reynolds, Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA, Nucleic Acids Res. 24 (14) (1996) 2807–2812.
[22] P. Gill, Application of low copy number DNA profiling, Croat. Med. J. 42 (3) (2001) 229–232.
[23] P.M. Schneider, K. Bender, W.R. Mayr, W. Parson, B. Hoste, R. Decorte, J. Cordonnier, D. Vanek, N. Morling, M. Karjalainen, et al., STR analysis of artificially degraded DNA-results of a collaborative European exercise, Forensic Sci. Int. 139 (2–3) (2004) 123–134.
[24] 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, et al., DNA commission of the international society of forensic genetics: recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2–3) (2006) 90–101.
[25] M. Fondevila, C. Phillips, N. Naverán, M. Cerezo, A. Rodríguez, R. Calvo, L. M. Fernández, Á. Carracedo, M.V. Lareu, D.N.A. Challenging, Assessment of a range of genotyping approaches for highly degraded forensic samples, Forensic Sci. Int. Genet. Suppl. Ser. 1 (1) (2008) 26–28.
[26] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Sci. Int Genet. 4 (1) (2009) 1–10.
[27] A.J. Gibb, A.L. Huell, M.C. Simmons, R.M. Brown, Characterisation of forward stutter in the AmpFlSTR SGM Plus PCR, Sci. Justice 49 (1) (2009) 24–31.
[28] A.A. Westen, J.H. Nagel, C.C. Benschop, N.E. Weiler, B.J. de Jong, T. Sijen, Higher capillary electrophoresis injection settings as an efficient approach to increase the sensitivity of STR typing, J. Forensic Sci. 54 (3) (2009) 591–598.
[29] A. Freire-Aradas, M. Fondevila, A.K. Kriegel, C. Phillips, P. Gill, L. Prieto, P. M. Schneider, A. Carracedo, M.V. Lareu, A new SNP assay for identification of highly degraded human DNA, Forensic Sci. Int. Genet. 6 (3) (2012) 341–349.
[30] S. Gittelson, A. Biedermann, S. Bozza, F. Taroni, Decision analysis for the genotype designation in low-template-DNA profiles, Forensic Sci. Int. Genet. 9 (2014) 118–133.
[31] C.D. Steele, M. Greenhalgh, D.J. Balding, Evaluation of low-template DNA profiles using peak heights, Stat. Appl. Genet. Mol. Biol. 15 (5) (2016) 431–445.
[32] H.R. Dash, P. Shrivastava, S. Das, Analysis of capillary electrophoresis results by geneMapper® ID-X v 1.5 software. Principles and Practices of DNA Analysis: A Laboratory Manual for Forensic DNA Typing, Springer,, 2020, pp. 213–237.
[33] 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 Bioinform. 16 (2015) 298.
[34] H. Haned, K. Slooten, P. Gill, Exploratory data analysis for the interpretation of low template DNA mixtures, Forensic Sci. Int. Genet. 6 (6) (2012) 762–774.
[35] O. Bleka, G. Storvik, P. Gill, EuroForMix: an open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts, Forensic Sci. Int. Genet. 21 (2016) 35–44.
[36] D. Taylor, J.A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int. Genet. 7 (5) (2013) 516–528.
[37] M.D. Coble, J.A. Bright, Probabilistic genotyping software: an overview, Forensic Sci. Int. Genet. 38 (2019) 219–224.
[38] T.W. Bille, S.M. Weitz, M.D. Coble, J. Buckleton, J.A. Bright, Comparison of the performance of different models for the interpretation of low level mixed DNA profiles, Electrophoresis 35 (21–22) (2014) 3125–3133.

[39] O. 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.
[40] E. Alladio, M. Omedei, S. Cisana, G. D'Amico, D. Caneparo, M. Vincenti, P. Garofano, DNA mixtures interpretation – a proof-of-concept multi-software comparison highlighting different probabilistic methods' performances on challenging samples, Forensic Sci. Int. Genet. 37 (2018) 143–150.
[41] J.S. Buckleton, J.A. Bright, K. Cheng, B. Budowle, M.D. Coble, NIST interlaboratory studies involving DNA mixtures (MIX13): a modern analysis, Forensic Sci. Int. Genet. 37 (2018) 172–179.
[42] P.A. Barrio, M. Crespillo, J.A. Luque, M. Aler, C. Baeza-Richer, L. Baldassarri, E. Carnevali, P. Coufalova, I. Flores, O. Garcia, et al., GHEP-ISFG collaborative exercise on mixture profiles (GHEP-MIX06). Reporting conclusions: results and evaluation, Forensic Sci. Int. Genet. 35 (2018) 156–163.
[43] L. Prieto, H. Haned, A. Mosquera, M. Crespillo, M. Aleman, M. Aler, F. Alvarez, C. Baeza-Richer, A. Dominguez, C. Doutremepuich, et al., Euroforgen-NoE collaborative exercise on LRmix to demonstrate standardization of the interpretation of complex DNA profiles, Forensic Sci. Int. Genet. 9 (2014) 47–54.
[44] T.R. Moretti, R.S. Just, S.C. Kehl, L.E. Willis, J.S. Buckleton, J.A. Bright, D. A. 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.
[45] J.A. Bright, R. Richards, M. Kruijver, H. Kelly, C. McGovern, A. Magee, A. McWhorter, A. Ciecko, B. Peck, C. Baumgartner, et al., Internal validation of STRmix – a multi laboratory response to PCAST, Forensic Sci. Int. Genet. 34 (2018) 11–24.
[46] J.A. Bright, K. Cheng, Z. Kerr, C. McGovern, H. Kelly, T.R. Moretti, M.A. Smith, F. R. Bieber, B. Budowle, M.D. Coble, et al., STRmix collaborative exercise on DNA mixture interpretation, Forensic Sci. Int. Genet. 40 (2019) 1–8.
[47] M. Crespillo, P.A. Barrio, J.A. Luque, C. Alves, M. Aler, F. Alessandrini, L. Andrade, R.M. Barretto, A. Bofarull, S. Costa, et al., GHEP-ISFG collaborative exercise on mixture profiles of autosomal STRs (GHEP-MIX01, GHEP-MIX02 and GHEP-MIX03): results and evaluation, Forensic Sci. Int. Genet. 10 (2014) 64–72.
[48] C.C.G. Benschop, E. Connolly, R. Ansell, B. Kokshoorn, Results of an inter and intra laboratory exercise on the assessment of complex autosomal DNA profiles, Sci. Justice 57 (1) (2017) 21–27.
[49] J.M. Butler, M.C. Kline, M.D. Coble, NIST interlaboratory studies involving DNA mixtures (MIX05 and MIX13): variation observed and lessons learned, Forensic Sci. Int. Genet. 37 (2018) 81–94.
[50] P. Gill, C. Benschop, J. Buckleton, Ø. Bleka, D. Taylor, A review of probabilistic genotyping systems: euroForMix, DNAStatistX and STRmixTM, Genes 12 (10) (2021) 1559.
[51] K. Cheng, Ø. Bleka, P. Gill, J. Curran, J.A. Bright, D. Taylor, J. Buckleton, A comparison of likelihood ratios obtained from EuroForMix and STRmixTM, J. Forensic Sci. 66 (6) (2021) 2138–2155.
[52] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria (2020) URL https://www.R-project.org/.
[53] C.R. Hill, D.L. Duewer, M.C. Kline, M.D. Coble, J.M. Butler, US population data for 29 autosomal STR loci, Forensic Sci. Int. Genet. 7 (3) (2013) e82–e83.
[54] H. Haned, C.C. Benschop, P.D. Gill, T. Sijen, Complex DNA mixture analysis in a forensic context: evaluating the probative value using a likelihood ratio model, Forensic Sci. Int. Genet. 16 (2015) 17–25.
[55] J. Buckleton, J. Curran, J. Goudet, D. Taylor, A. Thiery, B. Weir, F.S.T. Population-specific, values for forensic STR markers: a worldwide survey, Forensic Sci. Int. Genet. 23 (2016) 91–100.
[56] B.S. Weir, The second national research council report on forensic DNA evidence, Am. J. Hum. Genet. 59 (3) (1996) 497.
[57] J. Buckleton, H. Kelly, J.-A. Bright, D. Taylor, T. Tvedebrink, J.M. Curran, Utilising allelic dropout probabilities estimated by logistic regression in casework, Forensic Sci. Int. Genet. 9 (2014) 9–11.
[58] T. Tvedebrink, P.S. Eriksen, H.S. Mogensen, N. Morling, Estimating the probability of allelic drop-out of STR alleles in forensic genetics, Forensic Sci. Int. Genet. 3 (4) (2009) 222–226.
[59] S. Riman, H. Iyer, P.M. Vallone, Examining performance and likelihood ratios for two likelihood ratio systems using the PROVEDIt dataset, PLOS One 16 (9) (2021), e0256714.
[60] L. Russell, S. Cooper, R. Wivell, Z. Kerr, D. Taylor, J. Buckleton, J.A. Bright, A guide to results and diagnostics within a STRmixTM report, WIREs Forensic Sci. 1 (2019) 6.