Science for Communities

Article:

Bright, J.-A., Cheng, K., Kerr, Z., Mcgovern, C., Kelly, H., Moretti, T. R., Smith, M. A., Bieber, F. R., Budowle, B., Coble, M. D., Alghafri, R., Allen, P. S., Barber, A., Beamer, V., Buettner, C., Russell, M., Gehrig, C., Hicks, T., Charak, J., Cheong-Wing, K., Ciecko, A., Davis, C. T., Donley, M., Pederson, N., Gartside, B., Granger, D., Greer-Ritzheimer, M., Reisinger, E., Kennedy, J., Grammer, E., Kaplan, M., Hansen, D. Larsen, H. J., Laureano, A., Li, C., Lien, E., Lindberg, E., Kelly, C., Mallinder, B., Malsom, S., Yacovone-Margetts, A., McWhorter, A., Prajapati, S. M., Powell, T., Shutler, G., Stevenson, K. Stonehouse, A. R., Smith, L., Murakami, J., Halsing, E., Wright, D. Clark, L., Taylor, D. A., Buckleton, J. (2019). **STRmix™ collaborative exercise on DNA mixture interpretation**. *Forensic Science International: Genetics*, *40*, 1–8.

INSTITUTE OF ENVIRONMENTAL SCIENCE AND RESEARCH LIMITED
Kenepuru Science Centre: 34 Kenepuru Drive, Kenepuru, Porirua 5022 | PO Box 50348, Porirua 5240, New Zealand
T: +64 4 914 0700 **F:** +64 4 914 0770
www.esr.cri.nz

STRmix™ collaborative exercise on DNA mixture interpretation

Jo-Anne Bright[1]*, Kevin Cheng[1], Zane Kerr[2], Catherine McGovern[1], Hannah Kelly[1], Tamyra R. Moretti[3], Michael A. Smith[3], Frederick R. Bieber[4], Bruce Budowle[5] , Michael D. Coble[5], Rashed Alghafri[6], Paul Stafford Allen[7], Amy Barber[8], Vickie Beamer[9], Christina Buettner[10], Melanie Russell[11], Christian Gehrig[12], Tacha Hicks[13], Jessica Charak[14], Kate Cheong-Wing[15], Anne Ciecko[16], Christie T. Davis[18], Michael Donley[19], Natalie Pedersen[20], Bill Gartside[21], Dominic Granger[22], MaryMargaret Greer-Ritzheimer[23], Erick Reisinger[24], Jarrah Kennedy[25], Erin Grammer[26], Marla Kaplan[27], David Hansen[28], Hans J. Larsen[29], Alanna Laureano[30], Christina Li[31], Eugene Lien[32], Emilia Lindberg[33], Ciara Kelly[34], Ben Mallinder[35], Simon Malsom[36], Alyse Yacovone-Margetts[37], Andrew McWhorter[38], Sapana M. Prajapati[39], Tamar Powell[40], Gary Shutler[41], Kate Stevenson[1], April R. Stonehouse[42], Lindsey Smith[43], Julie Murakami[44], Eric Halsing[45], Darren Wright[46], Leigh Clark[47], Duncan A. Taylor[48,49], John Buckleton[1,50]

[1] Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland, 1142 New Zealand
[2] STRmix™ Limited, Auckland, 1142 New Zealand
[3] DNA Support Unit, Federal Bureau of Investigation Laboratory, 2501 Investigation Parkway, Quantico, VA 22135, USA
[4] Center for Advanced Molecular Diagnostics, Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, 75 Francis Street, Boston, MA 02115 USA email: fbieber@bwh.harvard.edu
[5] Center for Human Identification, Department of Microbiology, Immunology, and Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, Texas 76107, USA email: Bruce.Budowle@unthsc.edu
[6] General Department of Forensic Sciences and Criminology, Dubai Police G.H.Q., Dubai, UAE.
[7] Cellmark Forensic Services, UK. email: pstaffordallen@cellmark.co.uk
[8] Massachusetts State Police Crime Laboratory
[9] Scottsdale Police Department Crime Laboratory
[10] Wyoming State Crime Laboratory
[11] CT DESPP Division of Scientific Services

[12] University Center of Legal Medicine, Lausanne - Geneva (CURML)
[13] School of Criminal Justice, University of Lausanne
[14] Las Vegas Metropolitan Police Department
[15] Northern Territory Police, Fire and Emergency Services
[16] Midwest Regional Forensic Laboratory
[17] Onondaga County Center for Forensic Sciences
[18] Helix Analytical, Inc
[19] Harris County Institute of Forensic Sciences
[20] Victoria Police Forensic Science laboratory
[21] San Bernardino County Sheriff's Dept.
[22] Laboratoire de sciences judiciaires et de médecine légale, Montréal
[23] DuPage County Sheriff's Crime Laboratory, Illinois
[24] Orange County Crime Laboratory
[25] Kansas City Police Crime Laboratory
[26] Indiana State Police Laboratory
[27] Oregon State Police Portland Metro Crime Laboratory
[28] Forensic Analytical Crime laboratory
[29] University of Copenhagen
[30] Westchester County Department of Labs and Research
[31] Government Laboratory (Hong Kong)
[32] New York City Office of Chief Medical Examiner (OCME)
[33] NBI, Forensic laboratory Finland
[34] Forensic Science Ireland
[35] Scottish Police Authority
[36] Key Forensic Services
[37] Palm Beach County Sheriff's Office
[38] Texas Department of Public Safety
[39] Signature Science, LLC Austin, Texas
[40] San Francisco Police Department Crime Lab
[41] Washington State Patrol
[42] Mesa Police Department--Forensic Services
[43] US Army Criminal Investigation Laboratory
[44] PathWest
[45] California department of Justice
[46] Idaho State Police Forensic Services
[47] Florida Department of Law Enforcement
[48] Forensic Science SA, 21 Divett Place, Adelaide, SA 5000, Australia
[49] School of Biological Sciences, Flinders University, GPO Box 2100 Adelaide SA, Australia 5001
[50] University of Auckland, Department of Statistics, Auckland, New Zealand

* Corresponding author at: Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland, 1142, New Zealand. Email address: Jo.bright@esr.cri.nz .

An intra and inter-laboratory study using the probabilistic genotyping (PG) software STRmix™ is reported. Two complex mixtures from the PROVEDIt set, analysed on an Applied Biosystems™ 3500 Series Genetic Analyzer, were selected. 174 participants responded.

For Sample 1 (low template, in the order of 200 rfu for major contributors) five participants described the comparison as inconclusive with respect to the POI or excluded him. Where *LR*s were assigned, the point estimates ranging from 2×10^4 to 8×10^6. For Sample 2 (in the order of 2000 rfu for major contributors), *LR*s ranged from 2×10^{28} to 2×10^{29}. Where *LR*s were calculated, the differences between participants can be attributed to (from largest to smallest impact):

- varying number of contributors (*NoC*),
- the exclusion of some loci within the interpretation,
- differences in local CE data analysis methods leading to variation in the peaks present and their heights in the input files used,
- and run-to-run variation due to the random sampling inherent to all MCMC-based methods.

This study demonstrates a high level of repeatability and reproducibility among the participants. For those results that differed from the mode, the differences in *LR* were almost always minor or conservative.

Highlights

- Inter-laboratory study with 174 participants using STRmix™
- CE analysis settings resulted in larger differences in *LR* than PG software
- Differences in log(*LR*) due to MCMC variation were less than one order of magnitude

Keywords: Forensic DNA interpretation; probabilistic genotyping; STRmix, Inter-laboratory study; Intra-laboratory study

Introduction

In forensic DNA analysis there is valid interest in the reliability of the interpretation reported. Reliability, in part but not fully, relates to the scientific concepts of repeatability, reproducibility, and accuracy. Before attempting to address repeatability and reproducibility, and the challenge in context of determining accuracy in context, we give a brief summary of the current process that leads to the reported result.

The predominant DNA analysis process takes a sample through a series of processing stages such as extraction, amplification, separation by capillary electrophoresis (CE), and detection of fluorescent tags after laser excitation to produce a raw output that is a trace of signal with migration time. Since several fluorescent dyes are used, multiple traces are produced simultaneously. The outputs are the raw CE data. These traces are then processed by software which applies baselining and other algorithms such as smoothing. Detection time is converted to base length by alignment with base pair size standards. Peaks are then designated as alleles by comparison of their base length with those of an allelic ladder. The output from this process is a DNA profile known as an electropherogram (or epg). After production of the epg, the data are interpreted. Decisions are made at this point regarding the suitability of interpretation, single source or mixture, removal of artifacts, number of contributors, and so on. The epg may be compared with one or more persons of interest (POI). In some instances, it is possible to assume the presence of the DNA of a contributor(s), say the victim in intimate samples, which facilitates the interpretation process.

Each of the stages of the DNA analysis process will introduce variability [1] which will impact repeatability and reproducibility. Repeatability is the variation between results generated under the same conditions of measurement. This assessment is made with the same operator on the same machine at closely similar times. Reproducibility is similar to repeatability except under changed conditions of measurement such as different operators and different instruments. The study herein describes an intra- and inter-laboratory study using the probabilistic genotyping (PG) software STRmix™ for the interpretation of the epgs to assess contributions of variability of the interpretation. The output from STRmix™ is a likelihood ratio, *LR*. Reproducibility in this regard could be considered to be the precision of the *LR* created from the same raw CE output by different people working independently in the same (intra) or in different (inter) laboratories.

Accuracy describes how close a measurement is to the correct or true answer. In casework we may never know if the POI is truly a contributor to the sample; only in controlled studies can we know whether or not a POI is a true donor. The *LR* is the product of a number of modelling assumptions and the correct answer exists only within the specifications of the model. In some situations the *LR* can be predicted from the models [2], but in most instances the reasonableness of the *LR* derives from our belief in the correctness of the model(s). Just as there is no true model, there is no true *LR* [3]. Indeed many forensic scientists deliberately bias the models, and their decision making, to understate the *LR*, a behaviour described as being conservative, that is, to the favour the defendant. Thus, accuracy is not attained routinely but instead a reasonable answer is sought that does not overstate the value of the comparison. Later in this paper reasonableness of an answer will be addressed based on the models, and in some cases, knowledge of the true donors.

There have been a number of inter-laboratory studies undertaken to assess variability, and some of these studies are summarised by Butler et al. [4]. A general concern about the outcome of these studies is limited reproducibility. This lack of reproducibility has fuelled calls for standards of operation and standardisation. Standards of operation and general

features that often are addressed by various working groups (usually more so as guidelines) and will not be described herein. Standardisation, performing in the same manner, cannot begin until the sources of variation are identified. While a degree of conservatism is desired, a search for standardisation should not drive reproducibility in which everyone generates an unreasonable result. To emphasise this point by exaggeration, everyone would obtain reproducibility if an *LR* of 1 was reported in every situation. However, an *LR* of 1 in every situation would not serve stakeholders, since it denies the prosecution and defense valuable information.

Forensic scientists have a marked and deliberate tendency to concede uncertainty in the direction of lowering the *LR*, the aforementioned conservativeness behaviour. In an inter-laboratory study this decision(s) translates into a long left-hand tail to the distribution of reported *LR*s, where some analysts have made concessions that others have not and thus contributing to variation in the final result(s). The ranges given in such studies are usually not the range of the laboratories' best (i.e. accurate) assignment but rather the ranges in *LRs* assigned by the laboratories to deliberately understate the evidence as a deliberate policy.

In routine casework peer (or technical) review of results before reporting is routine. Peer review may or may not be performed when participating in inter-laboratory studies. The effect of peer review is to ensure proper interpretation of the evidence (within a laboratory's guidelines), a reduction of the rate of trivial errors, such as transcription errors, and proper documentation, such as maintaining chain of custody to name a few. There has been a general hope that the advent of PG solutions will improve reproducibility in both the approach to interpretation and assigned *LR*s. PG solutions automate some, but not all, aspects of decision making and this was expected to reduce variability in *LR* both within and between laboratories. However, as mentioned above there are other aspects of the DNA typing process that contribute to variation which PG software cannot control and stakeholders should discern the sources of variation due to the analytical process, those due to the performance of PG software, and those due to the operator.

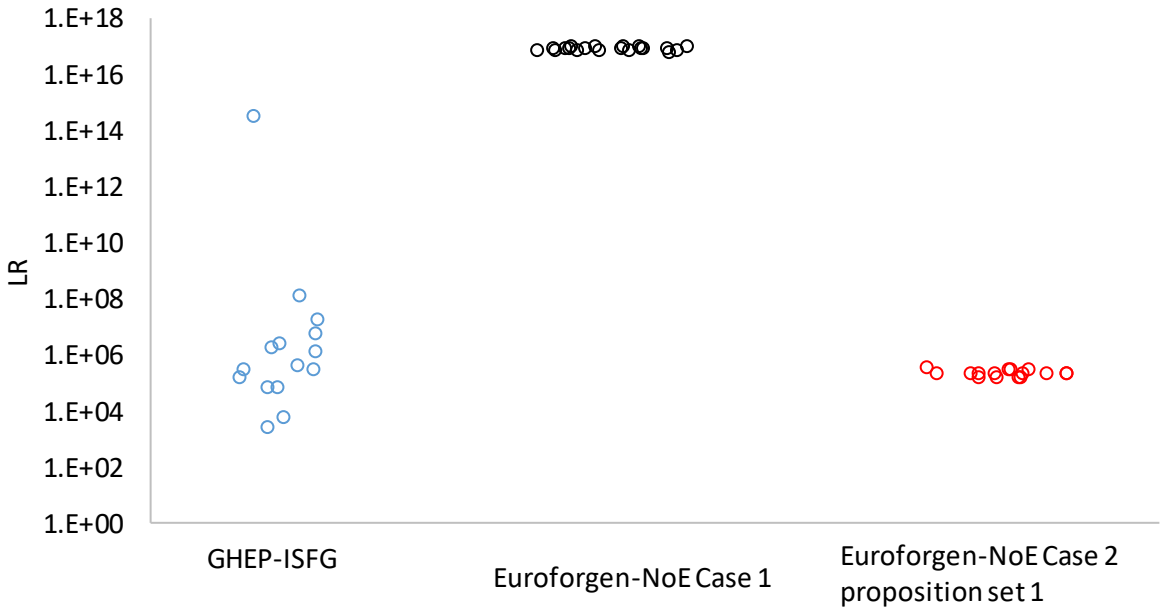
We briefly discuss here three published inter-laboratory studies that used PG software.

The Euroforgen-Network of Excellence [5] (hereafter NoE study) and the Spanish and Portuguese-Speaking Group of the International Society for Forensic Genetics (hereafter GHEP-ISFG study) [6] report inter-laboratory studies predominantly using the PG software LRmix and LRmix Studio.

The NoE study demonstrated little variation (see Figure 1) although for case 2 some labs did report alternative proposition sets (data given in [5]) either as well as or instead of the majority set. The GHEP-ISFG study had been specifically designed to present challenges[1]. Despite the use of a PG software there was a considerable spread of results even if attention was restricted to the one PG software, LRmix Studio (refer to Figure 1, derived from Table 1 of [6]).

[1] Donors were deliberately selected who possessed alleles differing by a single base pair resulting in unresolved allelic peaks. The authors also modified the epg provided to participants using Adobe® Photoshop®. Specifically, small modifications were made to the heights of some alleles while other alleles were removed entirely (P.A. Barrio 2018, pers, comm., 11 August). 44% of participating labs reported that their methodology had been validated to international requirements, 20% answered negatively, and 36% stated that they were in the process of validation.

Figure 1. The distribution of results in the NoE and GHEP-ISFG studies. The *x*-axis is included simply to spread the data so that they can be seen.

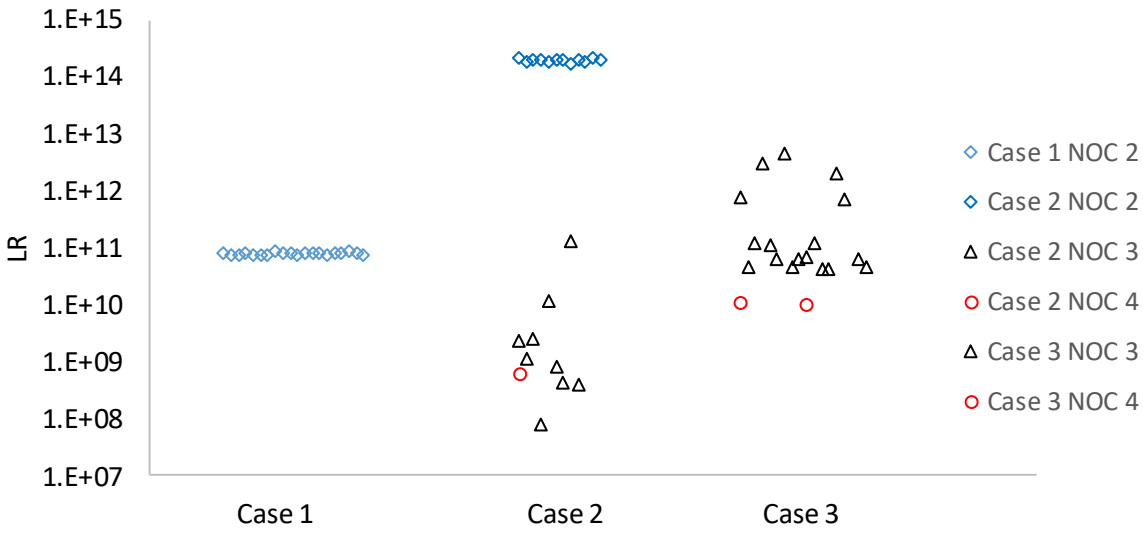


The authors of the GHEP-ISFG study reported that the large variance observed occurred due to subjective decisions made prior to use of the software (see Figure 1 of [6]). These decisions included whether peaks were stutter or allelic, and how to treat the two instances of non-resolution of allelic peaks separated by one base pair. These are the 16.3 and 17 alleles at locus D1S1656 and the 19.3 and 20 alleles at locus D12S391. In both cases shoulders can be seen on the detected peak suggesting the presence of an unresolved peak (particularly so at locus D12S391). These authors, and many of the participants[2], mention the need for and the lack of training.

The third PG study was reported by Cooper et al. [7] who surveyed 20 participants from Australia, New Zealand, the US, Canada, and the UK using the STRmix™ software (hereafter early STRmix™ study). The exercise involved the interpretation of Identifiler™ DNA profiles generated from three questioned samples. As the samples were casework samples, the true number of contributors to each sample was unknown. A reference Caucasian allele frequency file was provided. The results are shown in Figure 2.

[2] See last paragraph of section 3.2.1 of [6]

Figure 2. A display of the results from Cooper et al. [7]. Three cases were presented with 20 responses received. Some respondents did not provide a numerical response and do not appear on this graph. Others provided two values for different numbers of assigned contributors (*NoC*); in such cases both responses were plotted. In case three one respondent used only one replicate rather than the two provided. This response was removed.



The variation in the early STRmix™ study at least in regards to case 2 arises from different assignments of the number of contributors (*NoC*), with eleven participants assuming 2 contributors, seven participants assuming 3 contributors, and the remaining two participants not progressing an interpretation. Six participants also undertook a secondary interpretation, varying *NoC* hence in Figure 2 there are 24 data points for case 2. The increased spread of results for cases 2 and 3 *NoC* 3 and 4 arises because the extra contributor allows the minor to split into two contributors. A feature in STRmix™ developed subsequent to this trial allows the user to specify that the extra contributor is a trace by specifying the mixture proportion for that contributor. This would be expected to alleviate this effect. While the profiles provided originate from casework samples and the ground truth is not known, this variation indicates that subjective decisions prior to application of the software can lead to a wide range in the reported *LR*s.

Since the time of this study a body of knowledge has been developed about the effects of uncertainty in the number of contributors and management of that uncertainty. It has been shown that the correct number of contributors[3] usually gives the higher *LR* and that incorrect assignments, either over or under, tend to give conservative *LR*s or cause false exclusions [8, 9] for the true donors and slightly more adventitious matches of false donors usually giving *LR*s only slightly above 1. Later STRmix™ versions allow the user to inform the software of the approximate mixture proportions of the contributors prior to

[3] The correct number of contributors is not trivial to determine. The true number of contributors to casework profiles is always unknown and unknowable. In the case of constructed mixtures it is known how many donors were selected for use in the mixture. As long as each are present in a sufficient amount to contribute in some detectable way to the final signal they could be described as contributors. We acknowledge the subjective element to this.

deconvolution (referred to as M_x priors). This attribute has proven useful when dealing with mixtures believed to originate from closely-related individuals or if a very small trace contribution is suspected of being present. This policy has been implemented by several laboratories. Management of uncertainty in *NoC* is a recurrent theme within the forensic community. One approach to manage this uncertainty in *NoC* is described in Taylor et al. [10] who describe a mathematical approach for interpreting a DNA profile without specifying the number of contributors. Another way forward is to interpret the profile considering all reasonable options and report the range or the lowest *LR*. Note that this approach only addresses the uncertainty in *NoC* and does not address other decisions made by operators such as stutter management.

The National Institute of Standards and Technology's (NIST) MIX13 inter-laboratory study was mostly undertaken using manual interpretation methods and not PG [4], but subsequently was reanalyzed by Buckleton et al. [11] with several PG software. Originally, raw CE output for five mixed DNA profiles were distributed for interpretation. Variability in reported match statistics was due in part to differences in the interpretation methods used, but also due to the use of different allele frequencies, values for theta, and analytical and stochastic thresholds among laboratories.

Case 05 from this study, in particular, attracted considerable attention as it was over-engineered so that the profile was no longer indicative of the majority of casework and was intentionally designed to be difficult to interpret. This case consisted of a four-person mixture constructed to masquerade as a two-person mixture based on allele count alone, and resulted in 69% of participating laboratories including a known non-contributor to the mixture [4]. Buckleton et. al. [11] argue that it is very difficult or impossible to exclude this non-contributor manually (one participant did manage this however). However this case does demonstrate why the implementation of more sophisticated interpretation methods, such as PG, should be prioritised within the forensic community.

In the study herein, we aimed to refine the sources of variation in the reported *LR*. In order to facilitate this study, the key known variables were set such as the allele frequency database, values for theta, and the various STRmix™ parameters controlling the biological modelling of peaks that in normal casework were defined by internal validation studies. Propositions were set by the participants based on the same case information.

Methods

Two mixed GlobalFiler™ DNA profiles were submitted to participating laboratories. The profiles were taken from the PROVEDIt data set [12]. Sample 1 was RD14-0003-44_45_46_47-1;1;4;1-M3a-0.105GF-Q0.8 and Sample 2 was RD14-0003-30_31_32-1;4;4-M2a-0.75GF-Q0.6. Profiles were supplied as raw analysis (.hid) files as well as analysed text files (for participants unable to analyse 3500 data). Each profile was supplied with brief case scenarios and two reference profiles, detailed in Table 1. The epgs and provided STRmix™ input files are provided in the supplementary material.

Table 1: Case scenarios and reference profiles supplied.

Sample	Scenario	References
1	The DNA profile was obtained from a semen stained sample from underwear collected from the complainant after an alleged sexual assault. The male complainant alleges he has been sexually assaulted by two male individuals. DNA swabs from the complainant and a person of interest have been taken for analysis.	Two reference profiles have been submitted, one described as having come from the complainant and one from the suspect
2	The DNA profile was obtained from a semen stained anal swab after an alleged sexual assault. The male complainant alleges he has been sexually assaulted by two male individuals. DNA swabs from the complainant and a person of interest have been taken for analysis.	Two reference profiles have been submitted, one described as having come from the complainant ("Sample 2 complainant") and one from the suspect ("Sample 2 suspect").

Laboratories were asked to analyse both GlobalFiler™ .hid samples using ILS GeneScan™ 600 LIZ™ Size Standard using the per dye analytical thresholds (ATs) of 75, 100, 60, 80 and 100 rfu for the blue, green, yellow, red, and purple dyes, respectively, following those previously published [13]. Participants were asked to label and model allelic, back, and forward stutter peaks, review the case circumstances provided (Table 1), assign the number of contributors to each sample, and develop suitable propositions for a *LR* calculation (where appropriate).

STRmix™ kit and stutter files were provided for versions 2.4 and 2.5. Both back and forward stutter peaks were modelled as per Kelly et al. [13]. Profiles were interpreted with drop-in modelled, with a maximum cap of 150 rfu, and using a saturation threshold of 30,000 rfu.

174 participants from 42 laboratories submitted results for collation. A total of 349 interpretations were submitted for both profiles. 35.5% (124/349) used STRmix™ version 2.5 and the remaining 64.5% (225/349) used STRmix™ version 2.4. A summary of all the *LR*s from the submitted interpretations can be found within the supplementary material.

All of the STRmix™ version 2.4 interpretations were undertaken using the default number of MCMC accepts (100,000 burn-in accepts and 400,000 post burn-in accepts across all chains). The majority of STRmix™ version 2.5 interpretations (94/124) were undertaken using the default settings (8 chains of 100,000 burn-in accepts, 50,000 post burn-in accepts per chain). Thirty interpretations were undertaken with an increased number of accepts. Five of these interpretations with increased accepts originated from the same laboratory. This particular laboratory increases the number of accepts when interpreting a profile with 4 or more contributors. Increasing the number of MCMC accepts allows each of the chains to more thoroughly explore the sample space; this approach is expected to improve precision at the expense of longer run-times [14] but for more robust presentations of profiles in a mixture, the precision is not significantly improved.

All *LR*s were assigned using the FBI extended Caucasian allele frequencies ([15], also supplied to participants) using F_{ST}=0.01.

Results and discussion

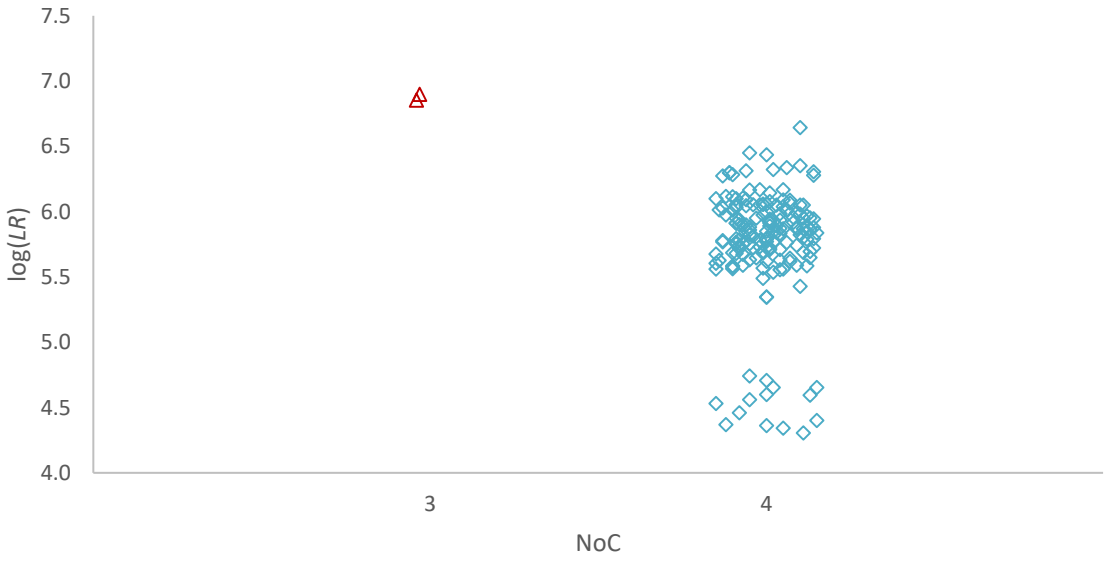
The experimental (ground truth) *NoC* for Sample 1 was four in the ratio 1:1:4:1, with total input DNA 0.105 ng. The complainant was the third contributor and the suspect the fourth. The experimental *NoC* for Sample 2 was three in the ratio 1:4:4, with total input DNA 0.75 ng. The suspect was the second contributor and the complainant the third.

A total of 173 STRmix™ results were submitted for Sample 1 with some participants submitting multiple STRmix™ interpretations varying the *NoC*. The majority of submissions (162/173) interpreted Sample 1 assigning *NoC* =4 whilst 11 submissions assigned Sample 1 as *NoC* =3. Two participants submitted interpretations assigning both *NoC* =3 and *NoC* =4 with comments indicating both *LR* values would be included in their reporting. Five participants chose to not progress an interpretation for Sample 1. A plot of log(*LR*) versus *NoC* is given in Figure 3.

Of the 11 responses where Sample 1 was interpreted assigning *NoC* =3, nine reported *LR*=0 for the POI (not plotted in Figure 3). This result was due to a single-locus exclusion at D18S51. The remaining two responses ignored locus D18S51 which resulted in an inclusionary *LR* for the suspect. No explanation was provided with the submissions as to why the D18S51 locus was ignored other than being indicated in the STRmix™ report. Both of these responses originated from different laboratories within the same multi-laboratory system.

Of the 173 results provided, all but one conditioned on the complainant. The resulting log(*LR*) for the non-conditioned interpretation was 5.7 (*NoC* =4) which fell within the distribution of log(*LR*)s produced for interpretations assuming the same *NoC* but conditioned on the complainant. Given that the complainant aligns with the major contributor to Sample 1 it would be expected that conditioning on his profile does not greatly assist with resolving the genotypes of the unknown contributors.

Figure 3: A plot of the log(*LR*) versus *NoC* for Sample 1. Submissions were interpreted assuming three (triangles) or four (diamonds) contributors.



The electropherogram for the green channel for Sample 1 at the D18S51 locus is shown in Figure 4 where it can be seen that seven peaks were detected. Given the results observed it

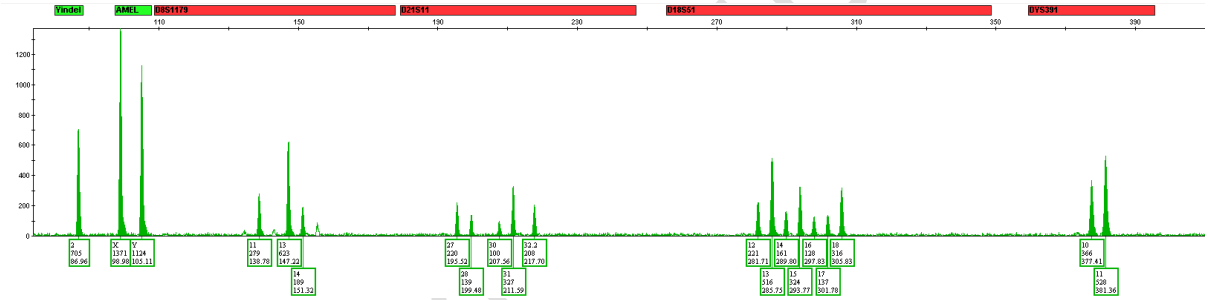
would be reasonable to assign each of these peaks as being at least partly-allelic in origin and either:

1. Assign the minimum number of contributors required to explain the mixture data as four, or

2. Assign the minimum number of contributors as three and model either the 16 or 17 peak at D18S51 as drop-in.

The ground truth *NoC* was four. Even though there are four contributors to this mixture, STRmix™ was able to proceed with an interpretation assuming three contributors by modelling drop-in at D18S51. Review of the genotype weights indicated that STRmix™ proposed that either the 16 or 17 peaks were drop-in events. It was able to do so as the heights of these peaks were below the drop-in cap used during interpretation (150 rfu, see Methods). This locus shows 6 alleles, hence for *NoC* = 3 all contributors must be heterozygotes with no shared alleles and hence this excludes the ground truth situation with is complainant (genotype = 13,18) and suspect's genotype of 13,15.

Figure 4: Sample 1 green channel. The D18S51 locus is the second from the right and shows seven peaks above the analytical threshold

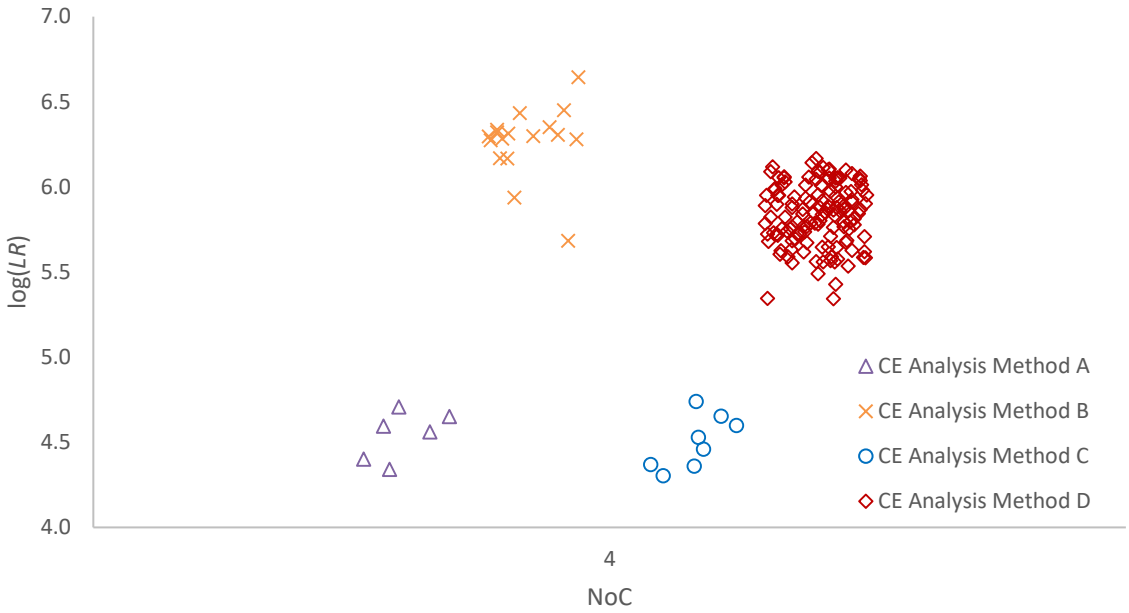


Examination of Figure 3 revealed that the log(*LR*)s for Sample 1 when run assuming four contributors clustered into two distinct groups. Further investigation revealed that this was largely due to differences in peak heights between input files. This is explored further in Figure 5. Differences in peak heights were due to differences in the Peak Detector tab settings of the local Analysis Method within GeneMapper® *ID-X*. Depending on the laboratory's GeneMapper® analysis settings for peak smoothing, normalisation, peak window size, and baseline window size, additional peaks were present or missing compared to the supplied STRmix™ text input file (see Table 2). The peak smoothing parameter smooths the outline of peaks and reduces the number of false peaks that are detected [16]. The normalisation parameter utilises the signal information of the size standards to adjust the observed peak heights across the entire profile [17]. This is useful for binary interpretation, or the comparison of samples run between multiple CE instruments or injection parameters. The peak window size, as well as the polynomial degree, affect the resolution of peaks. A higher polynomial degree and small peak window can help resolve shoulder peaks and increase sensitivity of peak detection [16]. The baseline window controls the baselining for the range of data points selected. A small baseline window can lead to smaller peak heights [16]. The impact of the differences in analysis settings on Sample 1 is demonstrated in Table 2 where the peaks detected at a single locus (D8S1179) are provided, along with the total number of peaks detected in the profile and the range of log(*LR*)s produced. The supplied electropherogram and input files were analysed using Method D.

Table 2. A summary of the GeneMapper® *ID-X* analysis settings used, peak heights observed at D8S1179, total number of peaks within input file (including Amelogenin), and range of observed inclusionary log(*LR*)s for Sample 1. The asterisk indicates one participant using analysis Method B removed the 15 peak at D8S1179 (=86 peaks) and another retained an 11.1 peak at D5S818 (=88 peaks). The differences between methods (compared to method A are indicated in bold).

Analysis method	Allele peak heights (rfu) for D8S1179				Total number of peaks/profile	Number of POI alleles aligning with a peak above AT	Range of log(*LR*)s
	11	13	14	15			
Method A: light smoothing, no normalisation, baseline window 51	263	602	181	N/A	82	30	4.3 to 4.7
Method B: light smoothing, **normalisation on**, baseline window 51	395	906	272	102	87 (range 86-88*)	32	5.7 to 6.6
Method C: light smoothing, no normalisation, **baseline window 33**	258	592	176	N/A	82	30	4.3 to 4.7
Method D: **no smoothing**, no normalisation, baseline window 51	279	623	189	N/A	85	32	5.3 to 6.1

Figure 5: A plot of the log(*LR*) for Sample 1 interpreted assuming *NoC* =4. The *x*-axis is included simply to spread the data so that they may be better observed. Yellow triangles are profiles analysed using CE analysis method A. Green crosses are profiles analysed using CE analysis method B. Blue circles are profiles analysed using CE analysis method C. Red diamonds are profiles analysed using CE analysis method D. The input file and epg provided by the authors was produced using CE analysis method D. Refer to Table 2 for detailed settings pertaining to each CE analysis method.



It seems likely that the two extra allelic peaks detected using methods B and D are the cause of the higher *LR*s for these methods. Although beneficial in this example this study is not adequate to suggest that these methods are beneficial in most or all instances.

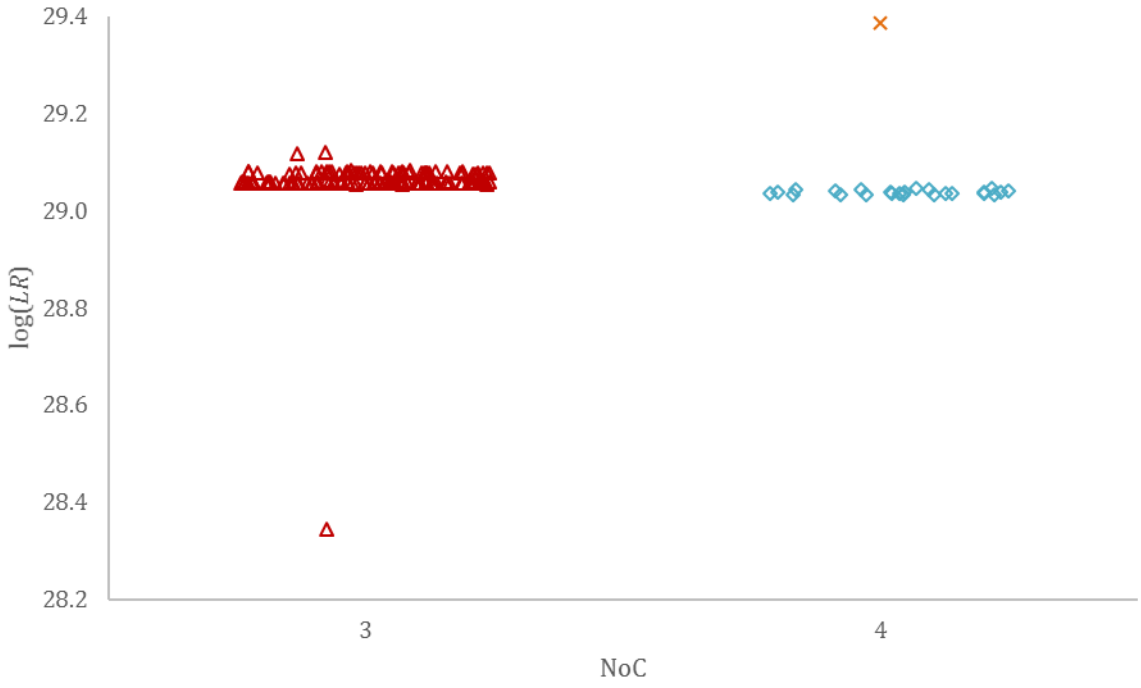
176 submissions were received for Sample 2. The experimental *NoC* for Sample 2 was three. 151 submissions assigned *NoC* =3 whilst the remaining 25 submissions assigned *NoC* =4. All interpretations conditioned on the complainant. One participant did not progress an interpretation of Sample 2. A number of participants submitted multiple STRmix™ interpretations. Two participants interpreted Sample 2 assuming both three and four contributors. One participant interpreted Sample 2 assuming three contributors using the provided text input file and then repeated the interpretation after carrying out his/her own analysis of the raw .hid file.

The log(*LR*)s produced when the suspect was compared with Sample 2 are plotted in Figure 6. The log(*LR*)s reported were highly reproducible and were largely unaffected by variations in the number of contributors assumed during interpretation. Examination of the mixture data revealed that Sample 2 appears to originate from two major contributors with at least one minor contributor. The major components of Sample 2 are consistent with originating from the complainant and the suspect. When interpreted as a four-contributor mixture, STRmix™ added the fourth contributor at trace levels (mixture proportion approximately 0 to 1%). This trace component, when added, did not interact with the genotype weights of the unknown major contributor resulting in minimal impact on the *LR*s produced for the suspect. The minimum log(*LR*) produced was 28.3 and resulted from one

participant ignoring the CSF1PO locus during interpretation. No explanation was provided as to why this locus was ignored. The maximum log(*LR*) produced was 29.4, as shown by the green cross in Figure 6. The participant's laboratory utilises a bespoke method to manage artifacts. The analyst created a pseudo-reference DNA profile with the genotype [1,1] at all but one locus. At the D7S820 locus where a putative artifact peak was observed, this peak was added to the pseudo-reference profile. The analyst interpreted the mixture as originating from five individuals and conditioned on the pseudo-reference profile in addition to the complainant from this case. This approach allowed STRmix™ to consider the putative artifact peak as having originated from the pseudo-contributor or from one of the other four contributors.

All of the remaining log(*LR*)s ranged from 29.0 to 29.1 and are consistent with run-to-run variability inherent in the Markov chain Monte Carlo process utilised by STRmix™.

Figure 6: A plot of the log(*LR*) versus *NoC* =3 (Red triangles) and *NoC* =4 (Blue diamonds) for Sample 2. The datum at *NoC* =4, log(*LR*) = 29.4 indicated by the green cross, is from a laboratory that uses a bespoke method to manage artifacts. The datum at *NoC* =3, log(*LR*) = 28.3 was from one participant ignoring a locus during interpretation. Such a difference would not have any practical impact, contrary to what was seen for Sample 1 (which was low template, with the POI being a minor contributor).



The effect on the *LR* by varying *NoC* has previously been explored [13, 18, 19]. The magnitude of *LR* variability due to differences in CE analysis methods has not been previously reported. The variability observed when participants used identical input files and assigned the same *NoC* for STRmix™ interpretation is attributable to run-to-run Markov chain Monte Carlo variability [1]. The *LR* variability due to MCMC variability was within one order of magnitude.

Some participating laboratories have chosen not to internally validate STRmix™ beyond three person mixtures or have had little experience with either GlobalFiler® 3500 data or STRmix™ use in casework. These factors likely led to increased uncertainty in those respective laboratories' approaches to the interpretation of the profiles in this study, which explains why five participants elected not to interpret the Sample 1 profile.

Sample 1 represented a complex DNA profile with low peak heights. As part of the PROVEDIt dataset, it is known that this mixture was prepared using DNA from four contributors mixed in the ratio 1:1:4:1 and amplified with 0.105 ng total template. The Complainant 1 and Suspect 1 references represented the third and fourth contributor positions respectively in the listed ratio. The profile was challenging in its complexity due to high *NoC* and a high level of stochastic variation observed between/among loci. Of the 174 participants, a few chose not to report either Sample 1 (5 participants) or reported multiple interpretations. Several participants commented that in a casework situation they would not interpret Sample 1 due to its complexity, i.e. assigned *NoC* > 3 and exhibiting low level peaks below their locally established stochastic threshold. Some participants commented they would prefer a PCR replicate to address the uncertainty in *NoC*, and to see if the imbalance is due to stochastic effects or not. One participant indicated the laboratory is required by law to use replicates.

The requirement for allelic dropout of one of Suspect 1's alleles at two loci under the contributor proposition (inclusion of the POI) led a few participants to consider a comparison to Sample 1 as inconclusive (1 participant) or an outright exclusion (2 participants). Furthermore, the complexity of Sample 1 is reflected in the range of *LR* values calculated for Suspect 1.

Taking a close look at the Sample 1 epg (supplementary data) assuming that Suspect 1 had contributed DNA to Sample 1 (i.e. contributor hypothesis), then allelic dropout must have occurred at both the vWA (allele 18) and D1S1656 (allele 16) loci. No sub-threshold peaks are visible in these positions (AT 75 rfu and 100 rfu, respectively). A common theme to participant explanations provided with Sample 1 was a reluctance to consider dropout given the heights of the surviving peaks. We note the absolute absence of even sub-threshold peaks where ground truth is that this allele was present amongst the donors. This decision making behaviour is an understandable carry over from binary methodology, and likely accumulated over years from using older CE technology and more robust kits with fewer PCR cycles. One of the strengths of moving to a PG system is the ability to model allelic dropout probabilistically, allowing the probability of dropout to be incorporated into a continuous model *LR*, rather than being required to make a binary decision upfront about presence or absence. If the probability of the evidence given the hypothesis supports exclusion (i.e. the non-contributor or alternate proposition) then this will be reflected in an *LR* less than 1.

Sample 2 represented a profile with less ambiguity than Sample 1. The mixture in Sample 2 was constructed using three contributors mixed in the ratio 1:4:4 amplified with 0.75 ng total template. Overall, peak heights were higher, with less stochastic peak height variation observed compared with Sample 1. Suspect 2 and Complainant 2 references represented the second and third contributor positions in the listed ratio. Only one participant chose not to submit an interpretation of Sample 2 (explanation not provided), and two participants submitted two interpretations of Sample 2 under different assigned *NoC*. Assigned *NoC* ranged from three to four for Sample 2. Despite this range, it had little impact on the *LR* as shown by the log(*LR*) range of 1.04. One submission ignored the CSF1PO locus with no explanation, resulting in a lower *LR*. This outcome is expected when a locus with an *LR* greater than 1 is omitted from the calculation. The log(*LR*) range for all submissions that did not exclude any loci or utilised a unique method to deal with artifacts is 0.20.

Again, lack of familiarity with the GlobalFiler® multiplex caused some participants increased uncertainty with their interpretation. In particular, several participants commented they would have liked more information about expected stutter rates with this system to better inform their assignment of the *NoC*. The inter-laboratory study showed the differences in log(*LR*) due to the interpretation software STRmix™ were smaller than those introduced by differences in peak height due to differences in the analysis software settings. Varying *NoC* was also shown to result in differences in *LR*. Most evidently, under-assigning *NoC* can result in an exclusion for a true contributor, which is a known outcome [18].

In the supplementary results table, the range of log(*LR*)s for intra-laboratory comparisons is provided. The largest single laboratory (intra-laboratory) range of log(*LR*)s for Sample 1 was 2.09, which appears to be due to the use of different CE analysis methods by analysts. For the other laboratories where multiple participants submitted results for Sample 1, the intra-laboratory range did not exceed one order of magnitude. For Sample 2, with the exception of one laboratory where one participant ignored the CSF1PO locus, the intra-laboratory range of log(*LR*)s for each laboratory system did not exceed 0.05. These results support that PG and concomitant interpretation guidelines contribute to the reproducibility of the *LR* within each laboratory.

Conclusion

Most of the decisions outlined as troublesome in the GHEP-ISFG study are automated in the PG solution STRmix™. The only decision from their list that is not automated is the non-resolution of peaks separated by one base pair.

Five participants either called Sample 1 inconclusive with respect to the POI or excluded them without use of STRmix™. Nine participants reported an *LR* of 0 from STRmix™ arising from assigning *NOC*=3. Non-zero point estimate *LR*s ranged from 2.02×10^4 to 7.92×10^6. The *LR*s for Sample 2 ranged from 2.21×10^{28} to 2.43×10^{29}. Where *LR*s were calculated, the differences between participants can be attributed to (from largest to smallest effect):

- Too few contributors (*NoC)* for sample 1, the effect is downwards on the *LR*, provoking a false exclusion.
- The exclusion of one locus for sample 2 within the interpretation, the effect is downwards on the *LR*.
- Differences in local CE data analysis methods leading to variation in the peaks present and their heights in the input files used, the effect is downwards on the *LR.*
- Too few contributors (*NoC)* for sample 1 and the exclusion of one locus, the effect is slightly upwards on the *LR*,
- Run-to-run variation due to the random sampling inherent to all MCMC-based methods. The effect is minor variability in the *LR.*

Because the intra-laboratory *LR*s are a subset of the inter-laboratory *LR*s, the variability in the *LR* within each laboratory can be attributed to the same sources of variation observed in the inter-laboratory results – loci exclusion, CE analysis methods, number of contributors, and MCMC variation – as described above.

In conclusion, the study herein supports that there is a high level of repeatability and reproducibility among the participants. In those results that differed from the mode, the

differences in *LR* were almost always minor or conservative. We attribute this inter/intra-laboratory convergence to the use of PG software. This is a highly desirable and pleasing outcome for the forensic biology field.

Acknowledgements

This work was supported in part by grant NIJ 2017-DN-BX-0136 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 organisations.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.fsigen.2019.01.006.

References

[1] Bright J-A, Stevenson KE, Curran JM, Buckleton JS. The variability in likelihood ratios due to different mechanisms. Forensic Science International: Genetics. 2015;14:187-90.

[2] Bright J-A, Evett IW, Taylor D, Curran JM, Buckleton J. A series of recommended tests when validating probabilistic DNA profile interpretation software. Forensic Science International: Genetics. 2015;14:125-31.

[3] Evett IW, Berger CEH, Buckleton JS, Champod C, Jackson G. Finding the way forward for forensic science in the US—A commentary on the PCAST report. Forensic Science International. 2017;278:16-23.

[4] Butler JM, Kline MC, Coble MD. NIST Interlaboratory Studies Involving DNA Mixtures (MIX05 and MIX13): Variation Observed and Lessons Learned. Forensic Science International: Genetics. 2018;37:81-94.

[5] Prieto L, Haned H, Mosquera A, Crespillo M, Alemañ M, Aler M, et al. Euroforgen-NoE collaborative exercise on LRmix to demonstrate standardization of the interpretation of complex DNA profiles. Forensic Science International: Genetics. 2014;9:47-54.

[6] Barrio PA, Crespillo M, Luque JA, Aler M, Baeza-Richer C, Baldassarri L, et al. GHEP-ISFG collaborative exercise on mixture profiles (GHEP-MIX06). Reporting conclusions: Results and evaluation. Forensic Science International: Genetics. 2018;35:156-63.

[7] Cooper S, McGovern C, Bright JA, Taylor D, Buckleton J. Investigating a common approach to DNA profile interpretation using probabilistic software. Forensic Science International: Genetics. 2015;16:121-31.

[8] 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 Science International: Genetics. 2018;34:11-24.

[9] Bright J-A, Taylor D, Curran J, Buckleton J. Searching mixed DNA profiles directly against profile databases. Forensic Science International: Genetics. 2014;9:102-10.

[10] Taylor D, Bright J-A, Buckleton J. Interpreting forensic DNA profiling evidence without specifying the number of contributors. Forensic Science International: Genetics. 2014;13:269-80.

[11] Buckleton JS, Bright J-A, Cheng K, Budowle B, Coble MD. NIST Interlaboratory Studies Involving DNA Mixtures (MIX13): A modern analysis. Forensic Science International: Genetics. 2018;37:172-9.

[12] 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 Science International: Genetics. 2018;32:62-70.

[13] Kelly H, Bright J-A, Kruijver M, Cooper S, Taylor D, Duke K, et al. A sensitivity analysis to determine the robustness of STRmix™ with respect to laboratory calibration. Forensic Science International: Genetics. 2018;35:113-22.

[14] 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 Science International: Genetics. 2016;23:226-39.

[15] Moretti TR, Moreno LI, Smerick JB, Pignone ML, Hizon R, Buckleton JS, et al. Population data on the expanded CODIS core STR loci for eleven populations of significance for forensic DNA analyses in the United States. Forensic Science International: Genetics. 2016;25:175-81.

[16] Applied Biosystems. GeneMapper®ID-X Software Version 1.5. 2015. https://assets.thermofisher.com/TFS-Assets/LSG/manuals/100031707_GeneMapIDX_ver1_5_ReferenceGuide.pdf Accessed 22 August 2018.

[17] Shewale JG, Qi L, Calandro L. Principles, Practice, and Evolution of Capillary Electrophoresis, as a Tool for Forensic DNA Analysis. In: Shewale JG, editor. Forensic DNA Analysis: Current Practices and Emerging Technologies. Boca Raton: CRC Press; 2013. p. 131-62.

[18] Bright J-A, Curran JM, Buckleton JS. The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation. Forensic Science International: Genetics. 2014;12:208-14.

[19] Moretti TR, Just RS, Kehl SC, Willis LE, Buckleton JS, Bright J-A, et al. Internal validation of STRmix for the interpretation of single source and mixed DNA profiles. Forensic Science International: Genetics. 2017;29:126-44.