Forensic Science International: Genetics 37 (2018) 81–94

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

NIST interlaboratory studies involving DNA mixtures (MIX05 and MIX13): Variation observed and lessons learned

John M. Butler[a,*], Margaret C. Kline[b], Michael D. Coble[b,1]

[a] National Institute of Standards and Technology, Special Programs Office, Gaithersburg, MD 20899, United States
[b] National Institute of Standards and Technology, Applied Genetics Group, Gaithersburg, MD 20899, United States

ARTICLE INFO

Keywords:
Forensic DNA
DNA mixture
Mixture interpretation
Interlaboratory study
Collaborative exercise
MIX05
MIX13
Forensic science

ABSTRACT

Interlaboratory studies are a type of collaborative exercise in which many laboratories are presented with the same set of data to interpret, and the results they produce are examined to get a "big picture" view of the effectiveness and accuracy of analytical protocols used across participating laboratories. In 2005 and again in 2013, the Applied Genetics Group of the National Institute of Standards and Technology (NIST) conducted interlaboratory studies involving DNA mixture interpretation. In the 2005 NIST MIX05 study, 69 laboratories interpreted data in the form of electropherograms of two-person DNA mixtures representing four different mock sexual assault cases with different contributor ratios. In the 2013 NIST MIX13 study,108 laboratories interpreted electropherogram data for five different case scenarios involving two, three, or four contributors, with some of the contributors potentially related. This paper describes the design of these studies, the variations observed among laboratory results, and lessons learned.

1. Introduction

Interlaboratory comparisonstudies, which are sometimes referred to as collaborative exercises or round-robin studies, provide a useful way to demonstrate that multiple laboratories can generate comparable results with the same provided samples, and are cited as valuable methods for assessing measurement reproducibility in accredited laboratories [1]. Interlaboratory studies are regularly used in clinical DNA diagnostics (e.g., [2]) and other scientific communities. Given that DNA databases used in criminal investigations compile data from many jurisdictions, it is valuable to assess the degree to which laboratories across jurisdictions produce comparable analytical results.

Interlaboratory studies, which are typically voluntary, assess progress on the standardization of methods across laboratories and enable technical and statistical issues to be ascertained and discussed. Most analysts focus on their own laboratory protocols and rarely get an opportunity to determine how their laboratory performs relative to others. Intralaboratory evaluations examine performance across analysts within the same laboratory and can be useful in assessing whether further training on following protocols is needed to improve consistency. Both intra- and inter-laboratory studies can help better understand causes of variability among laboratories and analysts – and hopefully lead to improvement of the entire community.

It is important to recognize that interlaboratory studies tend to be research-focused and are not meant to evaluate the performance of individual analysts. Although errors made by laboratories are noted in interlaboratory study publications, finding these errors is not typically the primary objective of a study. Any errors detected reveal opportunities for improvement (see Ref. [3]) based on the research question being explored and cannot normally be used to formally assess operational error rates for a general activity as has been advocated for proficiency test data that is produced under standard conditions (see Ref. [4]).

DNA mixtures arise when biological material from two or more individuals contributes to the sample being tested, and different types or categories of mixtures have been proposed [5]. Methods for deconvoluting mixtures were first described about two decades ago [6]. In 2006, the DNA Commission of the International Society of Forensic Genetics (ISFG) stated in their "Recommendations on the interpretation of mixtures" article that "our discussions have highlighted a significant need for continuing education and research into this area" [7]. Interlaboratory studies can enable monitoring of variability in practice and overall laboratory performance with different types of DNA mixtures.

About a dozen interlaboratory studies exploring DNA mixture interpretation with short tandem repeat (STR) markers have been performed over the past two decades [8–18] to examine various aspects of

* Corresponding author at: NIST Special Programs Office, 100 Bureau Drive, Mail Stop 4701, Gaithersburg, MD 20899-4701, United States.
E-mail address: john.butler@nist.gov (J.M. Butler).
[1] Current address: University of North Texas Health Science Center, Fort Worth, TX 76107, United States.

https://doi.org/10.1016/j.fsigen.2018.07.024
Received 26 April 2018; Received in revised form 24 July 2018; Accepted 31 July 2018
Available online 01 August 2018

Table 1
Interlaboratory studies involving STR multiplexes and DNA mixtures.

Study (when conducted)	Publication	# Labs (# data sets)	# Mixtures	Samples Provided and Study Purpose
STR triplex CTT (Dec 1995–May 1996)	Kline et al. [8]	34 (46)	0	4 single-source DNA extracts, 4 single-source stains to explore factors affecting sizing variability
NIST Mixed Stain Study (MSS) 1 (April–Nov 1997)	Duewer et al. [9]	22 (37)	5	Buffy coat cells on S&S 903 paper; 6 single-source, 4 two-source mixtures, and 1 three-source mixture to explore donor types obtained given a complete set of reference sources
NIST MSS2 (Jan–May 1999)	Duewer et al. [9]	45 (70)	2	Part A: 4 single-source stains, 1 two-source stain, 1 three-source stain; Part B: 5 vials of a four-level DNA concentration series to explore donor types obtained given an incomplete set of reference sources (Part A) and to examine performance of DNA quantitation assays (Part B)
NIST MSS3 (Dec 2000–Oct 2001)	Kline et al. [10] and Duewer et al. [11]	74 (117)	6	DNA extracts; 1 single-source, 5 two-source mixtures (3:1 to 10:1 component ratios), and 1 three-source mixture (4:2:1) to explore the effect of quantitation on STR typing performance
GHEP-MIX01 (2010)	Crespillo et al. [12]	32 (32)	4	Questionnaire and data (.fsa files) provided with 2 STR kits (Identifiler and PP16) for 4 two-source mixtures (1F:5 M, 1F:10 F, 1F:1 M, 5F:1 M) to explore errors (discrepancies) obtained during mixture interpretation
GHEP-MIX02 (2011)	Crespillo et al. [12]	24 (24)	2	Questionnaire and data (.fsa files) provided with 1 STR kit (Identifiler) for 1 two-person mixture (1M:5 F) and 1 three-person mixture (2F:1M:1 M) to explore statistical treatment of results under a common set of hypotheses
GHEP-MIX03 (2012)	Crespillo et al. [12]	17 (17)	3	Questionnaire and data (.fsa files) provided with 2 STR kits (Identifiler Plus and NGM) for 2 two-person mixtures (1F:5 M, 1F:10 F) and 1 three-person mixture (1F:3M:7 M) to explore statistical treatment of results under an open set of hypotheses for the likelihood ratios used
EuroForGen-NoE (2013)	Prieto et al. [13]	18 (20); 18 (22)	2	Data (csv format) provided for 16 STR loci with case scenarios; two exercises each involving a two-person mixture were supplied along with victim and suspect profiles, population allele frequencies, and LRmix software to explore impact of training and whether standardization of an approach could be demonstrated
UK Forensic Science Regulator (2014)	Unpublished report	8 (18)	5	DNA extracts provided with case scenarios for 2 two-person mixtures (4:1, 2:1) and 3 three-person mixtures (6:4:1, 6:3:1, 7:1.5:1) to explore variability across UK forensic science providers
DFSC Mixture Study (2014–2015)	Aranda [14] *(presentation only)*	55 (185)	6	Data provided for 15 STR loci (Identifiler Plus) involving 4 two-person mixtures and 2 three-person mixtures to explore intra- and inter-laboratory variation in genotype determinations to better understand the current state and potential limitations of mixture interpretation
STRmix study (2014)	Cooper et al. [15]	12 (20)	3	Data provided for 15 STR loci (Identifiler) involving three casework samples (ground truth not known) to explore the improved level of agreement that was possible within and between laboratories using a common probabilistic genotyping software program
22nd GHEP-ISFG IE Basic (2014)	Toscanini et al. [16]	72	1	Two-source stain: 2:1 mixture (v/v) saliva/blood; results generated with autosomal STRs, Y-STRs, X-STRs, and mtDNA; to explore various approaches being used for mixture interpretation and technical difficulties observed
22nd GHEP-ISFG IE Advanced (2014)	Toscanini et al. [16]	52	1	Two-source stain: 4:1 mixture (v/v) saliva/semen; results generated with autosomal STRs, Y-STRs, X-STRs, and mtDNA; to explore various approaches being used for mixture interpretation and technical difficulties observed
GHEP-MIX06 (2015)	Barrio et al. [18]	25	2	Data (pdf files) provided for 15 STR loci (NGM) involving a three-person (7M:3F:1 M) mixture and 17-YSTRs (Yfiler) involving a two-male (3:1) mixture; participants were provided with mock case information and analytical, stochastic, and stutter thresholds used; to explore how results would be reported if this exercise were a real case
NFI-led study (2016)	Benschop et al. [17]	3 (26)	10	Data (pdf files) provided for 15 STR loci (NGM) with replicates involving 2 two-person (1:1, 5:1), 4 three-person (1:1:1, 5:1:0.2, 10:1:1, 10:1:1), 2 four-person (5:1:1:1, 5:1:1:1), and 2 five-person (2:2:1:1:1, 2:2:1:1:1) mixtures and some person of interest reference profiles to explore intra- and inter-laboratory variability
NIST MIX05 (Feb–Sept 2005)	*This article (and several presentations)*	69 (75)	4	Data (.fsa files) provided from 6 STR kits; 4 two-person mixture "evidence" profiles (3F:1 M, 1F:3 M, 1F:1 M, 7F:1 M) with female "victim" reference profiles; no "suspect" male reference profiles supplied for comparison purposes to explore mixture deconvolution approaches
NIST MIX13 (Aug–Dec 2013)	*This article (and several presentations)*	108 (163)	5	Data (.fsa files) provided from 2 STR kits with case scenarios; 5 "cases" involving two- (1:1, 3.5:1), three- (6:1.5:1, 7:2:1), or four- contributors (1:1:1:1) with "person of interest" reference profiles, some of which were not in the mixtures to explore variability in overall mixture interpretation

Table 2
Study design and overview for NIST studies MIX05 and MIX13.

	MIX05 (2005)	MIX13 (2013)
Responses received	69 labs (1 lab providing results from 7 analysts)	108 labs (4 labs providing results from 8, 10, 16, or 25 analysts)
Data supplied	Electronic (.fsa) ABI 3100 files for six STR kits: Identifiler, Profiler Plus, COfiler, SGM Plus, PowerPlex 16, and FMBIO files for PowerPlex 16 BIO	Electronic (.fsa) ABI 3130xl files for two STR kits: Identifiler Plus and PowerPlex 16HS
Data collection timeframe	February to September 2005	August to December 2013
Results announced	ISHI 2005 poster and workshop presentation (September 26–28, 2005); additional presentations given 2006 to 2008 to inform stakeholders and the community	NIST/FBI-sponsored DNA Technical Leader Summit (November 20–21, 2013); additional presentations given 2014 to 2016 to inform stakeholders and the community
Number of "cases" provided	4 cases with no case scenarios	5 cases with case scenarios
Case types being mimicked	Sexual assault evidence without "suspect" profiles for comparison	Sexual assault & touch evidence with potential persons of interest
Reference profiles provided?	Female "victim" reference profile was given in each case; no male "suspect" references were provided for comparison	Multiple reference profiles were provided including ones that were not in the mixture
Mixture complexity	2-person mixtures (male/female); all samples were unrelated; various major/minor ratios and degrees of allele overlap	2, 3, > 3-person mixtures; involved profiles from related individuals, low-template data, and inclusion/exclusion challenges
Challenges provided	Amelogenin X null allele (Case 3) and tri-allelic pattern at TPOX (Case 4)	Non-contributor reference given with a four-person mixture that exhibited no more than four alleles at any locus; case scenario involving a potential brother of the person of interest

DNA analysis (Table 1). Previous interlaboratory studies conducted by the U.S. National Institute of Standards and Technology (NIST) have demonstrated: (1) that laboratories have instruments with different sensitivities, (2) different levels of experience and training play a part in effective mixture interpretation, and (3) the amount of input DNA affects the ability to detect the minor component in a mixture [8–11].

The interlaboratory studies described in this paper were conceived and conducted with the goal of better understanding the "lay of the land" regarding analysis of DNA mixtures at the time (Table 2). While the general findings of the NIST 2005 (MIX05) and 2013 (MIX13) studies have been presented numerous times by the authors over the years (e.g., [19–21]) and are widely known, we are including full details of these landmark studies here for historical purposes and as an opportunity to reflect on lessons learned.

Findings from the MIX05 study influenced development of the Scientific Working Group on DNA Analysis Methods (SWGDAM) "SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Laboratories" released in 2010 [22] and updated in 2017 [23]. Findings from the MIX13 study were initially shared at a DNA Technical Leader's Summit in November 2013 and have influenced the U.S. forensic community in recent years to move towards probabilistic genotyping approaches for complex DNA mixtures (see Ref. [24]). Findings from both studies have brought awareness of differences in approaches to DNA mixture interpretation and have highlighted the need for improved training and validation, which have hopefully led to improved protocols over the years.

2. Materials and methods

The goal of the MIX05 and MIX13 studies was to examine sources of variability in interpretation rather than instrument sensitivity or amount of DNA being examined. Therefore, these studies involved sharing electronic files of DNA profiles with study participants rather than sharing of biological samples. The STR profiles used for these two studies are available on the NIST STRBase website at https://strbase.nist.gov/interlab/MIX05.htm and https://strbase.nist.gov/interlab/MIX13.htm. These profiles have been downloaded and used over the years for training purposes by many laboratories. Study design for MIX05 and MIX13 built on previous NIST work with earlier interlaboratory studies and is summarized in Table 2. Participants in each study are listed in Supplemental Table S1 (MIX05) and Supplemental Table S2 (MIX13).

2.1. MIX05

2.1.1. Participant enrollment

An invitation letter was prepared announcing the MIX05 interlaboratory study and explaining the purpose and plan for distribution of results. Initial enrollment through announcements and handouts were made at the following forensic DNA meetings: National CODIS Conference (held in Washington, D.C., November 15, 2004), the International Forensic Y User's Group (held in Berlin, Germany, November 20, 2004), and the Scientific Working Group on DNA Analysis Methods (held in Quantico, VA, January 18, 2005). Emails were sent to previous participants in NIST interlaboratory studies such as Mixed Stain Study 3 [10] and DNA Quantitation Study 2004 [25]. A total of 94 laboratories enrolled in MIX05, and 69 supplied results for analysis and comparison before the study closed in September 2005. Not every laboratory supplied results on every case.

All enrolled participants were provided with electronic data files on a CD-ROM. Data generated with the ABI 3100 (Thermo Fisher Scientific,[2] South San Francisco, CA) were also made available on the NIST STRBase website https://strbase.nist.gov/interlab/MIX05.htm. A handful of labs requested FMBIO data, which was generated at the Pennsylvania State Police (Greensburg, PA) or the Arkansas State DNA Laboratory (Little Rock, AR) from PowerPlex 16 BIO (Promega Corporation, Madison, WI) using polymerase chain reaction (PCR) products amplified at NIST. Most labs were supplied with the data for this study by early February 2005 and returned results by mid-March 2005. Labs were asked to interpret the provided data using their own protocols, and to supply those protocols and their reasons for making specific allele calls and mixture interpretation conclusions. Collection of MIX05 results was completed at the end of September 2005 with publication of the correct answers via a poster at the International Symposium on Human Identification held September 26–28, 2005. This poster is available at https://strbase.nist.gov/interlab/MIX05/MIX05poster.pdf.

2.1.2. Sample selection

Samples were selected for the MIX05 study based on review of all possible genotype combinations from 40 females and 660 males previously examined with the 15 STRs present in the Identifiler kit [26]. Genotypes for these samples may be found at http://strbase.nist.gov/NISTpopdata/JFS2003IDresults.xls. David Duewer, from the NIST

[2] ThermoFisher Scientific was known as Applied Biosystems in 2005 and Life Technologies in 2013 (see Ref. [36], p. 26).

Chemical Sciences Division, developed a Microsoft Excel-based computer program dubbed *Virtual MixtureMaker* to perform these pair-wise comparisons. The program is available at https://strbase.nist.gov/software.htm. Sample combinations were selected to explore the ability to reliably deduce alleles and genotypes from two-person mixtures with various contributor ratios and overlap of alleles. We typically aimed for a moderate degree of overlap meaning 3–5 loci with resolvable heterozygotes (e.g., 4 alleles observed in a two-person mixture).

2.1.3. Data generation

After various allele combinations were selected with a plan to mix one male and one female, mixture ratios were chosen to reflect some possible mock sexual assault casework scenarios. DNA extracts, which had previously been extracted and initially quantified as described previously [26], were re-quantified using the Quantifiler kit (Applied Biosystems) according to manufacturer recommendations on an ABI 7500 (Applied Biosystems). Based on these quantitation values, the "perpetrator" and "victim" DNA extracts were mixed in bulk at the desired ratios (e.g., 1:3 or 1:7) to form the "evidence" mixture.

Aliquots of the mixture or the single source victim and perpetrator DNA samples were used to generate PCR products following manufacturer recommended full-volume amplification conditions for the STR typing kits **Profiler Plus** (Applied Biosystems), **COfiler** (Applied Biosystems), **Identifiler** (Applied Biosystems), **SGM Plus** (Applied Biosystems), **PowerPlex 16** (Promega Corporation), and **PowerPlex 16 BIO** (Promega Corporation). All amplifications were performed in GeneAmp 9700 thermal cyclers with the manufacturer-recommended number of cycles (e.g., 28 cycles for Identifiler). The PowerPlex 16 BIO samples were run on an FMBIO II instrument (Hitachi Genetic Systems, Alameda, CA). Samples from all other kits were evaluated on an ABI 3100 Genetic Analyzer (Applied Biosystems) using a 36-cm array, POP-6 polymer, 10 s at 3 kV electrokinetic injections, and Data Collection software 1.0.1 (i.e., no variable binning of the various dye colors). PowerPlex 16 BIO gel images were evaluated with FMBIO software (Hitachi Genetic Systems). All other data evaluation was performed with GeneScan 3.7 and Genotyper 3.7 or GeneMapper*ID* 3.2 software (Applied Biosystems).

2.1.4. Sample details

Genomic DNA samples with specific allele combinations ("evidence") were mixed in the following ratios: **Case #1 evidence**, where the victim was the major contributor, was a mixture of three parts female DNA and one part male DNA; **Case #2 evidence**, where the perpetrator is the major contributor, was a mixture of one part female DNA and three parts male DNA; **Case #3 evidence**, with a fairly balanced mixture of approximately one part female and one part male, contained a male sample that lacked the amelogenin X amplicon; **Case #4 evidence**, was designed to be a more extreme mixture with seven parts female and one part male DNA (the male contained tri-allelic pattern at TPOX). Single-source female "victim" reference DNA profiles and the mixture "evidence" DNA profiles for each case (along with allelic ladder, positive and negative controls) were supplied. Labs were then asked to deduce the perpetrator DNA profile without any suspect (s) reference DNA profile(s) being supplied (Supplemental Table S3).

2.1.5. Scenarios provided

No mock cases scenarios were provided.

2.1.6. Data supplied

Enrolled participants were supplied with all STR profiles and could choose what kit data to examine based on their experience and laboratory protocols. Generally, Identifiler data were of poorer quality in the electropherograms provided, which caused some labs to not return results (they indicated a desire for higher quality data through sample re-injection to reduce pull-up prior to data interpretation). FMBIO data, which were generated in the Pennsylvania State Police Lab and Arkansas State Crime Lab using NIST-created PCR products, were supplied separately to laboratories requesting it.

2.1.7. Information requested for study

MIX05 participants were asked to provide the following information: "a) Report the results as though they were from a real case including whether a statistical value would be attached to the results. *Please summarize the perpetrator(s) alleles in each "case" as they might be presented in court—along with an appropriate statistic (if warranted by your laboratory standard operating procedure) and the source of the allele frequencies used to make the calculation.* Please indicate which kit(s) were used to solve each case; b) Estimate the ratios for samples present in the evidence mixture and how this estimate was determined; and c) Provide a copy of your laboratory mixture interpretation guidelines and a brief explanation as to why conclusions were reached in each scenario." Several participants noted that they did not routinely determine the ratios of mixture components according to their laboratories' standard operating procedures. Some of the participants returned only part of the information requested.

2.1.8. Results collation and summary

For collation of results in an Excel spreadsheet, laboratory participants were deidentified through assigning a number based on the order in which laboratories expressed an interest in participating. Note that a rank-ordered list of laboratory identification numbers is not necessarily sequential and goes beyond the total number of data sets received because not all laboratories that expressed interest submitted data. Information provided was summarized and entered into a spreadsheet to enable comparison (see Supplemental File S1). As with all previous NIST interlaboratory studies, and because the purpose of this study is not to serve as a proficiency test of any particular laboratory's or analyst's performance, but rather to gain an appreciation of variation that may exist in approaches taken at the time, results are reported in an anonymous fashion. MIX05 participants received a certificate for their participation in the study and a copy of the poster presented in September 2005 displaying "correct" results for the "perpetrator" STR profile in each case scenario. Laboratories were then invited to assess their own individual performance against the correct result.

2.2. MIX13

2.2.1. Participant enrollment

In early 2013, NIST and the FBI Laboratory's CODIS Unit planned the first in-person meeting of all DNA Technical Leaders, to be held jointly with CODIS Administrators, at the CODIS Conference in November 2013 in Norman, Oklahoma. An email invitation to participate in the MIX13 interlaboratory study was sent to all U.S. and Canadian Technical Leaders in July 2013 announcing the purpose and goals of the study. A total of 108 laboratories supplied results for analysis and comparison before the study closed, including labs from 46 states, three Federal laboratories, and three Canadian laboratories (Supplemental Table S2). Thirty-four of these laboratories participated previously in the MIX05 study. Again, not every laboratory supplied information for every case.

2.2.2. Sample selection

As with MIX05, the samples selected for Cases 1, 3, 4, and 5 were part of the NIST population data collection and the genotypes for 660 males and 40 females that had previously been published for the Identifiler kit (Butler et al. [26] also Hill et al. [27]). Again, *Virtual MixtureMaker* was used to explore possible allele combinations for these synthetic mixtures. One mixture (Case 2) used an electropherogram generated by Boston University (BU) researchers Catherine Grgicak and Robin Cotton. Their set of mixture profiles are available at http://www.bu.edu/dnamixtures/.

2.2.3. Data generation

Case scenarios representing 2-, 3-, and 4-person mixtures were generated to represent both straight-forward and complex examples. DNA extracts from the four NIST-generated examples (Cases 1, 3, 4, and 5) were quantified separately in triplicate using Quantifiler (Thermo Fisher Scientific) following the manufacturer guidelines, and the average quantity of each sample was used to create the target mixture ratios (Supplemental Table S3). Profiles were amplified with either PowerPlex 16HS (Promega Corporation) or Identifiler Plus (ThermoFisher Scientific) at full reaction volumes and following the recommended amplification cycles by each manufacturer. PCR amplification was performed on the ABI 9700 thermal cycler and STR alleles were separated on an ABI 3130xl Genetic Analyzer using POP-4 polymer and a 36-cm array. The BU sample (Case 2) was amplified with either PP16HS or Identifiler (see http://www.bu.edu/dnamixtures/pages/help/introduction/).

2.2.4. Sample details

Five mock cases were created to represent types of casework commonly encountered. Genomic DNA samples with specific allele combinations ("evidence") were mixed as noted in Supplemental Table S3. No DNA profiles were used more than once to create the various mixtures in MIX13 (and none were repeated from the MIX05 study). Each case consisted of an electropherogram file (.fsa format) from Identifiler or PowerPlex 16HS and typed reference profiles from individuals described as victims, consensual partners, or persons of interest (POIs). The mixed profile electropherograms were made available as both Identifiler Plus (or Identifiler for case 2) or PowerPlex 16 HS profiles. Brief case scenarios were provided to put each case into perspective.

2.2.5. Scenarios provided

2.2.5.1. MIX13 case 1. A female meets a male acquaintance at a bar after work and they return to her apartment for a nightcap. She recalls the drink tasting funny and then wakes up 14 h later after a co-worker has her landlord open her apartment. She is confident that she did not have consensual sex and was probably drugged. She reports the incident to the police and goes to the hospital for an examination. Evidence is the sperm fraction from a vaginal swab. The accused male (Reference 1A) gives a buccal swab for comparison.

2.2.5.2. MIX13 case 2. A convenience store employee was murdered after a robbery. Video from the store's security camera show two perpetrators enter the store, with one individual holding the gun on the victim and the other empties the cash register and takes two plastic bags full of cigarettes from behind the counter. Before leaving the store, the employee triggers an alarm, and is shot three times with the handgun. The police find the handgun in the parking lot near the entrance of the store, apparently dropped by the shooter during the escape. Ballistics comparison of bullets fired from the weapon confirms the gun was used to commit the homicide. Evidence is a DNA profile generated from a swab from the grip of the recovered handgun. DNA has been collected from four suspects (References 2A, 2B, 2C, and 2D) identified during the investigation.

2.2.5.3. MIX13 case 3. The female victim and her boyfriend host a party for a small group of friends on a recent Saturday. The victim had too much alcohol to drink and decided to go to bed around midnight. At some point in the middle of the night, she awoke with someone on top of her performing intercourse. She tried to resist and scream, but wasn't able to stop the assault and soon blacked out. She awoke at 5 a.m. and found her boyfriend passed out on the couch, unaware of what had happened. Evidence is a DNA profile generated from the sperm fraction from a vaginal swab collected from the victim. The police obtained DNA samples from the two men remaining in the house according to the boyfriend before he remembers passing out: his brother (Reference 3A) and one other unrelated male (Reference 3B). Both men claimed that they left together at 2 a.m. after the boyfriend passed out on the couch. Neither suspect locked the door before they left the house. About 12 h prior to the assault, the victim and her boyfriend confirmed that they had consensual sex.

2.2.5.4. MIX13 case 4. A female waiting at a bus stop in the late evening is attacked from behind and pushed to the ground. A motorist driving by witnesses the attack, pulls his car over, and runs to her aid. As the Good Samaritan comes upon the scene, the perpetrator bites the victim on the back of her neck before running away. Evidence is a DNA profile generated from saliva found when swabbing the bite mark on the victim. The motorist is able to give a good description of the perpetrator and a few days later, the police arrest a suspect (Reference 4A). He is positively identified in a police lineup by the witness.

2.2.5.5. MIX13 case 5. Several gang-related robberies have targeted multiple banks in the city. The robberies have typically involved two or three perpetrators. A ski mask was recovered in a trash can one block away from the latest bank robbery and is submitted for DNA testing. Evidence is a DNA profile developed from a ski mask recovered near a bank robbery scene. A confidential informant has implicated two suspects (References 5A and 5B) in at least three of the armed robberies. Police have obtained buccal swab references from the two suspects identified from the informant, and another known accomplice of the suspects (Reference 5C).

2.2.6. Data supplied

Three webcasts were held for DNA Technical Leaders in late-July 2013 (92 individuals signed up and 81 attended one of the three events) to provide more details on the purpose and goals of the study, the kits used to generate the data, and a planned timeline for the study. Initially, participating laboratories were provided electronic data from the five case examples along with references and case scenarios using a secured ftp site in mid-August. A webpage was also created so that the MIX13 data were available for anyone to download for training or study purposes (https://strbase.nist.gov/interlab/MIX13.htm).[3]

2.2.7. Information requested for study

MIX13 participants were requested to treat the ".fsa" data as though they were actual casework samples. Laboratories were asked to produce a report of their analysis including any statistical evaluation of the data. A table of alleles/genotypes used to generate the statistical results was requested. Many laboratories provided a printed output from PopStats, which is the statistical package supplied by the FBI Laboratory as part of the CODIS software. In addition, participants were asked to analyze cases 1, 2, 3, and 5 using their own standard operating procedures or the recommended analytical thresholds (AT) and stochastic thresholds (ST) provided by the study organizers. For case 4, laboratories were instructed to use an AT of 50 relative fluorescence units (RFUs) and a ST of 150 RFUs. When reporting results, study participants were invited to provide information on their AT and ST values, resolved locus genotypes, calculated match statistics, the allele frequency database used, and a copy of laboratory interpretation protocols. Although only a portion of the MIX13 participants provided a copy of their protocols, this information was helpful in some cases to understand why particular approaches were taken that may have led to differences among laboratories. Although it would be ideal to use the AT and ST thresholds from the laboratory generating the data (i.e., NIST or BU), most interlaboratory participants chose to use their own threshold values for interpretation, which contributed to some of the variation observed.

[3] In 2013, the URL was http://www.cstl.nist.gov/strbase/interlab/MIX13.htm.

2.2.8. Results collation and summary

Results received were collated in an Excel spreadsheet and laboratory participants were deidentified through numerical assignment (13–1 to 13–108). Information compiled included STR kit data used, analytical and stochastic thresholds applied, inclusion or exclusion of the person of interest, alleles determined from the mixture, population allele frequencies used, theta value used for population substructure, and the results of any statistical calculation performed (see Supplemental File S2).

3. Results and discussion

Interlaboratory studies involve receiving and trying to organize a great deal of information. While recognizing that we cannot describe every aspect of our data in a published manuscript, we have chosen to highlight some observations. The spreadsheets used to summarize information received from laboratories are included in the Supplemental Files, should any readers wish to inspect those further. Both inter-laboratory and intra-laboratory results were received for MIX05 and MIX13 and will be discussed separately.

3.1. Interlaboratory results

3.1.1. Summary of MIX05 responses

Of the 106 laboratories that expressed initial interest in participating in MIX05, 94 laboratories formally enrolled and received data, and 69 labs returned results. Of those, 50 assigned allele calls, 39 provided estimates of mixture ratios, and 29 included statistical reports. Generally, laboratories returned what they were comfortable with sharing at the time. For example, an estimation of mixture ratios was requested and several responded that their laboratory typically does attempt to calculate mixture ratios in casework. Only two laboratories explicitly stated that their protocols at the time allowed estimation of mixture component ratios.

An important lesson learned was that it is important to provide scenarios with the data. In some situations, participants did not analyze the provided data because no scenarios were included, such as “these mixture profiles were from intimate sexual assault samples,” which may have influenced some laboratories to subtract out the female victim genotypes given that victim’s DNA would be expected on intimate items.

Many analysts were uncomfortable analyzing data that was not collected by their laboratory under their protocols. In fact, some laboratories refused to provide responses because they claimed that the provided electropherograms were inadequate for their interpretation protocol. For example, the response received from one MIX05 participant stated: “It was determined the results from the four cases provided would not be interpreted by our laboratory. Several samples are overloaded (too much DNA was amplified) and would require re-injection for a shorter period of time or re-amplification with less DNA. In case examples 1 and 2 all samples are overloaded and the characteristic artifacts from this are evident. In case 3 the evidence sample exhibits-A [incomplete adenylation] and pull up and the victim sample is overloaded. In case 4 the evidence sample is interpretable but the victim sample is overloaded.” Unfortunately, the analyst providing this response did not state which STR kit data was examined to make this determination.

From the six different STR kits provided in MIX05, most results were returned using Profiler Plus and COfiler (to obtain the 13 CODIS core STR loci required at the time), followed by PowerPlex 16, PowerPlex 16 BIO, Identifiler, SGM Plus, or some combination of multiple kits.

A source of some variability in the responses was the use of different STR kit data due to peak height differences that naturally arise even when generating profiles from aliquots of the same DNA mixture solution. Fig. 1 illustrates peak height variation seen in the MIX05 Case 1 mixture at the D3S1358 locus, which is common to the five STR kits

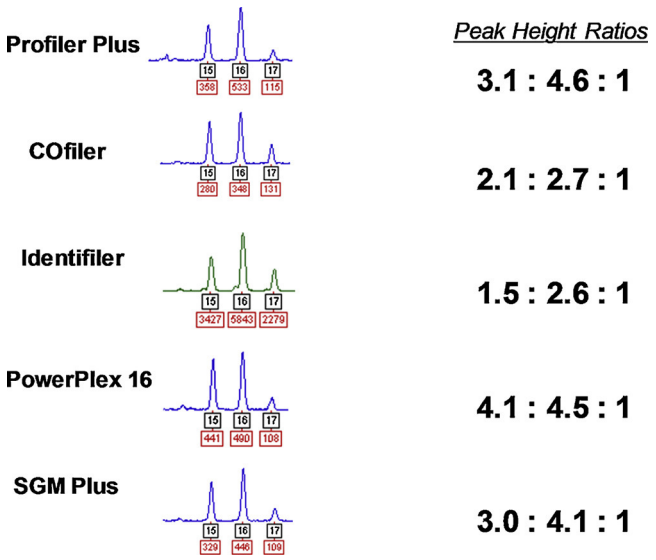

Fig. 1. MIX05 Case 1 mixture results at the D3S1358 locus from five different STR kits, which illustrates variation that can occur when amplifying the same DNA mixture. Values under each peak correspond to allele calls and peak heights. Relative peak height ratios were calculated by dividing the peak height of the lowest peak (allele 17) into the other peak heights.

shown. This particular mixture was created by combining three parts of a sample possessing a 15,16 genotype and 1 part of another sample with a 16,17 genotype. Thus, under ideal PCR amplification conditions, the relative peak height ratio should have been 3:4:1 for the 15, 16, and 17 alleles, respectively. The SGM Plus data of 3.0:4.1:1 in this set of test samples was closest to the expected peak height ratios for a 3:1 mixture of the DNA samples examined.

Samples used, mixture ratios explored, and degree of allele overlap investigated are described in Table S3. For example, with MIX05 Case 1, a NIST DNA sample with 26 alleles in Identifiler was mixed in a 3:1 ratio with another sample having 26 alleles in Identifiler to create a mixture with a moderate degree of allele overlap (39 total alleles and 29 unique alleles) containing 2 loci exhibiting four alleles, 5 loci showing three alleles, 6 loci with two alleles, and 2 loci with only a single allele (see “MIX05 selected samples” tab in Supplemental File S1 spreadsheet). With MIX05 Case 2, there was less allele overlap (55 total alleles with 53 unique alleles) containing 10 loci exhibiting four alleles, 5 loci exhibiting three alleles, and no loci exhibiting a high degree of allele sharing with two or one alleles using a NIST DNA sample with 31 alleles in Identifiler mixed in a 1:3 ratio with another NIST sample having 29 alleles in Identifiler). For Case 1, participants were asked to deduce the minor component genotypes when there was a substantial amount of allele sharing. For Case 2, participants were requested to decipher the major component genotypes when most loci had genotypes that were fully resolvable (i.e., four allele loci), which is a much easier task than that which was explored with Case 1.

Accurate genotype determination was achieved by almost all laboratories when a clear major contributor was being deduced (Case 2). However, mistakes in deducing genotypes increased when seeking a minor contributor especially in mixtures where loci exhibited significant allele overlap (Case 1) or where the contributor ratios were more balanced (Case 3). And when a low-level minor component was being sought (Case 4), often only sporadic alleles and genotypes were deduced by participants (see Supplemental File S1).

A review of responses for MIX05 Case 1 is instructive. With the FGA locus, where the perpetrator genotype is 20,22 – and easily resolved from the victim genotype of 19,21 – all reported genotypes were deduced correctly. However, with D3S1358, where there is a shared allele 16 between the victim “15,16” and the perpetrator “16,17” which

Table 3
MIX05 case 1 (1:3 mixture) variation in statistics observed after deducing genotypes present in the minor contributor. The seven laboratories were examining the same Profiler Plus and COfiler data. All laboratories (except 05–9) were reported as being accredited by ASCLD/LAB in 2005.

Lab ID	Statistical Approach Utilized	Statistical Value (U.S. Caucasian)	Detection Threshold	Solved Loci Listed?
05–90	Random match probability calculation from deduced minor contributor profile	1.18×10^{15}	75 RFUs	Results correct for all 13 STR loci
05–34	Random match probability calculation from deduced minor contributor profile	2.40×10^{11}	Not provided	8 STR loci, 2 partial, 3 inconclusive
05–33	Details not provided (likely CPI)	2.94×10^{8}	75 RFUs	No deduced genotypes reported
05–6	Used selected loci and summed all possible genotypes for loci not completely deduced	4×10^{7}	Not provided	3 STR loci, 6 partial, 4 inconclusive
05–9	Used 1/CPI	4.14×10^{7}	100 RFUs	No deduced genotypes reported
05–79	Details not provided (likely CPI)	9.30×10^{5}	150 RFUs	2 STR loci, 5 partial, 6 inconclusive
05–16	Details not provided (likely CPI)	4.35×10^{5}	Not provided	No deduced genotypes reported

makes it harder to unambiguously decipher the minor component, many laboratories simply designated the foreign, obligate allele "17" in attempting to deduce the minor contributor's genotype. Not designating a full genotype at a locus impacts the statistical weight of the overall profile being deduced.

Table 3 compares MIX05 Case 1 statistical results from seven laboratories calculating CPI or RMP, using the Profiler Plus and COfiler data (13 STR loci) with Caucasian allele frequencies. These laboratories, looking at the same data, reported values ranging over almost ten orders of magnitude from 1.18×10^{15} down to 434,600 or 4.35×10^{5} when attempting to decipher the minor component of this two-person mixture. However, in a different situation, when reporting statistics on the major contributor in MIX05 Case 2, these same two laboratories reported much more consistent results of 3.36×10^{20} and 4.08×10^{20} (see Supplemental File S1). Thus, variation observed differs depending on the type and complexity of the mixture being evaluated as well as whether the reference profile being compared is a major or a minor contributor to the mixture.

To better understand reasons for the large degree of variation in reported statistical responses, detection thresholds used by each laboratory were examined. Unfortunately, not all laboratories provided details on the protocols they used in the study. With some mixture data, use of a higher detection (analytical) threshold and stochastic (interpretation) threshold could lead to reporting results from fewer loci and thus a smaller reported match probability. For example, one laboratory used a detection threshold of 75 RFUs and reported accurate genotypes for all 13 STR loci under consideration while another laboratory used a higher detection threshold of 150 RFUs, only fully deduced genotypes at two loci and reported partial results at five loci and inconclusive results at 6 loci. With fewer genotypes in a DNA mixture being deduced when a higher detection threshold is utilized, there are fewer points of comparison between the mixture and the reference profile(s) and thus the resulting statistic will be lower.

Some possible reasons for variability in the reported statistics with the MIX05 mixtures include use of (1) different types of calculations including random match probabilities (RMP) versus combined probability of inclusion (CPI), (2) different combinations of loci included in the calculations due to different thresholds that may have been applied by the laboratories, (3) different allele frequency population databases (although most laboratories at the time used PopStats), (4) improper use of the victim's profile (e.g., major component in Case 1) to report statistics for the case, which was done by several laboratories, and (5) the possibility of an analyst missing an allele call (e.g., designating an allele as an artifact or vice versa) or miscalculating a statistic particularly if calculations were performed manually and not technically reviewed.

Protocol specificity and training of analysts may play a role in accurate mixture interpretation. One protocol possessed a set of specific, detailed mixture interpretation guidelines with worked examples and a detailed flowchart whereas the protocol received from another laboratory was fairly scant beginning with the following text: "…mixture interpretation is not always straightforward. Analysts must depend on their knowledge and experience…"

In summary, MIX05 participants were highly accurate in deducing genotypes and fairly consistent in reporting statistics for a major contributor in a two-person mixture when there was very little overlap in genotypes present from the other contributor (Case 2). Accurately deducing minor contributor genotypes appeared to be more challenging and led to a larger spread in reported statistical values (e.g., Table 3) for the other cases examined (Case 1, Case 3, and Case 4).

3.1.2. Summary of MIX13 responses

A total of 108 laboratories returned some information for the MIX13 study. Table 4 contains a summary count of participant laboratories and

Table 4
Summary results from 108 laboratories participating in the MIX13 interlaboratory study. For each of the five mixture cases, the number of laboratories providing conclusions for each person of interest (POI) are listed. False inclusions are shown in bold font (1 for reference 2D, 1 for reference 3B, and 74 for reference 5C).

Mixture	Person of Interest (POI) Considered	Included in Mixture	Approach Used When Including POI			Types of Non-Inclusions		
			CPI	LR	mRMP	Excluded	Inconclusive	Not Reported
Case 1	Reference 1A	Y	22	16	70	–	–	–
Case 2	Reference 2A	Y	41	3	28	–	36	–
	Reference 2B	Y	36	2	1	14	55	–
	Reference 2C	Y	12	2	1	32	61	–
	Reference 2D	**N**	**1**	–	–	73	33	1
Case 3	Reference 3A	Y	37	9	15	11	35	1
	Reference 3B	**N**	**1**	–	–	90	14	3
Case 4	Reference 4A	Y	25	20	61	–	1	1
Case 5	Reference 5A	Y	76	2	4	2	24	–
	Reference 5B	Y	77	2	4	2	23	–
	Reference 5C	**N**	**70**	–	**4**	7	27	–

statistical approaches applied in the five MIX13 cases scenarios and if the provided reference profile(s) was included, excluded, designated as inconclusive, or not reported. Note that three reference profiles (2D, 3B, and 5C) were not included in the mixtures to which they were compared (see indication in the column designated "*Included in Mixture*"). The number of laboratories that excluded the particular POI or determined the mixture to be inconclusive are also provided. The final column in Table 4, "Not Reported" refers to the number of laboratories that did not report a statistic or conclusion. For example, one laboratory did not give a result (e.g. inclusion with a statistic, inconclusive, or exclusion) for Reference 2D in MIX13 Case 2 (Table 4).

Most participants in the MIX13 study used the FBI allele frequency databases (82/108 = 76%), with another 9% (10/108) using a combination of the FBI allele frequencies in addition to local or state population databases (e.g., with indigenous or regional populations). Twelve laboratories (11%) exclusively use a state or regional database and three labs used the NIST allele frequency data [26]. One laboratory did not report the population database used for their statistical calculations. Most laboratories used PopStats software (65%) provided within CODIS to calculate their statistical results for MIX13. Approximately 30% of the laboratories reported the use of an in-house spreadsheet program to calculate statistics. Four laboratories used a commercially available software program that calculated statistics after the resolution of the mixture. Two laboratories did not indicate their statistical calculation program.

There was a range of statistical values provided by MIX13 participants using multiple statistical approaches: combined probability of inclusion (CPI), likelihood ratio (LR), or a random match probability after deducing genotypes from the mixture (listed as mRMP or RMP). The strength of the evidence is presented as the $\log_{10}$ of the statistic on the y-axis. For example, a statistic of 10^6 (1 million) gives a $\log_{10}$ number of 6. For CPI and mRMP statistics, the results are reported as the $\log_{10}$ of 1/CPI or 1/mRMP in order to plot the data on the same scale. Fig. 2 illustrates the ranges reported for MIX13 Case 1. Reported statistical ranges for the other MIX13 cases are available in Supplemental Figs. S1–S4. Results obtained for each of the five MIX13 cases are explored in more detail below.

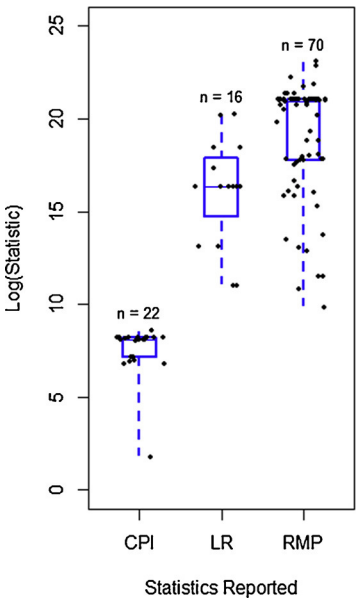

Fig. 2. Variation in reported results for MIX13 case 1 (2-person, 1:1 ratio) with reference 1A across statistical approaches of combined probability of inclusion (CPI), likelihood ratio (LR), and random match probability (RMP) using Caucasian allele frequencies. The vertical axis is in powers of 10 to reflect orders of magnitude as a $\log_{10}$ (statistic).

3.1.2.1. MIX13 case 1 (2-person mixture, 1:1 ratio). The question being explored with this case scenario, which involved a two-person mixture at roughly a 1:1 balanced ratio of contributors, was whether participants would attempt to infer the genotype of the unknown contributor, or would they simply use a CPI statistic without attempting to deduce any possible genotypes? All participants correctly included the reference profile "1A" and provided a statistic (Table 4). Most of the laboratories inferred the genotype of the unknown contributor and provided either mRMP or LR statistics. However, a wide range of variation between methods was observed in the statistical values reported (Fig. 2). Some of this variation between MIX13 participants can be explained using different interpretation thresholds or population data allele frequencies or a decision to not use some loci in the statistical calculations (e.g., Penta D and Penta E). Note that results obtained with Identifiler Plus and PowerPlex 16 HS data sets are grouped together and only distinguished by statistical approach.

3.1.2.2. MIX13 case 2 (3-person mixture, 6:1.5:1 ratio). The question being explored with this case scenario, which involved a low-level, three-person mixture from touch evidence with roughly a 6:1.5:1 ratio of contributors, was whether participants would attempt to solve this mixture, or would it or its minor components be treated as too complex for interpretation due to the potential of allele drop-out? Four reference profiles were provided for comparison – 2A, 2B, 2C, and 2D (Table 4). DNA from individuals 2A, 2B, and 2C were used to create the mixture while the reference 2D profile was a decoy and not included in the mixture. The total quantity of DNA amplified for Case 2 was 300 pg of DNA with sample 2A at 6 parts (approximately 210 pg of the mixture), sample 2B at 1.5 parts (approximately 55 pg) and sample 2C at 1 part (approximately 35 pg). Most of the laboratories included the reference 2A profile as a major contributor in the mixture. CPI was the most commonly used statistic in Case 2 (Table 4; Fig. S1). Fewer participants were willing to include references 2B and 2C as minor contributors in the mixture opting instead for a report of "inconclusive" on these profiles. The non-contributor, reference 2D, was falsely included in the mixture by one laboratory using a low CPI statistic (1 in 2.8 in the U.S. Caucasian population). Most of the remaining laboratories either excluded or gave an inconclusive result. One laboratory did not provide a result (inclusion, exclusion, or inconclusive) in their report for 2D.

3.1.2.3. MIX13 case 3 (3-person mixture, 7:2:1 ratio). This sexual assault case scenario, which involved another three-person mixture with the possibility of allele dropout from its minor contributor (≈ 100 pg), also contained a potential relative to explore how participants would handle someone that was not an unrelated individual. Several of the laboratories in their responses recognized the issue of a related person in the mixture and responded with something like "due to the relatedness of the exemplars submitted for comparison, a statistical analysis cannot be provided at this time." One participant falsely included reference 3B with a CPI statistic of 1 in 2.2 in the U.S. Caucasian population. Most laboratories (83%) correctly excluded 3B with the remaining responses reporting an inconclusive result (13%) or no statistic (3%) (Table 4).

3.1.2.4. MIX13 Case 4 (2-person mixture, 3.5:1 ratio). This case scenario was composed of two contributors with a mixture ratio of approximately 3.5 to 1. It was designed to determine if laboratories would choose to deconvolute the mixture since the mixture ratio is close to the limit of 4:1 that some laboratories use to distinguish a major from a minor contributor (see Ref. [5]). Participants were also requested to use a NIST-provided analytical threshold of 50 RFUs and a stochastic threshold of 150 RFUs to determine if similar statistical results might be obtained to avoid a range of variation observed like in Table 3. One laboratory used a probabilistic genotyping software which

obviates the need for a stochastic threshold. Another laboratory provided an inconclusive result with this case since the 150 RFU threshold was below their laboratory-developed interpretation protocol, and therefore they would not have made an inclusion or exclusion for this example (Table 4). All other laboratories included 4A and provided statistical weight to their conclusion. Most of the laboratories (≈ 75%) inferred the genotype of the unknown contributor and provided either mRMP or LR statistics while the remaining laboratories (≈ 23%) provided CPI statistics. One laboratory indicated in their report that 4A would be included in the mixture, but did not provide a statistic and was counted as "not reported" (Table 4).

For the 25 laboratories reporting CPI results (Fig. S3), most of the statistics were grouped around the median log CPI of 4.2. The two highest outliers (log CPI of 5.9 and 6.5) failed to drop most or all of the loci with alleles below the ST and included these loci as part of the statistic. The other outlier laboratory (log CPI of 5.4) included loci, for example, where the two minor alleles are both below the ST, but assumed a two-person mixture and included all four alleles in the CPI calculation as recommended [28]. The one laboratory that provided an outlier statistic with a log CPI of 1.8 did not provide any details on how this number was determined.

Despite use of the same AT and ST, there was a broad range of variability in results reported from laboratories that deconvolved the MIX13 Case 4 profile to infer the minor contributor (Fig. S3). Since differing thresholds cannot explain this variation, we first note that some laboratories used a restricted versus unrestricted approach (see Ref. [22]) that explains some of the differences. For example, at the FGA locus, there are two major alleles and two minor alleles (both minor alleles are below 150 RFU). Most laboratories (using either LR or mRMP) restricted the minor contributor to having the two minor alleles. Some laboratories considered all possible genotypes (unrestricted) minus any homozygous combinations. Neither approach is incorrect, but this difference in interpretation alone can generate differences in laboratories using the same statistical approach.

Additional variation can be explained through the assumptions used by MIX13 participants according to their protocols. For example, we will use the D16S539 locus in the Identifiler multiplex to present the constellation of results observed in this study (see Supplemental File S3). The victims's genotype at this locus is "12,12." The minor contributor "11,12" shares an allele with the victim's "12" and an allele in the stutter position "11." If a 10% stutter percentage is assumed, the 11 allele (at 233 RFU) is higher than expected (163 RFU), suggesting that the "11" allele is partially from the unknown contributor and/or partially stutter artifact.

Using the results from the 61 mRMP reports (those that provided data), we observed the following approaches to uncertainty at D16S539:

(1) **Drop the locus** (19% of the results). If, for example, 163 RFU of the 11 allele is attributed to stutter, then 70 RFU of the remaining allele belongs to the unknown minor contributor. These laboratories did not therefore use this locus in their statistical calculation because allele drop-out is a possibility. As described previously [29], dropping a locus can be anticonservative in some cases.
(2) **Use the "2p" rule** (38% of the results). Using the same logic of the previous example, if 70 RFU of allele 11 belongs to the minor unknown contributor, then this peak is between the AT and ST, so using the "2p" rule would be an accepted way to use this locus. This was the most popular strategy used. We note that it has been demonstrated that the 2p rule is not always conservative in some situations [30].
(3) **Infer "all" possible genotypes for the minor contributor** (8%). In this example, laboratories considered the possible genotypes of the minor contributor to be either "11,12" or "11,11" and then summed the 2pq and p^2 for each genotype, respectively. The implicit assumption here is that stutter is not necessarily assumed to be 10% and the 11 allele may very well be above 150 RFU, so drop-out considerations of approaches 1 and 2 are not considered. We noted that most labs reported only the "11,12" or "11,11" genotypes of the minor contributor. It is also possible if one considers all possible genotypes that the minor contributor could also be "12,12" like the major contributor, and the "11" allele is simply elevated stutter from both homozygous "12,12" contributors. This was the least used approach.
(4) **Infer only the "11,12" genotype for the minor contributor** (35%). Behind the laboratories that used the "2p" rule, this was the second-most popular strategy to infer the genotype of the unknown minor contributor. Laboratories applying this strategy could eliminate a "11,11" genotype possibility for the minor contributor by considering the mixture ratio in their interpretation. If, for example, the minor contributor was truly "11,11" – the mixture ratio of major "12,12" to minor "11,11" would be approximately 7 to 1 (1635 RFU from the 12 allele/233 RFU of the "11" allele = 7.01 to 1). This is beyond the estimated mixture ratio of 3.5 to 1 determined across the profile. If, one considers that the minor contributor is "11,12" – then 233 RFU of the 12 allele could belong to the minor contributor. This would leave 1635 – 233 = 1402 RFU to the major contributor and the mixture ratio of major to minor would be 1402/(233 + 233) = 3 to 1 mixture ratio which is very close to the estimated 3.5 to 1 ratio across the profile.

Interestingly, of the 25 laboratories that reported CPI results and provided the specific loci used in calculating their statistics, none considered the uncertainty of the stutter allele at D16S539 – that is, the minor allelic contribution may be below 150 RFU and drop-out would be possible at this locus (making it ineligible for CPI statistical calculations).

It is interesting to assess how different laboratories handle peaks where there is uncertainty – minor alleles in stutter positions, alleles between the AT and ST, and so forth – that can lead to differing results and a wide range of variation when using the same thresholds for interpretation. At least seven laboratories explicitly used a "source attribution" statement that capped the statistic at (typically) 1 in a few billion rather than report statistics of trillions or quadrillions.

Additional concerns were also observed with MIX13 Case 4. One laboratory used a strategy of performing CPI on some of the loci and then mRMP/2p on other loci in the same profile. This is against recommendation 4.6.2 of the 2010 SWGDAM autosomal STR interpretation guidelines [22] and is still prohibited by the 2017 guidelines (Section 3.2.5.1; [23]) because different assumptions are being made within the same profile. It was also concerning that one mRMP laboratory determined the mRMP for the minor contributor in Case 4 to be over 1 in 400 quintillion (log mRMP of 20.6) which is nearly the RMP of a single-source profile for this individual and may be an indication of "suspect-driven" interpretation where the statistics are calculated on the reference profile and not necessarily from the interpretation of the evidence profile. Problems with suspect-driven interpretation approaches have also been noted by others (see Refs. [31,32]).

3.1.2.5. MIX13 case 5 (4-person mixture, 1:1:1:1 ratio). MIX13 Case 5 involved a DNA profile developed from a discarded ski mask where the prepared mixture contained four contributors in roughly equal amounts (i.e., 1:1:1:1). However, this mixture was designed to contain no more than four alleles at any locus to appear as a two-person mixture if maximum allele count was used to infer the number of contributors [33]. In addition, only two (5A, 5B) of the four contributors were provided as reference samples. The profiles of the remaining two individuals in the mixture were not provided. Instead, a contrived profile of a reference not in the mixture (5C) was provided for comparison. The purpose of MIX13 Case 5 was to explore whether

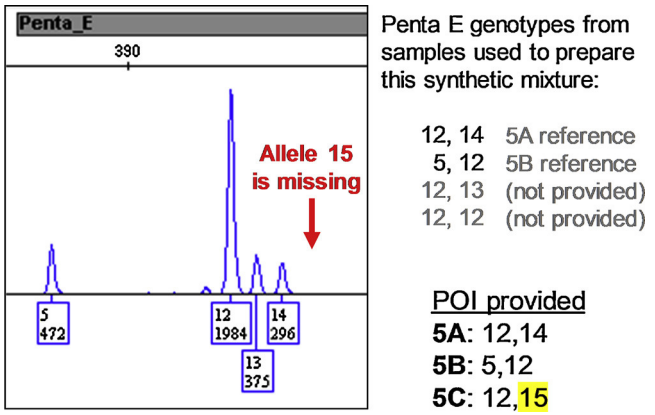

Fig. 3. Penta E locus results using PowerPlex 16 HS data from MIX13 case 5 (4-person mixture, 1:1:1:1 ratio). Allele 15, which is present in the person of interest 5C reference profile, is not present in the mixture. An extra allele 13 is present in the profile (suggesting perhaps an unknown contributor is in the mixture) and the "15" allele (obligate to the 5C reference profile) is absent suggesting a potential drop-out event at Penta E if reference 5C is a true contributor. This dilemma was created to evaluate how many laboratories would exclude reference 5C for a single discordance or would they assume allele drop-out and still include 5C.

laboratories would consider this mixture too complex to interpret, and whether they would include the non-contributing reference profile (5C) and provide a matching statistic. The exact scenario of genotypes provided is highly improbable as has been noted [24].

Of 108 laboratories contributing to the MIX13 study, a total of 74 (69%) included "suspect" 5C (along with suspects 5A and 5B) (Table 4) and provided CPI statistics in a Caucasian population ranging from 1 in 9 to 1 in 344,000. Another 23 laboratories (21%) declared the entire mixture "inconclusive" and did not provide any comparisons or statistics for 5A, 5B, or 5C. Four additional laboratories (4%) declared 5C to be "inconclusive" but included 5A and 5B. These laboratories used PP16HS data for 5C, which possessed a single allele at Penta E that was discordant (Fig. 3).

Seven laboratories (6%) correctly excluded 5C, but for a variety of reasons. Four of the laboratories mentioned the missing allele 15 at Penta E with PP16HS data. One laboratory (using Identifiler Plus data) assumed major and minor contributors and noted that the suspects did not fit the needed combination to produce the mixture data reported. Another laboratory (using Identifiler Plus data) turned in results with detailed manual genotype assessments and noted that 5C would not fit and therefore should be excluded. The single use of a probabilistic genotyping program in this study (True Allele) resulted in a negative log likelihood ratio and therefore correctly reported that evidence did not support 5C being in the MIX13 Case 5 mixture profile under an assumption of three or four contributors. Developers of the TrueAllele probabilistic genotyping software program submitted results that correctly excluded 5C, but they were not included in the 108-laboratory tally as they were not considered a forensic "laboratory" for the purposes of this study. Developers of the Lab Retriever program also submitted results for MIX13 cases 1-4, but not case 5 because their software at the time was limited to a maximum of three contributors. Following completion of the study, analysis with other continuous probabilistic genotyping software programs (e.g., STRmix and DNA-View Mixture Solution) obtained similar results with "excluding" 5C (personal communication).

The range of results for individuals 5A and 5B for log CPI and log mRMP statistics were similar to those of 5C (Fig. S4). Most often, a laboratory that included 5C also included 5A and 5B. Since the same loci were used for the CPI calculation, the statistics for 5A, 5B, and 5C were almost always the same. For reference 5C, 27 laboratories (25%) reported inconclusive results for case 5. Again, most laboratories that reported inconclusive results for 5C also reported inconclusive results for 5A and 5B.

The Case 5 mixture was a very challenging test for interpretation. The mixture was made from an equal contribution of four unrelated individuals (i.e., in a 1:1:1:1 ratio) from the NIST population dataset. We selected the four individuals so that in the mixture no more than four alleles would be present at any locus (including the supplemental loci: Penta D, Penta E, D2S1338, and D19S433). On the surface, simply by allele count, the mixture would appear to be a two-person mixture. However, no laboratories in this study determined Case 5 to be a two-person mixture. Based upon the variability in peak height ratios and potential mixture ratios, labs typically reported that the mixture may have been from more than two individuals (see Supplemental File S2).

Because we designed the mixture to have a great deal of allele sharing among 17 loci (the 13 CODIS and the four supplemental loci from PP16 and Identifiler kits), we were unable to find a fifth unrelated person for comparison. Therefore, we constructed the 5C reference to share alleles among the four individuals in the mixture. We felt it was important to keep the same CODIS loci consistent for both PP16 and Identifiler kits since (a) this was done in MIX13 Case 1 through Case 4, and (b) we wanted to allow one-to-one comparisons at each locus between the kits. Interestingly, if we only made "four-person mixtures that looked like a two-person mixture" separately for each kit (i.e., only using 15 loci instead of 17 loci) we would find several "fifth" unrelated persons that could be compared, so it is unlikely that the contrived 5C reference versus a sample from a real person would have made a difference in the final analysis.

Only four laboratories attempted to deconvolve the Case 5 mixture and apply mRMP or LR statistics to their conclusions (Table 4). We focus here on the results from the non-contributor, 5C (Fig. 4S). We found that 70 of the 108 participating laboratories (65%) determined 5C was included in the mixture and provided a CPI statistic to the weight of the evidence (Table 4). Most of the results for the log CPI clustered around the median statistic of 5.3.

Of the seven laboratories that excluded reference 5C, there was a substantial difference in the decision to exclude based upon the STR multiplex kit used. In MIX13, 90 laboratories used the Identifiler Plus data and 18 laboratories used the PowerPlex 16 HS data. Only 3 of 90 Identifiler laboratories excluded 5C (3%) whereas 4 of the 18 PP16 laboratories (22%) made an exclusion. These PP16 laboratories excluded based upon the presence of a non-matching allele at a single locus – Penta E.

Four alleles are present at the Penta E locus with the Case 5 mixture (Fig. 3): One major '12' allele and three minor alleles at 5, 13 and 14 with all alleles above the suggested ST of 150 RFUs. The genotypes of the references provided for Penta E were 12, 14 (5A); 5, 12 (5B) and 12, 15 (5C). The challenge at this locus was that none of the given references had the '13' allele (suggesting perhaps an unknown contributor is in the mixture) and the '15' allele at 5C is absent from the profile (suggesting a potential drop-out event at Penta E if reference 5C is a true contributor). Of the18 MIX13 participants who utilized PowerPlex 16 HS data, six included the non-contributor reference profile 5C and reported a CPI statistic ranging from 1 in 12.5 to 1 in 150 thousand, eight reported an "inconclusive" results, and only four correctly excluded 5C.

3.2. Intra-laboratory results

Significant variation reported between analysts within a laboratory may reflect lack of training or variation in understanding mixture interpretation principles. Observed variation in reported results may also reflect a lack of protocol specificity or sufficiency. The following sections describe intra-laboratory (among analyst) variation reported with the MIX05 and MIX13 studies.

3.2.1. Summary of MIX05 responses

Supplemental Table S4 shows variation observed with deducing genotypes in the MIX05 cases with seven analysts from one laboratory, which are coded "a" through "g" in Supplemental File S1. With MIX05 Case 1, the first four analysts deduced genotypes of the minor contributor and obtained the correct results at 10 of 14 loci examined: D3S1358, FGA, amelogenin, D21S11, D18S51, D5S818, D7S820, D16S539, TPOX, and CSF1PO. While Analyst "b" deduced all the minor component genotypes correctly, Analyst "a" did not report allele 17 in vWA, reported an extra "14" at D8S1179, an extra "11" at D13S317, and an extra "8" at TH01, Analyst "d" reported an extra "11" at D13S317 and an extra "8" at TH01, and Analyst "c" reported an extra "8" at TH01. In their reports, the remaining three analysts (Analysts "e", "f", and "g") provided various reasons for not deducing the minor contributor genotypes including that there were no four allele loci in the COfiler data. For example, Analyst "f" shared that she did not feel comfortable with the reliability of the outcomes, and therefore decided not to attempt a separation of the potential component genotypes. It is interesting that different levels of comfort were reported when analyzing the same data with the same protocol.

With MIX05 Case 2, all seven analysts reported the correct deduced genotypes for the major contributor, which, as described earlier, had mostly four allele loci (i.e., fully resolvable genotypes in this two-person mixture). In MIX05 Case 3, which was a balanced two-person mixture, Analysts "d" and "e" obtained the correctly deduced genotypes while the other five analysts reported some inconsistencies. Finally, in MIX05 Case 4, only one analyst attempted to report a few alleles for the "extreme" minor contributor.

3.2.2. Summary of MIX13 responses

Four participating laboratories in the MIX13 study provided intra-laboratory results, which are labeled here at "A", "B", "C", and "D". Each laboratory that provided intra-laboratory results gave the five mixture profiles to their analysts to interpret independently, and then compiled and reported their results. Information in Table 4 represents consensus results for the laboratory based on each intra-laboratory participant.

A detailed summary of intra-laboratory responses is present in Supplemental Table S5. There were eight analysts from Lab A, 10 analysts from Lab B, 16 analysts from Lab C, and 25 reports provided by analysts from Lab D. For example, of the eight analysts providing results from Lab A, six analysts reported mRMP results and two provided CPI results for MIX13 Case 1 when considering reference profile 1A. The log mRMP results ranged from 10.6 to 20.7, based on the loci and genotypes that were deemed appropriate for interpretation by the individual analyst.

The intra-laboratory results mirrored the inter-laboratory results by highlighting the wide range of variation from the reported results. The "spread" between the lowest and highest reported statistics ranged from 4.4 orders of magnitude (laboratory D, Table S5(d)) and 13.3 orders of magnitude (laboratory C, Table S5-(c)).

3.2.2.1. MIX13 Case 1 (2-person mixture, 1:1 ratio). A great deal of within-laboratory variation was observed in Laboratories A and C (up to 10 orders of magnitude between the lowest reported statistic and the highest reported statistic) compared to laboratory B, where the variation was only two to four orders of magnitude. Laboratory D provided results where all analysts reported the same statistic (Table S5-(d)).

3.2.2.2. MIX13 Case 2 (3-person mixture, 6:1.5:1 ratio). As in Case 1, there was a great deal of variation in the statistics reported and range of results within laboratories A and C for Case 2. For example, two analysts in laboratory A provided log mRMP results for reference 2A while four analysts used a CPI approach to include 2A and two analysts declared the comparison of reference 2A to the evidence mixture to be inconclusive (Table S5-(a)). For laboratory C, three analysts provided either mRMP or LR statistics for all three references 2A, 2B, and 2C while others used CPI to include these three reference profiles, or reported an exclusion or inconclusive result (Table S5-(c)). Laboratory B only had two analysts provide an inclusion and statistical weight for reference 2A and one analyst include 2B (Table S5-(b)). All other analysts provided either inconclusive results or exclusions. For laboratory D, all 25 analysts determined the mixture to be too complex and provided inconclusive results (Table S5-(d)).

3.2.2.3. MIX13 Case 3 (3-person mixture, 7:2:1 ratio). Case 3 was a complex mixture with a relative (sibling) in the mixture. Laboratories A and B mostly provided inconclusive/excluded/not reported results (Table S5-(a), (b)). Only a couple of analysts in each lab provided statistical conclusions. All but one analyst in Laboratory C provided statistical results as a log LR or log mRMP that ranged from 13.3 to 7.5 orders of magnitude, respectively (Table S5-(c)). For laboratory D, only 4 of 25 analysts reported inclusive results and gave a log mRMP between 0.30 to 0.36. Given the complexity of this example, we observed fewer analysts within the same laboratory providing inclusions for reference profile 3A except for one laboratory. Laboratories A, B, and D (Table S5) generally reported similar results for those analysts who made an inclusion. The exception was laboratory C where all but one analyst reported an inclusion and provided a statistic (either LR or mRMP). The range of results among the analysts reporting a LR for laboratory C was 13.3 orders of magnitude from low to high, and 7.5 orders of magnitude difference for those analysts that reported mRMP results. We highlight the consistency observed again in laboratory D where only 4 of the 25 reported results gave a mRMP statistic, but these were essentially the same value, which is indicative of the analysts being trained to interpret this type of complex mixture in the same way.

3.2.2.4. Case 4 (2-person mixture, 3.5:1 ratio). Nearly all analysts reporting results for Case 4 gave an inclusion and provided either a LR or mRMP statistic. However, one analyst from laboratory A provided CPI results while the other seven analysts from this same laboratory returned mRMP results. Without more information, it is not possible to ascertain if this outlier CPI result arose due to training differences of the analysts involved or is an artifact of the study itself. The amount of variation within each laboratory varied widely: 9.6 orders of magnitude for laboratory A, 8.3 orders of magnitude for LR results and 3.8 orders of magnitude for mRMP results in laboratory B, 4.0 orders of magnitude for LR results and 4.4 orders of magnitude for mRMP results in laboratory C, and 4.4 orders of magnitude for laboratory D (Table S5).

3.2.2.5. Case 5 (4-person mixture, 1:1:1:1 ratio). The most consistent intra-laboratory results were observed for the non-contributor reference profile 5C. All analysts from laboratory A included the reference and gave the same statistic (Table S5(a)). The 12 analysts that included reference 5C from laboratory C gave essentially the same CPI result (less than one order of magnitude difference, Table S5(c)). All 25 reporting analysts from laboratory D turned in an inconclusive result due to the complexity of the profile (Table S5(d)). For laboratory B, only 2 of the 10 analysts gave an inclusion and provided the same CPI statistic. More analysts in this laboratory reported exclusions or inconclusive results (Table S5(d)). This is most likely because laboratory B used the PP16 STR chemistry and faced the Penta E challenge of discordance (see Fig. 3).

3.3. Informing participating laboratories

In some cases, such as in the German DNA Profiling Group (GEDNAP) studies [34], a workshop is held each year to make results public from each study conducted. Generally, performance from each individual laboratory is kept anonymous through use of lab codes

known only to the specific participating laboratory and the study coordinator. The focus of an interlaboratory study is typically on the overall performance across the participating laboratories and lessons learned based on the study design.

Following the MIX05 study, NIST provided participating laboratories with a copy of the poster presented at the ISHI meeting in September 2005 so they could obtain the "correct" answers for each of the four mixture cases. Laboratories could then self-grade in terms of how well they did against the correct answers. This poster has been available on the NIST STRBase website since the ISHI 2005 meeting (https://strbase.nist.gov/interlab/MIX05/MIX05poster.pdf). Almost all of the participants in the MIX13 study attended the DNA Technical Leader Summit held in November 2013 and so learned of the "correct" answers there and the overall performance of participants on each of the five provided mixture cases. Laboratories were again encouraged to self-grade following the MIX13 study and to use the information gained as a teaching opportunity. Learning from mistakes can be beneficial [3].

These interlaboratory studies were not intended as a proficiency test but rather as a training tool and an opportunity to discover the general performance across the community with the mixture scenarios being explored. The focus of any NIST presentations about these studies has been on the overall variation observed across the community. In most cases, individual laboratories would not have known who they were (i.e., which laboratory number in the overall data set) unless their results specifically stood out for some reason.

4. Observations and lessons learned from MIX05 and MIX13 studies

It is important to keep in mind that interlaboratory studies like MIX05 and MIX13 may not always provide a full window into day-to-day performance in forensic laboratories. Variation observed and mistakes made in interlaboratory performance does not necessarily equate to innocent people being in jail – or the improper application of mixture interpretation in a specific case. Despite requests that the provided data be treated as if they involved real cases, results reported may not always have been handled as such. Some participants shared that results were provided back to NIST without the typical technical review that would be present before a real case report is released. Other laboratories may have conducted more extensive review than normal in reporting their results. For example, one laboratory, which did well in the MIX05 study, shared that "the Profiler Plus and COfiler sample files were evaluated by four different analysts [note: it is implied in the report that the initial interpretations were performed independently but this is not explicitly stated], using both [Window] NT and MAC analysis platforms. The analysts checked for concordance, and a single conclusion for each mock case has been issued." This laboratory response then went on to describe all assumptions made in their MIX05 work outside the course of routine casework and how a flowchart was used in their mixture interpretation process. Detailed genotype calculations were described by some participants in MIX05 and MIX13 while others simply listed conclusions and accompanying inclusionary statistics without any detail of how the conclusions or statistical values were derived.

An important purpose of MIX13 was to determine how well laboratories were progressing with interpreting more complex mixtures compared to MIX05 and with implementation of the SWGDAM 2010 autosomal STR interpretation guidelines [22]. We had previously observed in the MIX05 study that only a few laboratories at that time used some form of a stochastic threshold and that some results seemed to be widely distributed not only between analysts in differing laboratories but also within the same laboratory. With the MIX13 study conducted eight years later, we still observe a great deal of variation within and between laboratories when thresholds are included in the interpretation. Some of this variation is understandable, some of it is not. Some variation in results may be expected due to assumptions that laboratories have made in building their interpretation protocols. However, the intra-laboratory results suggest that training consistency may be an issue in some situations as different analysts in the same laboratory using the same protocol provided different results (Tables S4 and S5).

When laboratories and analysts were presented simple, straightforward two-person mixtures such as MIX05 Case 2 or MIX13 Case 1 and Case 4, participants drew correct conclusions and reference samples were correctly included or excluded. More complex samples and scenarios, or situations where a great degree of allele overlap existed in the mixtures, produced more variation in responses.

When provided mixtures of three or more individuals with either low-level contributors (MIX13 Case 2), relatives (MIX13 Case 3), or uncertainty in the number of contributors (MIX13 Case 5), laboratories responding at the time generally relied more and more on CPI statistics *as a method of interpretation* since deconvolution of mixture components was viewed as being too complex. That is, if the alleles in the profile were above the stochastic threshold, then the locus was used for statistics without an interpretation as to whether drop-out may be possible (see Ref. [28]). We observed this in MIX13 Case 2 where laboratories were including individuals where allele drop-out was evident and building a statistic on the three to four loci where all alleles were above the stochastic threshold.

The danger of using this strategy was evident in MIX13 Case 5, where despite the presence of extra alleles unattributed to the three references provided, and the near unanimity of laboratories reporting the high degree of allele sharing in this profile, most laboratories relied on the criteria that, "as long as the alleles were above the stochastic threshold, an inclusion and a CPI statistic can be provided." This overreliance on CPI as an interpretation method was also observed in the "easier" mixtures, such as the D16S539 locus from Case 4, where none of the labs that reported locus data as part of their report considered the possibility of drop-out of the allele 11 being in the stutter position of the major allele 12 (see Ref. [21]).

As we look to the future, the community may ask if there are obvious improvements necessary to achieve more reliable mixture interpretation. Is it possible to produce a "standard" mixture approach that all laboratories can implement to achieve consistency across the United States or around the world? Probably not. Protocols for interpretation are developed depending on different chemistries, different capillary electrophoresis platforms, different philosophies on interpreting mixtures, and the experience and training of analysts in the laboratory.

However, we should nevertheless strive to achieve consistency within each laboratory to avoid the possibility of different conclusions as highlighted by the intra-laboratory results from Laboratory B for MIX13 Case 5, where the PP16HS kit was used (Table S5-(b)). Under nominally the same interpretational protocol, 50% of the analysts effectively said, "I don't know"; 30% of the analysts said, "He's not there"; and 20% of the analysts said, "He's not only in the mixture, but I can exclude greater than 99.9% of the population." It would appear under this tested scenario at that time in that laboratory that presentation of DNA favorable or not favorable to the person of interest may depend upon the analyst assigned to the case when the evidence comes in the door. Clearly, this seemingly subjective variation is undesirable!

In contrast, one large laboratory showed a great deal of consistency in their results (Table S5-(d)). This laboratory proved that it is possible to achieve consistency within a laboratory through a commitment to training and technical leadership. We have heard that an important outcome of this collaborative exercise is that some laboratories participating in the MIX13 study have implemented a routine mixture challenge to their analysts to help achieve better consistency. A regular review of DNA mixture interpretation performance within and across laboratories is expected to highlight areas for potential improvement.

We are encouraged by the developments in probabilistic software systems that will no longer rely upon philosophies that "drop a locus" or "assume 0% stutter," but instead model parameters such as allele

drop-out and consider all information in the profile without relying on stochastic thresholds. The difficulty of interpreting evidence with relatives (MIX13 Case 3) is probably better realized with probabilistic software than simply using CPI.

At the time of the MIX13 study only two probabilistic genotyping software programs were available and used by three laboratories in this study (although not on all the mixtures). It is likely that such software systems can improve consistency within the laboratory. However, there are examples of differences occurring when using the same software within and between laboratories [15]. The bottom line is that analysts cannot blindly accept the results of a software analysis (i.e., submit data to the software then simply copy and paste the LR results) without the due diligence of human interpretation both before and after the software analysis step.

Inter-laboratory studies measure variation among laboratories while intra-laboratory studies enable assessment of variation among analysts in a single laboratory. Both studies can be valuable tools to understand measurement uncertainty as well as the effectiveness of laboratory training. Several important benefits come out of these studies. First, data sets from a variety of STR kits now exist with multiple mixture scenarios representing a range of situations that might be seen in forensic casework that can be used for training purposes. These data are available for download from the NIST STRBase website. A wide variety of approaches to mixture interpretation have been applied to the same data set enabling specific approaches and protocols to be evaluated as part of these studies. Additionally, with numerous forensic practitioners evaluating the same mixture data, best practices and poor practices were identified, which can aid in future training. A large dataset of mixtures to assist in this effort was recently released [35]. Finally, laboratories have seen the value of regular review of their analysts' work and some have implemented periodic internal mixture challenges to assess and verify their mixture protocol performance.

5. Conclusions

The results described in this article provide only a brief snapshot of DNA mixture interpretation as practiced by participating laboratories in 2005 and 2013. Any overall performance assessment is limited to participating laboratories addressing specific questions with provided data based on their knowledge at the time. Given the adversarial nature of the legal system, and the possibility that some might attempt to misuse this article in legal arguments, we wish to emphasize that variation observed in DNA mixture interpretation cannot support any broad claims about "poor performance" across *all* laboratories involving *all* DNA mixtures examined in the past. Some variation is to be expected due to use of different STR kits, different assumptions, and other variables mentioned above.

DNA mixture interpretation, as can be seen from this study, is a complex area. Variation in the chemistries, analytical approaches, and software to interpret inevitably lead to variation across participating laboratories. This study highlights the difference in agreement when dealing with simple mixtures, or instances when the genotype of interest is the major profile, compared with complex mixtures and situations where the genotype of interest is a minor profile. In addition, limitations in the use of CPI for complex mixtures were highlighted in several of the MIX13 cases.

We hope that presenting the overall variation observed in the MIX05 and MIX13 studies will contribute to the adoption of more uniform approaches to mixture interpretation. Despite improvements in protocols and interpretation guidelines across the United States and Canada since the SWGDAM interpretation guidelines were released in 2010, results of mixture interpretation were still highly variable several years later when the MIX13 study was conducted. Some of this variation was *a consequence of inappropriately using CPI to interpret complex mixtures*. As demonstrated in MIX13 Case 5, there is a risk of including a non-contributor when blindly applying CPI without interpretation of the DNA mixture itself. We recognize that many laboratories are implementing probabilistic genotyping software systems to assist in the deconvolution and statistical evaluation of complex mixtures. Future interlaboratory studies will be helpful in assessing how effective these software approaches are at improving performance across the community. These results, as with previous collaborative exercises, can be tools to draw attention to issues that can lead to improvements in the field.

Acknowledgments

Interlaboratory studies are not possible without the contributions of participating laboratories. We express our appreciation to the many forensic scientists and their supervisors who took time out of their busy schedules to examine the MIX05 and MIX13 data sets and to provide results and reports analyzed as part of this study.

Assistance in the study design was provided by Charlotte Word (MIX13). An Excel-macro entitled "VirtualMixtureMaker" developed by David Duewer enabled virtual mixture creation with desired allele combinations. Synthetic DNA mixtures with known mixture ratios were prepared by Becky Steffen (MIX13: Cases 1,3,4,5) or obtained as electropherograms (MIX13: Case 2) from the Boston University resource available at http://www.bu.edu/dnamixtures/ prepared by Catherine Grgicak and Robin Cotton. For MIX05, FMBIO data were generated by Chris Tomsey, Frank Krist, Kermit Channel, and Mary Robnett. Jan Redman assisted in shipping MIX05 data disks to enrolled laboratories. Brooke Morgan provided assistance in reviewing MIX13 results and compiling data tables. Jo-Anne Bright provided data interpretation assistance and helpful comments on MIX13. Hari Iyer helped generate the box and whisker plots. John Paul Jones II helped organize webcasts and the DNA Technical Leaders' Summit held in November 2013 where the initial results of MIX13 were shared. Rich Cavanagh, Rich Press, Pete Vallone, Sheila Willis, and Kathy Sharpless provided helpful feedback during the internal NIST review process as this manuscript was being prepared.

For MIX05, the National Institute of Justice funded through interagency agreement 2003-IJ-R-029 with the NIST Office of Law Enforcement Standards. For MIX13, funds were provided to the Applied Genetics Group by the NIST Special Programs Office.

Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice or the National Institute of Standards and Technology. Certain commercial entities are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that any of the entities identified are necessarily the best available for the purpose.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.fsigen.2018.07.024.

References

[1] ISO/IEC 17025, General Requirements for the Competence of Testing and Calibration Laboratories. International Organization for Standardization: Geneva, Switzerland. Section 7.2.2.1(e) Mentions Interlaboratory Comparisons As a Technique Used for Method Validation and Section 7.7.2(b) Encourages Laboratories to Monitor Their Performance by Participation in Interlaboratory Comparisons (2017).

[2] X.L. Pang, J.D. Fox, J.M. Fenton, G.G. Miller, A.M. Caliendo, J.K. Preiksaitis, Interlaboratory comparison of cytomegalovirus viral load assays, Am. J. Transplant. 9 (2009) 258–268.

[3] J. Metcalfe, Learning from errors, Annu. Rev. Psychol. 68 (2017) 465–489.

[4] J. Koehler, Proficiency tests to estimate error rates in the forensic sciences, Law Probab. Risk 12 (2013) 89–98.

[5] P.M. Schneider, R. Fimmers, W. Keil, G. Molsberger, D. Patzelt, W. Pflug, T. Rothämel, H. Schmitter, H. Schneider, B. Brinkmann, The German Stain Commission: recommendations for the interpretation of mixed stains, Int. J. Legal Med. 123 (1) (2009) 1–5.
[6] T.M. Clayton, J.P. Whitaker, R. Sparkes, P. Gill, Analysis and interpretation of mixed forensic stains using DNA STR profiling, Forensic Sci. Int. 91 (1) (1998) 55–70.
[7] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, N. Morling, M. Prinz, P.M. Schneider, B.S. Weir, DNA commission of the international society of forensic genetics: recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2006) 90–101.
[8] M.C. Kline, D.L. Duewer, P. Newall, J.W. Redman, D.J. Reeder, M. Richard, Interlaboratory evaluation of STR triplex CTT, J. Forensic Sci. 42 (1997) 897–906.
[9] D.L. Duewer, M.C. Kline, J.W. Redman, P.J. Newall, D.J. Reeder, NIST mixed stain studies #1 and #2: interlaboratory comparison of DNA quantification practice and short tandem repeat multiplex performance with multiple-source samples, J. Forensic Sci. 46 (2001) 1199–1210.
[10] M.C. Kline, D.L. Duewer, J.W. Redman, J.M. Butler, NIST mixed stain study 3: DNA quantitation accuracy and its influence on short tandem repeat multiplex signal intensity, Anal. Chem. 75 (2003) 2463–2469.
[11] D.L. Duewer, M.C. Kline, J.W. Redman, J.M. Butler, NIST mixed stain study #3: signal intensity balance in commercial short tandem repeat multiplexes, Anal. Chem. 76 (2004) 6928–6934.
[12] M. Crespillo, P.A. Barrio, J.A. Luque, C. Alves, M. Aler, F. Alessandrini, L. Andrade, R.M. Barretto, A. Bofarull, S. Costa, M.A. García, O. García, A. Gaviria, A. Gladys, A. Gorostiza, A. Hernández, M. Herrera, L. Hombreiro, A.A. Ibarra, M.J. Jiménez, G.M. Luque, P. Madero, B. Martínez-Jarreta, M.V. Masciovecchio, N.M. Modesti, F. Moreno, S. Pagano, S. Pedrosa, G. Plaza, E. Prat, J. Puente, F. Rendo, T. Ribeiro, A. Sala, E. Santamaría, V.G. Saragoni, M.R. Whittle, 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.
[13] L. Prieto, H. Haned, A. Mosquera, M. Crespillo, M. Aleman, M. Aler, F. Alvarez, C. Baeza-Richer, A. Dominguez, C. Doutremepuich, M.J. Farfán, M. Fenger-Grøn, J.M. García-Ganivet, E. González-Moya, L. Hombreiro, M.V. Lareu, B. Martínez-Jarreta, S. Merigioli, P. Milans Del Bosch, N. Morling, M. Muñoz-Nieto, E. Ortega-González, S. Pedrosa, R. Pérez, C. Solís, I. Yurrebaso, P. Gill, Euroforgen-NoE collaborative exercise on LRmix to demonstrate standardization of the interpretation of complex DNA profiles, Forensic Sci. Int. Genet. 9 (2014) 47–54.
[14] R. Aranda, A large-scale study of DNA mixture interpretation: inter- and intra-laboratory variability, in: J.M. Butler (Ed.), Proceedings of the 2015 International Symposium on Forensic Science Error Management (NIST Special Publication 1206), 2015, p. 60, , https://doi.org/10.6028/NIST.SP.1206 Available at.
[15] S. Cooper, C. McGovern, J.A. Bright, D. Taylor, J. Buckleton, Investigating a common approach to DNA profile interpretation using probabilistic software, Forensic Sci. Int. Genet. 16 (2015) 121–131.
[16] U. Toscanini, L. Gusmao, M.C. Alava Narvaez, J.C. Alvarez, L. Baldassarri, A. Barbaro, G. Berardi, H.E. Betancor, M. Camargo, J. Carreras-Carbonell, J. Castro, S.C. Costa, P. Coufalova, V. Domínguez, E. Fagundes de Carvalho, S.T.G. Ferreira, S. Furfuro, O. García, A. Goios, R. González, A.G. de la Vega, A. Gorostiza, A. Hernández, S. Jiménez Moreno, M.V. Lareu, A. León Almagro, M. Marino, G. Martínez, M.C. Miozzo, N.M. Modesti, V. Onofri, S. Pagano, B. Pardo Arias, S. Pedrosa, G.A. Penacino, M.L. Pontes, M.J. Porto, J. Puente-Prieto, R.R. Pérez, T. Ribeiro, B. Rodríguez Cardozo, Y.M. Rodríguez Lesmes, A. Sala, B. Santiago, V.G. Saragoni, A. Serrano, E.R. Streitenberger, M.A. Torres Morales, S.A. Vannelli Rey, M. Velázquez Miranda, M.R. Whittle, K. Fernández, A. Salas, Analysis of uni and bi-parental markers in mixture samples: lessons from the 22nd GHEP-ISFG intercomparison exercise, Forensic Sci. Int. Genet. 25 (2016) 63–72.
[17] C.C. 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.
[18] P.A. Barrio, M. Crespillo, J.A. Luque, M. Aler, C. Baeza-Richer, L. Baldassarri, E. Carnevali, P. Coufalova, I. Flores, O. García, M.A. García, R. González, A. Hernández, V. Inglés, G.M. Luque, A. Mosquera-Miguel, S. Pedrosa, M.L. Pontes, M.J. Porto, Y. Posada, M.I. Ramella, T. Ribeiro, E. Riego, A. Sala, V.G. Saragoni, A. Serrano, S. Vannelli, GHEP-ISFG collaborative exercise on mixture profiles (GHEP-MIX06). Reporting conclusions: results and evaluation, Forensic Sci. Int. Genet. 35 (2018) 156–163.
[19] J.M. Butler, Mixture interpretation: lessons learned from the MIX05 interlaboratory study, Presentation Given at the CODIS Conference (2006) Available at https://strbase.nist.gov/pub_pres/CODIS2006_MIX05.pdf.
[20] J.M. Butler, Mixture interpretation issues and insights, Presentation to the Scientific Working Group on DNA Analysis Methods (2007) Available at https://strbase.nist.gov/pub_pres/SWGDAM_Jan2007_MixtureInterpretation.pdf.
[21] M.D. Coble, MIX13: an interlaboratory study on the present State of DNA mixture interpretation in the U.S. presentation given at the America Bar Association's, 5th Annual Prescription for Criminal Justice Forensics (2014) Available at https://strbase.nist.gov/pub_pres/Coble-ABA2014-MIX13.pdf.
[22] SWGDAM, SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories, Available at (2010) http://www.forensicdna.com/assets/swgdam_2010.pdf.
[23] SWGDAM, SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories, Available at (2017) https://www.swgdam.org/publications.
[24] J.-A. Bright, D. Taylor, S. Gittelson, J. Buckleton, The paradigm shift in DNA profile interpretation, Forensic Sci. Int. Genet. 31 (2017) e24–e32.
[25] M.C. Kline, D.L. Duewer, J.W. Redman, J.M. Butler, Results from the NIST 2004 DNA quantitation study, J. Forensic Sci. 50 (3) (2005) 571–578.
[26] J.M. Butler, R. Schoske, P.M. Vallone, J.W. Redman, M.C. Kline, Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations, J. Forensic Sci. 48 (4) (2003) 908–911.
[27] C.R. Hill, D.L. Duewer, M.C. Kline, M.D. Coble, J.M. Butler, U.S. population data for 29 autosomal STR loci, Forensic Sci. Int. Genet. 7 (2013) e82–e83.
[28] F.R. Bieber, J.S. Buckleton, B. Budowle, J.M. Butler, M.D. Coble, Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion, BMC Genet. 17 (1) (2016) 125.
[29] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Sci. Int. Genet. 4 (2009) 1–10.
[30] J. Buckleton, C. Triggs, Is the 2p rule always conservative? Forensic Sci. Int. 159 (2–3) (2006) 206–209.
[31] B. Budowle, A.J. Onorato, T.F. Callaghan, M.A. Della, A.M. Gross, R.A. Guerrieri, J.C. Luttman, D.L. McClure, Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework, J. Forensic Sci. 54 (4) (2009) 810–821.
[32] J.M. Curran, J. Buckleton, Inclusion probabilities and dropout, J. Forensic Sci. 55 (5) (2010) 1171–1173.
[33] D.R. Paoletti, T.E. Doom, C.M. Krane, M.L. Raymer, D.E. Krane, Empirical analysis of the STR profiles resulting from conceptual mixtures, J. Forensic Sci. 50 (6) (2005) 1361–1366.
[34] S. Rand, M. Schürenkamp, B. Brinkmann, The GEDNAP (German DNA profiling group) blind trial concept, Int. J. Legal Med. 116 (2002) 199–206.
[35] L.E. Alfonse, A.D. Garrett, D.S. Lun, K.R. Duffy, C.M. Grgicak, A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt, Forensic Sci. Int. Genet. 32 (2018) 62–70.
[36] J.M. Butler, Advanced Topics in Forensic DNA Typing: Interpretation, Elsevier Academic Press, San Diego, 2015.