E-FILED
Wednesday, 06 February, 2019 05:12:11 PM
Clerk, U.S. District Court, ILCD

IN THE UNITED STATES DISTRICT COURT
CENTRAL DISTRICT OF ILLINOIS
URBANA DIVISION

UNITED STATES OF AMERICA,	)	
	)	
Plaintiff,	)	
	)	
vs.	)	Case No. 17-CR-20037
	)	
BRENDT A. CHRISTENSEN,	)	
	)	
Defendant.	)	

THE UNITED STATES OF AMERICA'S RESPONSE TO THE DEFENDANT'S MOTION TO EXCLUDE DNA AND SEROLOGY TEST RESULTS AND REQUEST FOR *DAUBERT* HEARING

NOW COMES the United States of America, by John C. Milhiser, United States Attorney for the Central District of Illinois, Eugene L. Miller and Bryan D. Freres, Assistant United States Attorneys, and James B. Nelson, Department of Justice Trial Attorney, and hereby requests that this Court deny the Defendant's Motion to Exclude DNA and Serology Test Results and Request for *Daubert* Hearing (R.119) because (1) the probabilistic genotyping software used in this case is a reliable methodology to assign a weight to a DNA match; (2) the defendant does not need and is not entitled to the source code for the proprietary software; (3) comparing alleles by size (*i.e.,* length) is a reliable methodology to determine a match or inclusion of a sample to a known source; (4) the presumptive serology test results using luminol and phenolphthalein are relevant and not unfairly prejudicial; and (5) the confirmatory serology test results were obtained by use of a reliable methodology.

BACKGROUND

A federal grand jury charged the defendant, Brendt A. Christensen, with kidnapping Yingying Zhang, and further alleged that he intentionally killed her in an especially heinous, cruel, and depraved manner after substantial planning and premeditation. (R.26) The defendant has made recorded statements that he took the victim to his apartment and engaged in conduct that would result in her bleeding in the apartment.

At trial, the United States intends to present expert testimony regarding the identification of deoxyribonucleic acid (DNA) and blood identified from samples taken from the defendant's apartment, which confirm the defendant's statements. Some of the samples were identified by the use of luminol, which can detect minute traces of blood even after an attempt has been made to wash the blood away. The samples were analyzed at the FBI Laboratory in Quantico, Virginia, using reliable principles and methods generally accepted in the scientific community. More specifically, in conducting its DNA analysis, the FBI Laboratory compared the length of alleles and used proprietary probabilistic genotyping software called STRmix™. Regarding the serology results, the FBI laboratory used phenolphthalein and Takayama hemochromogen testing.

Thereafter, the United States produced to the defendant multiple reports of examinations and tests under Rule 16(a)(1)(F) of the Federal Rules of Criminal Procedure, including reports concerning the results of the use of luminol, DNA testing, and serology testing. The United States also disclosed to the defendant under Rule

16(a)(1)(G) the written summary of the expected expert testimony of FBI Forensic Examiner Amanda Bakker, which the United States intends to use during its case-in-chief at trial, as well as proficiency testing regarding Ms. Bakker.

Additionally, the United States also provided the defendant with the notes (over 500 pages) of Ms. Bakker, the 1A Case File generated by the FBI Evidence Control Coordinator, a CD Rom containing relevant raw computer data files, including raw data collected in the course of a capillary electrophoretic run, and a CD Rom containing the Laboratory Operations Manual and the Standard Operating Procedure to include the FBI Approved Standards for Testimony and Report Language utilized by the DNA Unit and STR frequency tables. The United States also made relevant STR databases and data files related to the FBI's internal validation study available for the defense's inspection at the FBI Laboratory in Quantico, Virginia. Furthermore, the United States referred the defense to multiple journals detailing the samples used in the databases, allele frequencies, developmental validation studies of STRmix, and the FBI's internal validation study.

On July 11, 2018, after providing all of this information, the United States requested that the defendant provide to the United States, pursuant to Rule 16(b)(1)(C), a written summary of any testimony that the defendant intended to use under Rules 702, 703, or 705 of the Federal Rules of Evidence as evidence at trial, which described the witness's opinions, the bases and reasons for those opinions, and the witness's qualifications. (The request noted that, at the time, the Court's scheduling order required

disclosure of the defendant's non-Rule 12.2 expert witnesses, including rebuttal experts, on or before August 24, 2018.)

On August 24, 2018, the defendant disclosed that he intended to elicit expert testimony on the following subject:

> DNA, to rebut the government's expert testimony if necessary, and to challenge the reliability of the DNA and serology results from the FBI Laboratory in Quantico, Virginia. The specific expert on this topic has not yet been identified, although he/she will be associated with and/or employed by Forensic Bioinformatics, 2850 Presidential Drive, Suite 160, Fairborn, OH, 45324.

In the same disclosure, the defendant noted that he had not yet identified and/or retained the individual witness who would testify and that, recognizing his obligations under Rule 16(b)(1)(C), he would disclose all written summaries of any testimony that he intended to offer as soon as it became available. To date, the defendant has not identified his expert witness or disclosed a written summary of testimony beyond the paragraph quoted above.[1]

On that same date, the defendant filed a motion to exclude the DNA and serology test results and requested a *Daubert* hearing. (R.119) Regarding the DNA test results, the defendant alleged that there are "potential problems with the application of probabilistic genotyping software," including "the scientific validity of probabilistic genotyping

[1]The defendant recently filed a motion alleging that the lapse in government funding has prevented retention of his mental health experts. (R.213) Although the United States responded to the motion separately (R.219) and the motion does not discuss DNA experts, the United States would note that the defendant had identified the employer of their DNA expert by August 24, 2018. As the lapse in government funding did not begin until December 22, 2018, it is unclear why the defendant could not provide an expert report, or at the very least, identify an expert from the disclosed employer in the intervening four months.

algorithms" and that the algorithms "are not free of subjectivity." Therefore, the defendant requested the Court (1) require the United States to disclose the source code for STRmix to be examined for errors by a defense expert; and (2) conduct a *Daubert* hearing regarding the reliability of the probabilistic genotyping and STRmix.

Regarding the serology testing, the defendant requested that the Court (1) exclude the results of any luminol[2] or phenolphthalein tests from trial under Rules 401, 402, and 403 of the Federal Rules of Evidence; and (2) conduct a *Daubert* hearing regarding the reliability of any serology confirmatory testing, specifically, Takayama hemochromogen testing.

On January 30, 2019, the defendant filed a supplemental memorandum in support of his motion to exclude the DNA test results. The supplemental memorandum did not disclose the identity of the defense DNA expert or a written report. Instead, the defendant raised a new issue. He claimed that the FBI's DNA test results are unreliable because he alleges the FBI analyst should have used a new DNA analysis methodology referred to as next-generation sequencing (NGS), rather than the procedure that has been used for years at the FBI Laboratory and at forensic laboratories throughout the United States.

[2] Luminol was not used by the FBI Laboratory, but by the crime scene technicians at the defendant's apartment to identify locations from which to obtain suspected biological samples for further testing. Because this evidence has evidentiary value beyond the identification of blood itself and explains why and where the technicians obtained the samples they sent to the FBI Laboratory, the evidence should be admitted, as argued, *infra.* Moreover, the technicians who applied the luminol and obtained the samples would be testifying as fact witnesses, not expert witnesses.

The defendant still has not identified an expert who will testify regarding the reliability of the evidence he challenges at any *Daubert* hearing or at trial. In fact, at a previous hearing, the defendant indicated he intended merely to call the government expert at the scheduled *Daubert* hearing and cross-examine her. This is improper. The defendant should not be allowed, under the guise of a *Daubert* motion, to cross-examine the government's expert prior to trial, while shielding his own expert not only from cross-examination, but also from disclosure and a possible *Daubert* challenge. The defendant's motion should be denied without a hearing, or in the alternative, the hearing should be limited to relevant pre-trial matters, as argued, *infra*.[3]

APPLICABLE LAW

I. Scientific Testimony And Rule 702 Of The Federal Rules Of Evidence

In 1993, the Supreme Court interpreted Rule 702 of the Federal Rules of Evidence as abandoning the prior requirement that a necessary precondition to admissibility of scientific evidence was that it be generally accepted in the scientific community (the so-called "*Frye* test"). *Daubert v. Merrell Down Pharmaceuticals, Inc.*, 509 U.S. 579, 597 (1993). Instead, the Court held that the Federal Rules of Evidence require a trial judge to ensure "that an expert's testimony both rests on a reliable foundation and is relevant to the task at hand." *Id.* The Supreme Court later suggested the same analysis might apply in

3 Based on the defendant's claims, including his most recent claims made on January 30, 2019, the United States would intend to call FBI Forensic Examiner Jerrilyn M. Conway to rebut the defendant's claims, if an evidentiary hearing is held.

assessing the reliability of non-scientific expert testimony. *Kumho Tire Co. v. Carmichael,* 119 S. Ct. 1167 (1999).

In 2000, Rule 702 was amended in response to *Daubert* and *Kumho*, causing the Seventh Circuit to recently note that Rule 702 has superseded *Daubert. Kansas City S. Ry. Co. v. Sny Island Levee Drainage Dist.,* 831 F.3d 892, 900 (7th Cir. 2016); *but see Manpower, Inc. v. Ins. Co. of Pennsylvania*, 732 F.3d 796, 806 (7th Cir. 2013) (noting that "*Daubert* interpreted an earlier version of Rule 702, but it remains the gold standard for evaluating the reliability of expert testimony and is essentially codified in the current version of Rule 702"). Rule 702 allows a qualified expert to testify if (a) the expert's scientific knowledge will help the trier of fact to understand the evidence or to determine a fact in issue; (b) the testimony is based on sufficient facts or data; (c) the testimony is the product of reliable principles and methods; and (d) the expert has reliably applied the principles and methods to the facts of the case.

In evaluating whether the testimony is the product of reliable principles and methods, a district court should focus solely on the reliability of the principles and methods. The reliability of the ultimate conclusion is for the jury to decide, not the district court during a pretrial hearing:

> Reliability, however, is primarily a question of the validity of the methodology employed by an expert, not the quality of the data used in applying the methodology or the conclusions produced. "The soundness of the factual underpinnings of the expert's analysis and the correctness of the expert's conclusions based on that analysis are factual matters to be determined by the trier of fact . . . . Rule 702's requirement that the district judge determine that the expert used reliable methods does not ordinarily extend to the reliability of the conclusions those methods produce—that is, whether the conclusions are unimpeachable." The district court usurps the

> role of the jury, and therefore abuses its discretion, if it unduly scrutinizes the quality of the expert's data and conclusions rather than the reliability of the methodology the expert employed.

Manpower, 732 F.3d at 806 (citations omitted) (reversing district court where it "supplanted that adversarial process with its admissibility determination"); *see also Stollings v. Ryobi Techs., Inc.,* 725 F.3d 753, 766 (7th Cir. 2013) (finding that trial "judge's exclusion of . . . expert testimony on reliability grounds intruded too far into the province of the jury"); *In re Processed Egg Products Antirust Litigation,* 81 F. Supp. 3d 412, 416 (E.D. Pa. Jan. 26, 2015) ("Proponents of expert testimony do not 'have to prove their case twice—they do not have to demonstrate to the judge by a preponderance of the evidence that the assessments of their experts are correct, they only have to demonstrate by a preponderance of evidence that their opinions are reliable.'") (citations omitted).

In assessing the reliability of a scientific expert's principles and methods, a district court should look at factors such as (1) whether the scientific theory or technique can be and has been tested; (2) whether the theory or technique has been subjected to peer review and publication; (3) whether a particular technique has a known potential rate of error; and (4) whether the theory or technique is generally accepted in the relevant scientific community. *See Schultz v. Akzo Nobel Paints, LLC*, 721 F.3d 426, 431 (7th Cir. 2013) (citing *Daubert*).

II. DNA Testing Using Probabilistic Genotyping Software

As long ago as 2009, the Supreme Court stated that "DNA testing has an unparalleled ability both to exonerate the wrongly convicted and to identify the guilty. It has the potential to significantly improve both the criminal justice system and police investigative practices." *Dist. Attorney's Office for Third Judicial Dist. v. Osborne*, 557 U.S. 52, 55 (2009). The Supreme Court's acceptance of DNA testing has only grown over the years: "The advent of DNA technology is one of the most significant scientific advancements of our era. The . . . utility of DNA identification in the criminal justice system is already undisputed. Since the first use of forensic DNA analysis to catch a rapist and murderer in England in 1986, law enforcement, the defense bar, and the courts have acknowledged DNA testing's" reliability. *Maryland v. King,* 469 U.S. 435, 442 (2013) (citation omitted).

Thus, the results of DNA tests are universally admitted by courts in the United States, which have found that DNA analysis has the capacity to consistently, and with a high degree of certainty, demonstrate a connection between an evidentiary sample and a specific individual source. *Osborne,* 5557 U.S. at 80 (Alito, J. concurring) ("DNA tests can, in certain circumstances, establish to a virtual certainty whether a given individual did or did not commit a particular crime.") (citing *Harvey v. Horan*, 285 F.3d 298, 305 (4th Cir. 2002)). In fact, DNA testing is so well-accepted, that Congress had codified it in the United States Code. 18 U.S.C. § 3600. In 2013, the Supreme Court recognized that, although current DNA technology made it "possible to determine whether a biological

tissue matches a suspect with near certainty," "[f]uture refinements may improve present technology." *King*, 569 U.S. at 443.

Those future refinements have included the use of probabilistic genotyping and related software, such as STRmix. *See United States v. Morgan*, 675 F. App'x 53, 56 (2d Cir.), *cert. denied*, 138 S. Ct. 176 (2017) (noting that the New York Office of the Chief Medical Examiner was "discontinuing its use of LCN testing in favor of newer technology that produces reliable results," namely, probabilistic genotyping and new STR analysis software); *see also United States v. Lee*, No. 17-3559, 2018 WL 6600956 (2d Cir. Dec. 14, 2018) (rejecting defendant's challenge to use of STRmix probabilistic genotyping software).

Relevant here, probabilistic genotyping and STRmix software only come into play after an analyst finds a match between two DNA samples. At that point, the analyst assigns a weight to the match through statistical analysis. STRMix is a probabilistic genotyping software that calculates a likelihood ratio for DNA typing results. The FBI began using STRMix in its forensic laboratories in 2015 to assist in assigning statistical weights to its DNA typing results.

Every court to have considered the issue has found that the use of probabilistic genotyping and STRmix software are scientifically reliable and the results admissible. *See, e.g., People v. Smith,* No. 340845, 2018 WL 4926977, at *8 (Mich. App. Oct. 9, 2018); *People v. Muhammad,* No. 338300, 2018 WL 4927094, at *5 (Mich. App. Oct. 2, 2018); *People v. Blash,* No. 2015-CR-156, 2018 WL 4062322, at *8 (V.I. Super. Aug. 24, 2018); *United States v. Pettway*, No. 12-CR-103S, 2016 WL 6134493, at *3 (W.D.N.Y. Oct. 21, 2016); *People*

v. Bullard-Daniel, 42 N.Y.S.3d 714, 725-26 (N.Y. Cty. Ct. Mar. 10, 2016); *see also United States v. Oldman,* No. 18-CR-0020, Document 227, at 12 (D. Wyo. Dec. 31, 2018) (unpublished opinion) (finding probabilistic genotyping and STRmix software "are scientifically valid, proven, and tested"); *Smith v. State,* No. 12-16-139-CR, 2017 WL 1534048, at *2 (Tex. App.-Tyler Apr. 28, 2017) ("The new [STRmix] software can reliably consider all the available data"); *State v. Wakefield*, 9 N.Y.S.3d 540, 547 (N.Y. Sup. Ct. 2015) (holding as matter of first impression that DNA evidence using computerized probabilistic genotype analysis (*i.e.,* TrueAllele) is reliable and admissible).

III. DNA Testing Comparing Alleles By Length

Relevant to the defendant's supplemental memorandum, all forensic DNA analysts – including the analyst in this case – compare subject DNA samples with known DNA samples by comparing the size (*i.e.,* length) of alleles at different locations (or loci). The current approach is to compare the length of these alleles for the 23 short tandem repeat (STR) loci most commonly used in the United States and the world. For example, in this case, the analyst conducted a comparison between the victim's known DNA to the DNA recovered from various samples from the defendant's apartment and calculated a likelihood ratio. Given the DNA results, the likelihood ratios ranged from 1.4 quintillion (1.4×10^{18}) to 97 octillion (9.7×10^{28}). These likelihood ratios provide support that the victim was a contributor to the DNA in the defendant's apartment.

Next-generation sequencing (NGS) of the alleles is a potential improvement on current technology, much like probabilistic genotyping and STRmix software. Nonetheless, unlike those improvements, NGS has not yet been fully developed and

validated for forensic laboratories and is not commonly used in forensic analysis. Moreover, NGS is simply a more discriminating technology; it in no way suggests that the long-established methods of DNA analysis used in this case are unreliable. In fact, NGS would not result in an exclusion; it would be used simply to bolster the likelihood ratios determined through length analysis. Here, given the incredibly high likelihood ratios already generated (1.4 quintillion to 97 octillion), NGS would add little to the relevant analysis. More importantly, the development of NGS methodologies do not in any way draw into question the reliability of the DNA results in this case.

IV. Serology Testing

Forensic serology is the scientific identification and analysis of bodily fluids, including blood, on items of evidence. Today, serology is not used for source attribution because forensic DNA identity tests are more sensitive, specific, and easier to perform. Instead, serology is used to identify the tissue type of stain or biologic material. DNA, because it is present in all cells and tissues, cannot be used to distinguish between tissue types. At crime scenes, stains may or may not be obvious. Luminol sprayed over surfaces will cause blood to fluoresce bright blue under UV irradiation. The test is so sensitive it can detect minute traces of blood even after an attempt has been made to wash the blood away.

Serologic tests can be presumptive or confirmatory. Presumptive tests are typically chemical color tests, which are generally sensitive, but nonspecific. Confirmatory tests are highly specific, but can lack sensitivity. Serologic testing, which

does not interfere with DNA tests, is followed by DNA testing, which itself is specific to humans.

Presumptive blood tests take advantage of the peroxidase activity of blood. For example, in the presence of hydrogen peroxide solution, blood will cause phenolphthalein to turn pink. The confirmatory Takayama hemochromogen test procedure produces a positive result where the ferrous iron from hemoglobin reacts with pyridine to create red feathery crystals of pyridine ferroprotoporphyrin.

"[C]ourts . . . regularly admit evidence of Luminol testing for the presence of blood . . . ." *Cooper v. Brown*, 565 F.3d 581, 596 (9th Cir. 2009) (Fletcher, J. dissenting). For example, a district court has noted that "[w]e have held that luminol testing, as a scientific procedure, is sufficiently reliable for what it purports to do: presumptively indicate the possible presence of blood." *Dodd v. Workman*, No. CIV-06-140-D, 2011 WL 3299101, at *60 (W.D. Okla. Aug. 2, 2011) (citing *Robedeaux v. State*, 866 P.2d 417, 425, *cert. denied*, 513 U.S. 833 (1994)), *aff'd in part, re'd in part on other grounds sub nom. Dodd v. Trammell*, 753 F.3d 971 (10th Cir. 2013); *cf. Holtzer v. Davis,* No. 2:06-CV-169, 2009 WL 723881, at *1 (W.D. Mich. Mar. 11, 2009) (affirming conviction where testimony presented at trial was that luminol testing was a first line test for blood and could possibly detect the presence of blood).

Similarly, the confirmatory serological test used here (the Takayama hemochromogen test) to identify the biologic material as blood has been found reliable and admissible. *See, e.g., United States v. Williams*, No. 06-79, 2013 WL 4518215 at (*9 (D. Haw. Aug. 26, 2013) (rejecting *Daubert* challenge to the Takayama confirmatory test,

which "was developed between 1910 and 1912 and has become a standard confirmatory test for the presence of blood").

V. No Pretrial Hearing Is Required to Admit Scientific Testimony Under Rule 104

Under Rule 104(a) of the Federal Rules of Criminal Procedure, the proponent of expert testimony has the burden of demonstrating it satisfies Rule 702's requirements by a preponderance of the evidence. Fed. R. Evid. 702, advisory committee's note to 2000 amendment (citing *Bourjaily v. United States,* 483 U.S. 171 (1987)). This burden may be satisfied without an evidentiary hearing, as Rule 104(a) specifically provides that the court is not bound by evidence rules in making its determination; in other words, a defendant is not entitled to an evidentiary hearing under Rule 702 simply because he requests one.

In fact, where the proponent proffers sufficient evidence that the expert used reliable principles and methods, and the defendant presents no evidence to the contrary, the trial court should not hold an evidentiary hearing. *See, e.g., United States v. Eastman*, 645 F. App'x 476, 481 (6th Cir. 2016) (affirming admission of DNA testimony at trial without *Daubert* hearing where defendant "present[ed] no groundbreaking evidence that leads us to question" the reliability of DNA evidence); *United States v. Pettway*, No. 12-CR-103S, 2016 WL 6134493, at *2 (W.D.N.Y. Oct. 21, 2016) (denying *Daubert* hearing regarding DNA evidence based on defendant's failure to present any expert opinion or scientific evidence challenging reliability of STRmix); *United States v. Fell,* No. 5:01-CR-12-01, Document 914, at 9 (D. Vt. Sept. 19, 2016) (unpublished opinion) (refusing to hold evidentiary hearing on admissibility of DNA evidence because defense arguments went

to weight of evidence and did not "raise credible systemic concerns about the practice and use of DNA identification"); *United States v. McCluskey*, 954 F. Supp. 2d 1224, 1233-34 (D.N.M. 2013) (finding pretrial *Daubert* hearing on DNA not warranted because sufficient record to determine reliability); *see also, e.g., United States v. John*, 597 F.3d 263, 274–75 (5th Cir. 2010) (holding district court did not err in refusing to hold a *Daubert* hearing on fingerprint evidence because reliability of the technique had already been tested in the adversarial system); *Murray v. Marina Dist. Dev. Co.*, 311 F. App'x 521, 523 (3d Cir. 2008) (finding "no benefit in holding a *Daubert* hearing" where sufficient record to ascertain methodology and make reliability determination); *United States v. Crisp*, 324 F.3d 261, 268 (4th Cir. 2003) ("Under Daubert, a trial judge need not expend scarce judicial resources reexamining a familiar form of expertise every time opinion evidence is offered. In fact, if a given theory or technique is 'so firmly established as to have attained the status of scientific law,' then it need not be examined at all, but instead may properly be subject to judicial notice); *United States v. Nichols*, 169 F.3d 1255, 1263 (10th Cir. 1999) (finding trial court did not err in declining to hold a preliminary evidentiary hearing where the challenged evidence did not involve any new scientific theory or testing methodologies).

Finally, over 20 years ago, the Eighth Circuit held that district courts were not required to conduct *Daubert* hearings regarding the reliability of DNA analysis because they could take judicial notice of its reliability:

> Having considered all of Beasley's arguments, we conclude that the District Court did not abuse its discretion in admitting the government's evidence showing a "match" between the DNA in the hairs found in the rubber mask and Beasley's DNA. Moreover, we believe that the reliability of the PCR method of DNA analysis is sufficiently well established to permit the courts of this circuit to take judicial notice of it in future cases.

United States v. Beasley, 102 F.3d 1440, 1448 (8th Cir. 1996). Cf. *United States v. Beverly*, 369 F.3d 516, 528 (6th Cir. 2004) ("[t]he use of nuclear DNA analysis as a forensic tool has been found to be scientifically reliable by the scientific community for more than a decade").

RESPONSE

I. The Reliability Of DNA Testing Using Probabilistic Genotyping And Strmix

The specific issue raised by the defendant's motion is the reliability of the principles and methods used to analyze the DNA in this case. More specifically, the defendant's motion raises "concerns about the scientific validity of probabilistic genotyping algorithms" based primarily on one cited article. (R.119 at 7) In other words, the defendant has not challenged (1) the qualifications of Ms. Bakker; (2) whether her testimony would help the jury understand the evidence or to determine a fact in issue; (3) whether her expected testimony is based on sufficient facts or data; or (4) whether she has reliably applied the principles and methods to the facts of the case. Thus, the sole issue raised by the defendant's motion as to the DNA analysis is the reliability of probabilistic genotyping, STRmix, and the length analysis of alleles.

As noted earlier, there are four areas the courts have identified as particularly relevant to whether scientific principles and methods are reliable: (1) whether the

scientific technique can be and has been tested; (2) whether the theory or technique has been subjected to peer review and publication; (3) whether a particular technique has a known potential rate of error; and (4) whether the theory or technique is generally accepted in the relevant scientific community. In this case, application of each of those factors shows that the use of probabilistic genotyping and STRmix software is a reliable scientific methodology.

As the court addresses reliability of the methodology, it is helpful to realize that STRmix "represents an evolution in the process of DNA interpretation using existing accepted biological models and scientific principles rather than a completely new or novel approach to DNA analysis." *Pettway*, 2016 WL 6134493, at *2 (quoting John P. Simich, Ph.D., Director of the Erie County Forensic Laboratory).

A. Testing

Probabilistic genotyping using STRmix has been subjected to validation studies. First, the developer itself subjected STRmix to various different validation studies. *Muhammad,* 2018 WL 4927094, at *3; *Blash*, 2018 WL 4062322, at *7. The mathematics underlying the software involve a well-established method, and the development team performed the first 500 steps of the mathematics chain by hand, performed "true donor" and "false donor" tests, and tested STRmix against other software. *Muhammad,* 2018 WL 4927094, at *3; *see also Oldman,* No. 18-CR-0020, Document 227, at 12 (finding probabilistic genotyping process and STRmix software "are scientifically valid, proven, and tested").

Moreover, the FBI also conducted its own internal validation study prior to implementing STRMix. The validation study involved the use of over 300 mixtures of DNA from known contributors and around 200 non-contributor DNA samples that were analyzed against those mixtures. In total, the study included about 60,000 tests. The findings were published in the peer-reviewed journal, Forensic Science International: Genetics. *Blash*, 2018 WL 4062322, at *7. Other laboratories have also subjected STRmix to validation studies. *Muhammad,* 2018 WL 4927094, at *3.

Attached hereto are the published findings of several of these validation studies. *See, e.g., Scientific Working Group on DNA Analysis Methods (SWGDAM) Guidelines for the Validation of Probabilistic Genotyping Systems* (June 15, 2015) (Exhibit A); *Developmental validation of STRmix, expert software for the interpretation of forensic DNA profiles*, Forensic Science International: Genetics 23 (2016) 226-239 (Exhibit B); *Internal validation of STRmix for the interpretation of single source and mixed DNA profiles*, Forensic Science International: Genetics 29 (2017) 126-144 (Exhibit C); *Internal validation of STRmix – A multi laboratory response to PCAST,* Forensic Science International: Genetics 34 (2018) 11-14 (Exhibit D).

B. Peer Review and Publication

STRmix has been favorably reviewed in at least 19 peer-reviewed scientific journals, *Smith*, 2018 WL 4926977, at 8*, and peer-reviewed in more than 90 articles. *Pettway,* 2016 WL 6134493, at *2. Research regarding the reliability, validation, and underlying principles of STRMix has been published in a multitude of peer-reviewed journals such as *Forensic Science International, Australian Journal of Forensic Sciences, and Journal of Forensic Sciences*. *Blash*, 2018 WL 4062322, at *6. The National Institute for

Standards Technology presented scientific information on probabilistic genotyping, and SWGDAM published guidelines for probabilistic genotyping in June 2015; STRmix was presented to the New York Commission on Forensic Science, which adopted the DNA Subcommittee's recommendation to accept STRmix for casework. *Muhammad*, 2018 WL 4927094, at *4. The DNA Subcommittee consisted of scientists in the fields of molecular biology, population genetics, laboratory standards and quality assurance, and forensic science. *Bullard-Daniel*, 42 N.Y.S.3d at 723.

Additionally, the FBI has instituted several standard operating procedures and policies that govern the DNA Casework Unit's work, as well as the use of STRMix, and those policies and procedures align with the standards established by the International Organization for Standardization (ISO) and Quality Assurance Standards for Forensic DNA Testing Laboratories. *Blash*, 2018 WL 4062322, at *6. The FBI laboratories undergo accreditation by an outside accreditation body to ensure that the laboratories are satisfying ISO standards for testing laboratories, and the DNA Casework Unit specifically is audited against national DNA casework standards by other DNA experts in the field. *Blash*, 2018 WL 4062322, at *6.

C. Rate of Error

There is no evidence that STRmix's software has ever caused a false inclusion, and its developers only know of two errors causing false exclusions, both of which occurred during specific testing exercises rather than when used in an actual case. *Smith*, 2018 WL 4926977, at *8. In fact, STRmix has been subjected to "massive tests of false donors,

hundreds of millions," and the software had not made a "false positive" identification. *Muhammad*, 2018 WL 4927094, at *4.

D. Generally Accepted in Relevant Scientific Community

By 2018, STRmix was being used in at least 17 laboratories in the United States, including the laboratories of the FBI and the United States Army, and at least 65 additional laboratories have purchased the software are in the validation process for transitioning to its use. *Smith*, 2018 WL 4926977, at *8. Additionally, all but one of the states in Australia use STRmix, and there have been at least 40,000 cases processed in Australia without any discernable error. *Muhammad*, 2018 WL 4927094, at *4. To date, STRmix software is now being used by 43 labs in the United States, all 9 state and territory labs in Australia, and 11 labs elsewhere in the world. https://johnbuckleton. wordpress.com /strmix/.

E. No Evidentiary Hearing Is Warranted

Given that the reliability of probabilistic genotyping and STRmix has been established in numerous courts based on its extensive validation, favorable peer review, low error rate, and general acceptance in the relevant scientific community, the United States has met its burden of establishing the reliability of the methodology by a preponderance of the evidence. The defendant has offered no expert or other scientific evidence to support a challenge to the reliability of the methodology. Therefore, the Court should deny his motion without an evidentiary hearing.

Moreover, it would not be appropriate to allow the defendant the opportunity to cross-examine the government expert prior to trial as to the reliability of her conclusions,

as opposed to the reliability of the methodology. The defendant has already shown a general litigation strategy of using pre-trial hearings to cross-examine government witnesses on matters, such as the reliability of an expert's opinion, that should not be permitted until trial.

II. The Defendant's Request To Obtain The Source Code For Strmix

In his motion, the defendant requests that the Court order the United States to provide the source code for the STRmix software to the defendant for review by an unidentified "defense expert." This appears to be a discovery request. Prior to filing the motion, the defendant did not request the source code from the United States. On July 11, 2018, the United States informed the defendant in response to a prior request as to any commercial software programs used in the DNA testing in this case, that the United States used STRmix software and directed the defendant to its website at strmix.esr.cir.nz/.

The source code for STRmix commercial software is proprietary information that is not in the possession of the FBI or the United States. Therefore, the United States is unable to provide the defendant with the source code.[4] Regardless, for the numerous reasons set forth, *supra,* review of the proprietary source code is not necessary to find that using STRmix software to assist with probabilistic genotyping is a reliable methodology. *See., e.g., Blash,* 2018 WL 4062322, at *8 (finding STRmix scientifically valid,

4 The STRmix website provides information for defense legal teams on how to access STRmix software. *See* http://strmix.esr.cri.nz/assets/Uploads/Defence-Access-to-STRmix-April-2016.pdf.

reliable, and relevant despite defendant's argument that he needed access to its source code); http://strmix.esr.cri.nz/assets/Uploads/Defence-Access-to-STRmix-April-2016.pdf ("The developers consider that the STRmix™ software is best tested by examining the Extended Output for the compiled STRmix™ software, rather than the source code. The Extended Output of STRmix™ contains the intermediate steps of the STRmix™ interpretation process, allowing individual forensic laboratories, or experts for the defence, to verify the accuracy of STRmix™.")

III. Comparing Alleles By Length Is A Reliable Methodology

In his supplemental memorandum, the defendant requests that the Court bar the admission of the DNA test results simply because a more discriminating methodology, *i.e.,* next-generation sequencing (NGS), is being developed. As discussed, *infra,* the method used in this case of comparing the length of alleles is the standard method used by DNA forensic analysts. The United States has found no case that has found this method to be unreliable, and the defendant cites none in his supplemental memorandum. Moreover, the United States has found no case finding that NGS somehow renders prior DNA methodologies unreliable.[5] Again, the defendant cites to no such cases, but relies solely on citations to various literature. For example, no court has excluded fingerprint evidence as unreliable just because DNA evidence might be more discriminating.

[5] In fact, the United States has found no case finding that the use of NGS itself is reliable and admissible, nor has the defendant cited any. No doubt, if the United States attempted to introduce evidence of NGS, the defendant would argue (as he has regarding STRmix) that it has not yet been sufficiently validated. If the defendant's argument were accepted, no DNA test results would be admitted in courts throughout the United States.

Importantly, the literature cited by the defendant does not support his claim that "evaluating an allele by length alone is, in fact, not the most reliable way to interpret DNA." (Def. Memo. at 5) While NGS has the potential to be more *discriminating*, none of the literature suggests it is more *reliable* than prior methodologies. Moreover, where the likelihood ratios are already in the quintillions, as in this case, NGS offers little forensic improvement.

By way of analogy, a witness may identify a suspect that they already know by his height, weight, skin color, a scar on his face, a tattoo on his forehead, and a gold tooth. To say the witness was not as discriminating as he could be (*e.g.*, he could not number the hairs on the suspect's head), is not to question the reliability of the identification. None of the literature cited by the defendant supports excluding the DNA test results in this case, and no court has done so based on the ongoing development of "next-generation sequencing." The defendant's argument that the DNA results in this case should be excluded based on his unsupported claim that NGS is the only appropriate DNA testing method should be rejected.

IV. The Admissibility Of Luminol And Phenolphthalein Testing.

In his motion, the defendant also requests the Court to bar the admission of evidence of the results of the use of luminol or phenolphthalein testing as more prejudicial than probative under Rule 401, 402, and 403 of the Federal Rules of Evidence. The defendant first argues that, because these tests are not conclusive, they are irrelevant. This argument conflicts with the plain language of Rule 401(a), which defines

evidence as relevant if "it has any tendency to make a fact more or less probable than it would be without the evidence."

In other words, evidence need not be conclusive to be relevant; it only must make a fact of consequence more or less probable. The results of the luminol testing in this case make it more probable that blood was found in the defendant's apartment. Thus, the evidence is highly relevant. Here, the defendant "mistakenly equates a presumptive, *i.e.* inconclusive, scientific procedure with an unreliable, and therefore inadmissible, one. To be admissible, evidence need not be irrefutably conclusive of anything; it must only tend to make the existence of a particular fact of consequence more or less probable." *Dodd*, 2011 WL 3299101, at *60.

Moreover, in the context of the other evidence in this case, the results of these presumptive tests is highly relevant. First, where enough evidence remained for testing, the presumptive testing identifying blood was confirmed by later testing and corroborated by the results of DNA testing. Second, the identification, location, and pattern of the blood (including human hand prints) were consistent with the defendant's own later recorded statements as to what occurred in the apartment. Third, the evidence will show the defendant took extensive steps to clean the apartment after his alleged offense, thereby preventing confirmatory testing, but leaving trace amounts that were detected by the preliminary testing.

In 2010, the California Supreme Court addressed and rejected similar arguments that the results of presumptive blood testing are irrelevant and unfairly prejudicial:

> The circumstance that presumptive tests for blood on a jacket that might have been defendant's indicated the jacket might have had bloodstains in a pattern consistent with the murder in issue tends "in reason to prove or disprove any disputed fact that is of consequence to the determination of the action." The factors raised in defendant's challenge to this evidence—that the presumptive tests could not confirm the substance tested was human blood, that confirmatory tests failed to confirm the presence of blood, and that it is unknown when the jacket might have been exposed to the substance that created the positive results—do not mean the test results have no tendency in reason to establish that defendant shot Agent Cross. Those issues affect the probative weight of the evidence, not whether the test results meet the threshold requirement of relevancy. The trial court did not abuse its discretion in finding this evidence was relevant.
>
> Similarly, the trial court did not abuse its discretion by finding that the danger of undue prejudice, confusion of the issues, or misleading the jury did not substantially outweigh the evidence's probative value. " 'The "prejudice" referred to in Evidence Code section 352 applies to evidence which uniquely tends to evoke an emotional bias against the defendant as an individual and which has very little effect on the issues. In applying section 352, "prejudicial" is not synonymous with "damaging." ' " Evidence need not be excluded under this provision unless it "poses an intolerable ' "risk to the fairness of the proceedings or the reliability of the outcome." ' " The testimony regarding the presumptive blood tests had no particularly emotional component, nor did it consume an unjustified amount of time. Further, because the defense fully explored the limitations of the presumptive tests through cross-examination, there is no likelihood this evidence confused or misled the jury. The trial court did not err, and defendant's constitutional rights were not violated.

People v. Alexander, 235 P.3d 873, 924 (2010), *as modified on denial of reh'g* (Sept. 29, 2010); *see also Dyleski v. Grounds*, No. 12-CV-05336, 2016 WL 3194997, at *34 (N.D. Cal. June 9, 2016) (no error admitting testimony that a portion of the overcoat tested presumptively positive for blood, where it "responded fluorescently," because any potential issues went to the test's weight, not admissibility).

Other states have joined California in admitting evidence of presumptive blood tests. *See, e.g., Com. v. Hetzel*, 822 A.2d 747, 763 (Pa. Super. 2003) (phenolphthalein

testing); *State v. Canaan*, 964 P.2d 681, 694 (Kan. 1998) (luminol testing); *State v. Stenson*, 940 P.2d 1239, 1264 (Wash. 1997) (en banc) (phenolphthalein testing); *State v. Moseley*, 445 S.E.2d 906, 912 (N.C. 1994) (phenolphthalein testing); *Johnston v. State*, 497 So.2d 863, 870 (Fla. 1986) (luminol testing).

The non-controlling, minority state cases cited by the defendant present different factual situations, have been distinguished by their own courts, and do not counsel against admission of presumptive test results in this case. For example, in the Connecticut state case cited by the defendant, the police seized clothing and a pair of black shoes from the defendant's apartment; no blood was found on the clothing, but a stain was found on one of the soles of the defendant's shoes. *State v. Moody*, 573 A.2d 716, 722 (1990). The court, concerned that this isolated stain on the bottom of a shoe might only be animal blood, held that it was error to admit it into evidence. *Id.* This is far from the fact pattern presented here, where there is other substantial evidence tending to establish that the blood is the victim's blood, including DNA testing, the defendant's own statements, and the shape, pattern, and location of the stains revealed through the use of luminol.

In fact, Connecticut's own courts have distinguished *Moody* in precisely such a situation. For example, a court found that where the results of the presumptive testing of stains were also corroborated based on their shape, pattern, and location, "the facts in the present case are sufficiently distinguishable from those in *Moody* so as to render *Moody* inapplicable here." *State v. Downing*, 791 A.2d 649, 654 (Conn. App. 2002) (admitting evidence of presumptive blood tests). Moreover, multiple other courts have

likewise distinguished *Moody,* including the Connecticut Supreme Court itself. *See, e.g., State v. Grant,* 944 A.2d 947, 971 (Conn. 2008) (holding trial court properly admitted blood evidence because "*Moody* stands only for the proposition that, when the *sole* evidence that a substance was blood is the result of a presumptive testing method . . . the evidence is nonprobative") (emphasis in original); *State v. Jeffrey*, 601 A.2d 993, 998 (Conn. 1991) (declining to extend *Moody* and noting that "challenges to the evidence do not require its exclusion, however, because 'evidence need not exclude all other possibilities [to be relevant]; it is sufficient if it tends to support the conclusion, even to a slight degree.'"); *Weinberg v. Comm'r of Correction,* 962 A.2d 155, 166 (Conn. App. 2009) (finding that admission of evidence that presumptive blood found on knife seized from apartment "is easily distinguishable from *Moody*").

Here, the blood was not identified as a small stain on the bottom of a shoe; it was identified in significantly larger areas, including handprints, in the defendant's apartment. Some of the samples have been positively identified as blood (through confirmatory testing) with the victim's DNA located in those same samples. Moreover, unlike in *Moody,* the other evidence in the case, including the defendant's own statements, establish the presence of the victim's blood, making the presumptive test results tending to establish the location of the blood highly relevant.

Similarly, the Arkansas state cases cited by the defendant do not warrant exclusion of the presumptive testing in this case. In one case, the defendant wanted to offer into evidence the results of a *negative* luminol test to prove the absence of blood, but offered no evidence as to the likelihood of a false negative test. *Houston v. State*, 906

S.W.2d 286, 287 (Ark. 1995). The court held that the district court did not abuse its discretion in excluding the evidence of this novel use of luminol for lack of reliability (not relevance). *Id.* It also noted that "luminol test results are not relevant per se . . . *without* additional factors that relate that evidence to the crime . . . ." *Id.* (emphasis added). As noted above, there are many additional factors here that relate the luminol evidence to the defendant's alleged crime. Moreover, the Arkansas court's decision in *Brenk v. State,* 847 S.W.2d 1, 9 (Ark. 1993), was based on now-outdated science. The court was concerned with admitting preliminary testing because it "could not establish the blood type of the samples or connect the samples in any way with the victim." *Id.* With the advent of DNA testing, analysts do not conduct blood type testing to connect the samples to the victim; they use DNA testing. Here, such testing was used and confirms the presence of the victim's DNA in some of the same samples from the apartment that were presumptively identified as blood.

Finally, the Alaska case cited by the defendant simply affirmed the district court's decision to exclude the preliminary testing because no confirmatory tests whatsoever were conducted. *State v. Fukusaku*, 85 Haw. 462, 497, 946 P.2d 32, 67 (1997). Here, confirmatory and DNA tests established the presence of blood and the victim's DNA, respectively, in the defendant's apartment. Therefore, the evidence is not only relevant, but as found by the majority of the courts to address the issue, its probative value is not substantially outweighed by the danger of unfair prejudice. Fed. R. Evid. 403.

V. The Reliability Of Takayama Hemochromogen Testing

Interestingly, after arguing that presumptive tests should not be admissible without confirming blood tests (R.119 at 8-10), the defendant goes on to argue that confirmatory testing is unreliable.[6] The defendant does not present any expert opinion or scientific evidence, however, challenging the reliability of Takayama hemochromogen testing. Instead, the defendant claims that, even though the test has been a standard confirmatory test for blood for over 100 years, "nobody has ever taken the time to validate such a test as reliable." While he cites to articles that note that Takayama hemochromogen testing itself does not distinguish between human blood and blood of an animal, this does not indicate that the test is unreliable in identifying blood, only that it has limitations. Moreover, he fails to note that, when coupled with DNA testing, the test is capable of identifying a biological sample as human blood.

The Takayama hemochromogen test, which has long been in use to confirm the presence of blood, has been found reliable and admissible. *See, e.g.*, *Williams*, 2013 WL 4518215, at *9 (rejecting *Daubert* challenge). The defendant cites no case that supports his challenge to the reliability of the confirmatory testing.

In fact, the defendant in *Williams* advanced the same arguments as the defendant here, namely, that (1) the analyst failed to sufficiently document what she had done; and (2) the test had not been validated. *Id.* The district court rejected these arguments

[6] The defendant also suggests that there is insufficient evidence to show Ms. Bakker conducted the testing correctly, but as argued, *supra,* that is a matter for trial, not a pretrial hearing under Rule 702. *See Manpower*, 732 F.3d at 806 ("Reliability, however, is primarily a question of the validity of the methodology employed by an expert, not the quality of the data used in applying the methodology or the conclusions produced.").

because (1) the analyst recorded her observations in a report; and (2) given that the Takayama confirmatory test has been a standard confirmatory test for the presence of blood for over 100 years, the lack of recent validation is unsurprising. *Id.* Moreover, the court also noted that studies established that false positives using the Takayama test are unlikely. *Id.* Therefore, the court found the defendant's arguments went to the weight of the evidence, not its admissibility. *Id.* The same is true of the defendant's arguments here.

WHEREFORE, the United States of America respectfully requests that this Court deny the Defendant's Motion to Exclude DNA and Serology Test Results and Request for *Daubert* Hearing without an evidentiary hearing.

Respectfully submitted,

JOHN C. MILHISER
UNITED STATES ATTORNEY

s/ Eugene L. Miller
Eugene L. Miller, Bar No. IL 6209521
Assistant United States Attorney
201 S. Vine St., Suite 226
Urbana, IL 61802
Phone: 217/373-5875
Fax: 217-373-5891
eugene.miller@usdoj.gov

s/ James B. Nelson
James B. Nelson
Trial Attorney
Capital Case Section
United States Department of Justice
Washington, DC 20004
1331 F St. NW, Room 625
Washington, DC 20004
Phone: 202/598-2972
james.nelson@usdoj.gov

s/ Bryan D. Freres
Bryan D. Freres
Assistant United States Attorney
201 S. Vine St., Suite 226
Urbana, IL 61802
Phone: 217/373-5875; Fax: 217-373-5891
bryan.freres@usdoj.gov

CERTIFICATE OF SERVICE

I hereby certify that on February 6, 2019, I electronically filed the foregoing with the Clerk of the Court using the CM/ECF system, which will send notification of the filing to all CM/ECF participants.

s/ Eugene L. Miller
Eugene L. Miller, Bar No. IL 6209521
Assistant United States Attorney
201 S. Vine St., Suite 226
Urbana, IL 61802
Phone: 217/373-5875
Fax: 217-373-5891
eugene.miller@usdoj.gov

Scientific Working Group on DNA Analysis Methods

Guidelines for the Validation of Probabilistic Genotyping Systems

Table of Contents

SWGDAM Guidelinesfor the Validation of Probabilistic Genotyping Systems

The Scientific Working Group on DNA Analysis Methods, better known by its acronym of SWGDAM, is a group of approximately 50 scientists representing Federal, State, and Local forensic DNA laboratories in the United States and Canada. During meetings, which are held twice a year, Committees discuss topics of interest to the forensic DNA community and often develop documents to provide direction and guidance for the community. In some instances, an Ad Hoc Working Group may be empanelled to address a particular topic outside of the routine SWGDAM January/July meeting schedule.These Guidelines,drafted by the SWGDAM Ad Hoc Working Group on Probabilistic Genotyping,wereapproved by the SWGDAM Executive Board for public comment in March 2015.Following the public comment period, the Ad Hoc Working Group forwarded the Final Guidelines to the SWGDAM Executive Board and they were approved for posting on the SWGDAM web site on June 15, 2015.

Guidance is provided herein for the validation of probabilistic genotyping software used for the analysis of autosomal short tandem repeat (STR) typing results. These guidelines are not

EXHIBIT A

intended to be applied retroactively. It is anticipated that they will evolve with future developments in probabilistic genotyping systems.

Introduction

Probabilistic genotyping refers to the use of biological modeling, statistical theory, computer algorithms, and probability distributions to calculate likelihood ratios (LRs) and/or infer genotypes for the DNA typing results of forensic samples ("forensic DNA typing results"). Human interpretation and review is required for the interpretation of forensic DNA typing results in accordance with the FBI Director's Quality Assurance Standards for Forensic DNA Testing Laboratories[1]. Probabilistic genotyping is a tool to assist the DNA analyst in the interpretation of forensic DNA typing results. Probabilistic genotyping is not intended to replace the human evaluation of the forensic DNA typing results or the human review of the output prior to reporting.

A probabilistic genotyping system is comprised of software, or software and hardware, with analytical and statistical functions that entail complex formulae and algorithms. Particularly useful for low-level DNA samples (i.e., those in which the quantity of DNA for individuals is such that stochastic effects may be observed) and complex mixtures(i.e., multi-contributor samples, particularly those exhibiting allele sharing and/or stochastic effects), probabilistic genotyping approaches can reduce subjectivity in the analysis of DNA typing results.Historical methods of mixture interpretation consider all interpreted genotype combinations to be equally probable, whereas probabilistic approaches provide a statistical weighting to the different genotype combinations. Probabilistic genotyping does not utilize a stochastic threshold. Instead, it incorporates a probability of alleles dropping out or in.In making use of more genotyping information when performing statistical calculations and evaluating potential DNA contributors, probabilistic genotyping enhances the ability to distinguish true contributors and non-contributors.A higher LR is typically obtained when evaluating a person of interest (POI) who is a true contributor to the evidence profile, and a lower LR is typically obtained when the POI is not a true contributor. While the absence of an allele or the presence of additional allele(s)

[1] Probabilistic genotyping is to be distinguished from an Expert System.An Expert System, if NDIS approved and properly validated in accordance with the QAS, may only be used by a laboratory on database, known or casework reference samples to replace the manual review in accordance with the QAS and NDIS Operational Procedures. Expert Systems are not approved for use on forensic or forensic mixture DNA samples.

relative to a reference sample may support an exclusion, probabilistic genotyping approaches allow inclusion and exclusion hypotheses to be consideredby calculating a LR in which allele drop-out and drop-in may be incorporated.

The use of a likelihood ratio as a reporting statistic for probabilistic genotypingdiffers substantially from binary statistics such as the combined probability of exclusion. Prior to validating a probabilistic genotyping system, the laboratory should ensure that it possesses the appropriate foundational knowledge in the calculation and interpretation of likelihood ratios. Laboratories should also be aware of the features and limitations of various probabilistic genotyping programs and the impact that those items will have on the validation process. Depending on the performance characteristics of the software, prerequisite studies may be required to, for example, establish parameters forallele drop-out and drop-in, stutter expectations, peak height variation, and the number of contributors to a mixture.Each laboratory seeking to evaluate a probabilistic genotyping system must determine which validation studies are relevant to the methodology, in the context of its application, to demonstrate the reliability of the system and any potential limitations. The laboratory must determine the number of samples required to satisfy each guideline and may determine that a study is not necessary. Some studies described herein may also be suitable for evaluating material modifications to existing procedures.

Background

Please refer to the SWGDAM Validation Guidelines for DNA Analysis Methods andthe FBI Quality Assurance Standards for Forensic DNA Testing Laboratories and for DNA Databasing Laboratories (QAS) for general background information regarding validation and definition of terms.

Probabilistic genotyping may generate a number of possible genotype combinations for a given profile, where some genotypes may be assigned more weight than others. Allele drop-in and drop-out probabilities may be used in the determination of the weights associated with each of the possible genotypes. There are two main approaches to probabilistic genotyping: the semi-continuous method and fully continuous method. The semi-continuous method focuses only on the alleles present in the profile and addresses all possible genotype combinations of the

observed alleles in conjunction with a probability of drop-out and drop-in. Analysis parameters such as peak height variation, mixture ratios and stutter percentages are not typically utilized by semi-continuous software systems, although these elements may be considered during the initial manual evaluation of the data. The fully continuous method generally utilizesmore of the biological information in the profile, such as peak heights, stutter percentages and mixture ratios. The weighting of genotype combinations as more or less probable may be inferred from the data through methods such as Markov Chain Monte Carlo (MCMC) samplings from probability distributions.

The analyst will need to employ some level of interpretation before using the software to perform the calculations and should visually interpret allelic and non-allelic peaks and other characteristics of the DNA typing results, as necessitated by the software. For example, the analyst may be required to estimate and use a specific number of contributors in a statistical calculation when interpreting a DNA mixture, or to assess whether typing results should be interpreted or not based on quality.

Forensic DNA typing results interpreted by a DNA analyst using probabilistic genotyping software may be eligible for CODIS entry and upload to NDIS in accordance with the NDIS Operational Procedures if the probabilistic genotyping software has been properly validated pursuant to the QAS and these Guidelines.

1. **Validation of Probabilistic Genotyping Systems**
   - 1.1. The laboratory shall validate a probabilistic genotyping system prior to usage for forensic applications.
   - 1.2. The laboratory shall document all validation studies in accordance with the FBI Quality Assurance Standards for Forensic DNA Testing Laboratories.
   - 1.3. The laboratory should document or have access to documentation that explains how the software performs its operations and activities, to include the methods of analysis and statistical formulae, the data to be entered in the system, the operations performed by each portion of the user interface, the workflow of the system, and the system reports or

other outputs. This information enables the laboratory to identify aspects of the system that should be evaluated through validation studies.

2. **System control**

2.1. The laboratory should verify that the software is installed on computers suited to run the software, that the system has been properly installed, and that the configurations are correct.

2.2. The laboratory should, where possible, ensure the following system control measures are in effect:

2.2.1. Every software release should have a unique version number. This version number should be referenced in any validation documentation or published results.

2.2.2. Appropriate security protection to ensure only authorized users can access the software and data.

2.2.3. Audit trails to track changes to system data and/or verification of system settings in place each time a calculation is run.

2.2.4. User-level security to ensure that system users only perform authorized actions.

3. **Developmental Validation**

Developmental validation of a probabilistic genotyping system is the acquisition of test data to verify the functionality of the system, the accuracy of statistical calculations and other results, the appropriateness of analytical and statistical parameters, and the determination of limitations. Developmental validation may be conducted by the manufacturer/developer of the application or the testing laboratory. Developmental validation should also demonstrate any known or potential limitations of the system.

3.1. The underlying scientific principle(s) of the probabilistic genotyping methods and characteristics of the software should be published in a peer-reviewed scientific journal. The underlying scientific principles of probabilistic genotyping include, but are not limited to, modeling of stutter, allelic drop-in and drop-out, Bayesian prior assumptions such as allele probabilities, and statistical formulae used in the calculation and algorithms.

3.2. Developmental validation should address, where applicable, the following:

3.2.1. Sensitivity – Studies should assess the ability of the system to reliably determine the presence of a contributor's(s') DNA over a broad variety of evidentiary typing results (to include mixtures and low-level DNA quantities). This should be evaluated using various sample types (e.g., different numbers of contributors, mixture proportions, and template quantities).

3.2.1.1. Sensitivity studies should demonstrate the potential for Type I errors (i.e., incorrect rejection of a true hypothesis), in which, for example, a contributor fails to yield a LR greater than 1 and thus his/her presence in the mixture is not supported.

3.2.1.2. Sensitivity studies should demonstrate the range of LR values that can be expected for contributors.

3.2.2. Specificity – Studies should evaluate the ability of the system to provide reliable results for non-contributors over a broad variety of evidentiary typing results (to include mixtures and low-level DNA quantities). This should be evaluated using various sample types (e.g., different numbers of contributors, mixture proportions, and template quantities).

3.2.2.1. Specificity studies should demonstrate the potential for Type II errors (i.e., failure to reject a false hypothesis), in which, for example, a non-contributor yields a LR greater than 1 and thus his/her presence in the mixture is supported.

3.2.2.2. Specificity studies should demonstrate the range of LR values that can be expected for non-contributors.

3.2.3. Precision – Studies should evaluate the variation in Likelihood Ratios calculated from repeated software analyses of the same input data. This should be evaluated using various sample types (e.g., different numbers of contributors, mixture proportions, and template quantities).

3.2.3.1. Some probabilistic genotyping approaches may not produce the same LR from repeat analyses. Where applicable, these studies should therefore demonstrate the range of LR values that can be expected from

multiple analyses of the same data and are the basis for establishing an acceptable amount of variation in LRs.

3.2.3.2. Any parameter settings (e.g., iterations of the MCMC)that can reduce variability should be evaluated. For example, for some complex mixtures (e.g., partial profiles with more than three contributors), increasing the number of MCMC iterations can reduce variation in the likelihood ratio.

3.2.4. Case-type Samples – Studies should assess a range of data types exhibiting features that are representative of those typically encountered by testing laboratories. These features include those derived from mixtures and single-source samples, such as stutter, masked/shared alleles, differential and preferential amplification, degradation and inhibition.

3.2.4.1. These studies should demonstrate sample and/or data types that can be reliably evaluated using the probabilistic genotyping system.

3.2.5. Control Samples – If the software is designed to assess controls, studies should evaluate whether correct results are obtained with control samples.

3.2.6. Accuracy – Studies should assess the accuracy of the calculations performed by the system, as well as allele designation functions, where applicable.

3.2.6.1. These studies should include the comparison of the results produced by the probabilistic genotyping software to manual calculations, or results produced with an alternate software program or application, to aid in assessing accuracy of results generated by the probabilistic genotyping system. Calculations of some profiles (e.g., complex mixtures), however, may not be replicable outside of the probabilistic genotyping system.

3.2.6.2. If the software uses raw data files from a genetic analyzer as input data, the peak calling, sizing and allele designation functions should be compared to the results of another software system to assess accuracy. Allele designations should also be compared to known genotypes where available.

4. **Internal Validation**

Internal validation of a probabilistic genotyping software system is the accumulation of test data within the laboratory to demonstrate that the established parameters, software settings, formulae, algorithms and functions perform as expected. In accordance with the QAS, internal validation data may be shared by all locations in a multi-laboratory system.

Depending on the features and capabilities of the probabilistic genotyping system, some DNA typing results may or may not be determined to be suitable for such analysis. To identify data features (e.g., minimum quality requirements, number of contributors) that render a profile appropriate or inappropriate for probabilistic genotyping, the laboratory should test data across a range of characteristics that are representative of those typically encountered by the testing laboratory. Data should be selected to test the system's capabilities and to identify its limitations. In particular, complex mixtures and low-level contributors should be evaluated thoroughly during internal validation, as the data from such samples generally help to define the software's limitations, as well as sample and/or data types which may potentially not be suitable for computer analysis. In addition, some exclusions may be evident without the aid of probabilistic software.

If conducted within the same laboratory, developmental validation studies may satisfy some of the elements of the internal validation guidelines.

4.1. The laboratory should test the system using representative data generated in-house with the amplification kit, detection instrumentation and analysis software used for casework. Additionally, some studies may be conducted by using artificially created or altered input files to further assess the capabilities and limitations of the software. Internal validation should address, where applicable to the software being evaluated:

4.1.1. Specimens with known contributors, as well as case-type specimens that may include unknown contributors.

4.1.2. Hypothesis testing with contributors and non-contributors

4.1.2.1. The laboratory should evaluate more than one set of hypotheses for individual evidentiary profiles to aid in the development of policies regarding the formulation of hypotheses. For example, if there are two persons of interest, they may be evaluated as co-contributors and, alternatively, as each contributing with an unknown individual. The hypotheses used for evaluation of casework profiles can have a significant impact on the results obtained.

4.1.3. Variable DNA typing conditions (e.g., any variations in the amplification and/or electrophoresis parameters used by the laboratory to increase or decrease the detection of alleles and/or artifacts)

4.1.4. Allelic peak height, to include off-scale peaks

4.1.5. Single-source specimens

4.1.6. Mixed specimens

4.1.6.1. Various contributor ratios (e.g., 1:1 through 1:20, 2:2:1, 4:2:1, 3:1:1, etc.)

4.1.6.2. Various total DNA template quantities

4.1.6.3. Various numbers of contributors. The number of contributors evaluated should be based on the laboratory'sintended use of the software.A range of contributor numbers should be evaluated in order to define the limitations of the software.

4.1.6.4. If the number of contributors is input by the analyst, both correct and incorrect values (i.e., over- and under-estimating) should be tested.

4.1.6.5. Sharing of alleles among contributors

4.1.7. Partial profiles, to include the following:

4.1.7.1. Allele and locus drop-out

4.1.7.2. DNA degradation

4.1.7.3. Inhibition

4.1.8. Allele drop-in

4.1.9. Forward and reverse stutter

4.1.10. Intra-locus peak height variation

4.1.11. Inter-locus peak height variation

4.1.12. For probabilistic genotyping systems that require in-house parameters to be established, the internal validation tests should be performed using those same parameters. The data set used to establish the parameters should be different from the data set used to validate the software using those parameters.

4.1.13. Sensitivity, specificity and precision, as described for Developmental Validation

4.1.14. Additional challenge testing (e.g., the inclusion of non-allelic peaks such as bleed-through and spikes in the typing results)

4.2. Laboratories with existing interpretation procedures should compare the results of probabilistic genotyping and of manual interpretation of the same data, notwithstanding the fact that probabilistic genotyping is inherently different from and not directly comparable to binary interpretation.The weights of evidence that are generated by these two approaches are based on different assumptions, thresholds and formulae. However, such a comparison should be conducted and evaluated for general consistency.

4.2.1. The laboratory should determine whether the results produced by the probabilistic genotyping software are intuitive and consistent with expectations based on non-probabilistic mixture analysis methods.

4.2.1.1. Generally, known specimens that are included based on non-probabilistic analyses would be expected to also be included based on probabilistic genotyping.

4.2.1.2. For single-source specimens with high quality results, genotypes derived from non-probabilistic analyses of profiles above the stochastic threshold should be in complete concordance with the results of probabilistic methods.

4.2.1.3. Generally, as the analyst's ability to deconvolute a complex mixture decreases, so do the weightings of individual genotypes within a set determined by the software.

5. **Modification to Software**

   Modification to probabilistic genotyping software shall be addressed in accordance with the QAS.

   5.1. Modification to the system such as a hardware or software upgrade that does not impact interpretation or analysis of the typing results or the statistical analysis shall require a performance check prior to implementation.

   5.2. A significant change(s) to the software, defined as that which may impact interpretation or the analytical process, shall require validation prior to implementation.

   5.3. Data used during the initial validation may be re-evaluated as a performance check or for subsequent validation assessment. The laboratory must determine the number and type of samples required to establish acceptable performance in consideration of the software modification.

References and Suggested Readings

Federal Bureau of Investigation, (2015) *NDIS Operational Procedures Manual*, available at http://www.fbi.gov/about-us/lab/biometric-analysis/codis/ndis-procedures-manual.

Federal Bureau of Investigation (2011) *Quality Assurance Standards for Forensic DNA Testing Laboratories,* available at http://www.fbi.gov/about-us/lab/codis/qas-standards-for-forensic-dna-testing-laboratories-effective-9-1-2011

Gill, P. et al. (2012)*DNA Commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods*, Forensic Science International Genetics 6(6): 679-688.

Kelly. H. et al. (2014)*A comparison of statistical models for the analysis of complex forensic DNA profiles*, Science &Justice 54(1): 66-70.

Scientific Working Group on DNA Analysis Methods (2010) *Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories*, available at http://www.swgdam.org/Interpretation_Guidelines_January_2010.pdf.

Scientific Working Group on DNA Analysis Methods (2012) *Validation Guidelines for DNA Analysis Methods*, available at http://www.swgdam.org/SWGDAM_Validation_Guidelines_APPROVED_Dec_2012.pdf .

Scientific Working Group on DNA Analysis Methods (2014) *Guidelines for STR Enhanced Detection Methods*, available at http://swgdam.org/SWGDAM%20Guidelines%20for%20STR%20Enhanced%20Detection%20Methods%20FINAL%20100614.pdf.

Steele, C. D. and Balding, D. J. (2014) *Statistical Evaluation of Forensic DNA Profile Evidence,* Annu. Rev. Stat. Appl. 1:361-384.

Forensic Science International: Genetics 23 (2016) 226–239

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles

Jo-Anne Bright[a,*], Duncan Taylor[b,c], Catherine McGovern[a], Stuart Cooper[a], Laura Russell[a], Damien Abarno[b], John Buckleton[a]

[a] ESR, Private Bag 92021, Auckland 1142, New Zealand
[b] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[c] School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia

ARTICLE INFO

Article history:
Received 25 January 2016
Received in revised form 9 May 2016
Accepted 10 May 2016
Available online 12 May 2016

Keywords:
DNA mixtures
Probabilistic genotyping
Continuous method
Validation
STRmix

ABSTRACT

In 2015 the Scientific Working Group on DNA Analysis Methods published the SWGDAM Guidelines for the Validation of Probabilistic Genotyping Systems [1]. STRmix™ is probabilistic genotyping software that employs a continuous model of DNA profile interpretation. This paper describes the developmental validation activities of STRmix™ following the SWGDAM guidelines. It addresses the underlying scientific principles, and the performance of the models with respect to sensitivity, specificity and precision and results of interpretation of casework type samples. This work demonstrates that STRmix™ is suitable for its intended use for the interpretation of single source and mixed DNA profiles.

1. Introduction

The dominant method for forensic DNA analysis involves the amplification of short tandem repeats using PCR. Amplified products are separated via capillary electrophoresis (CE). Fluorescently labelled tags are used to colour code the markers or loci. A laser excites the primer tags as the different lengths of DNA travel through the capillaries of the electrophoresis instrument, which emit a signal that is recorded. The signals are visualised as peaks in a graph of fluorescence versus time, known as an electropherogram (epg). The height of the peaks is approximately proportional to the initial amount of DNA template and is measured in relative fluorescent units (rfu). In this way height can be used as an approximation of DNA quantity or template.

Manual techniques for DNA profile interpretation are heuristically based and may be difficult to apply consistently between laboratories, individual scientists and even a single scientist. Variable decisions often occur early in the manual interpretation process and can even occur at allele assignment. Divergence in these choices can have significant downstream consequences [2,3]. Phenomena such as stutter (artifactual amplicons produced as a consequence of the PCR process), allelic drop-in (the presence of low amounts of extraneous DNA) and dropout (which is a consequence of low template and/or degraded DNA and results in partial DNA profiles) [4] are all considered at profile analysis and interpretation. Interpretation of DNA profiles is also complicated by mixed samples (the presence of DNA from more than one individual).

The interpretation of an epg or evidentiary DNA profile should initially be undertaken 'blind'; in isolation of the person of interest's (POI) reference DNA profile, and where possible avoiding contextual effects [5,6]. Comparison with reference profiles of any POI or other relevant evidentiary profiles is undertaken after profile interpretation. Traditionally there are three primary conclusions that can be drawn: *cannot exclude* (or *inclusion*), *can exclude*, or *inconclusive* which is sometimes also called *uninterpretable* [7]. It is desirable when an association is reported (cannot exclude or inclusion) to present the evidence with the associated statistical weight [7]. When the evidence profile originates from a single individual, the weight of evidence can be presented as a match probability. This is an assignment of the probability that a random person might match the crime scene stain given the observation of that crime stain profile. A favoured alternative to the match probability, which can be extended to use for mixed DNA profiles, is the likelihood ratio (*LR*). The *LR* considers the probability of obtaining the evidence profile(s) given two competing propositions, usually aligned with the prosecution case and defence case. The *LR* is used throughout Australasia and the UK and is used in some laboratories within the US and Europe for criminal forensic work to express the weight of evidence. The *LR* is

* Corresponding author.
E-mail addresses: jodashanne@gmail.com, jo.bright@esr.cri.nz (J.-A. Bright).

http://dx.doi.org/10.1016/j.fsigen.2016.05.007
1872-4973/

accepted to be the most relevant and powerful statistic to calculate the weight of the evidence and is the only method recommended by the International Society for Forensic Genetics (ISFG) for ambiguous profiles [8]. Ambiguous profiles include all mixtures and single source profiles where dropout and drop-in are a consideration.

Known shortcomings of traditional methods of DNA profile interpretation have led to the development of improved models that factor in the probability of dropout [9–13]. The drop model (also known as the semi-continuous method) can optionally incorporate a probability for dropout, Pr(D), and/or a probability for drop-in, Pr(C). Semi-continuous methods do not use peak heights when generating possible genotype sets and do not model artifacts such as stutter. Continuous methods make assumptions about the underlying behaviour of peak heights across all profiles to evaluate the probability of a set of peak heights in a given profile. These methods are designed to be used in expert systems and reduce the requirement for the manual assignment of peaks as allelic within evidence profiles, and hence reduce the opportunity for inconsistency in interpretation to occur. The calculations are sufficiently complex that software is needed. STRmix™ is one such continuous method that employs a fully continuous approach for DNA profile interpretation (http://strmix.esr.cri.nz/ [14]).

In 2015 the Scientific Working Group on DNA Analysis Methods published the SWGDAM Guidelines for the Validation of Probabilistic Genotyping Systems [1]. The developmental validation of a probabilistic genotyping system has been described by SWGDAM as "the acquisition of test data to verify the functionality of the system, the accuracy of statistical calculations and other results, the appropriateness of analytical and statistical parameters, and the determination of limitations" [1].

The developmental validation of STRmix™ was initially undertaken in 2012 following the requirements outlined within the FBI Quality Assurance Standards [15] by analysts at Forensic Science South Australia (FSSA) and the Institute of Environmental Science and Research Limited (ESR; http://www.esr.cri.nz/). FSSA is the South Australian State Forensic Science Laboratory and is accredited by the National Association of Testing Authorities, Australia. ESR is the New Zealand Government Crown Research Institute that undertakes forensic services for the NZ Police. ESR forensic DNA laboratories are accredited by the Laboratory Accreditation Board of the American Society of Crime Laboratory Directors (ASCLD/LAB) under the International Testing Program (ISO 17025).

Within this paper we describe the developmental validation activities undertaken for STRmix™ following the SWGDAM recommendations [1]. Each of the guidelines is discussed in turn under their recommendation number.

1.1. Guideline 3.1 Publication of underlying scientific principles

All significant portions of the statistical algorithms and underlying scientific principles behind STRmix™ have been published in peer reviewed scientific literature. Within Table 1 we provide a summary of these models and algorithms and their references aligned with the software version in which they were introduced.

STRmix™ uses the quantitative information from an electropherogram (epg) such as peak heights (O), to calculate the probability of the profile given all possible genotype combinations (S_j). A value, or weight (w_i), is assigned to the normalised probability density $p(O|S_j)$. STRmix™ assigns a relative weight to the probability of the epg given each possible genotype combination at a locus. The weights across all combinations at that locus are normalised so that they sum to one. Therefore, a single unambiguous genotype combination at any locus would be assigned a weight of one.

STRmix™ describes the fluorescence observed in one or more epgs using a number of models that describe various properties of DNA profile behaviour. These are described as mass parameters and include a template for each contributor, a locus specific amplification efficiency for each locus, a replication efficiency for each PCR replicate, and a degradation for each contributor. This biological model is described in Bright et al. [16]. Profile degradation is modelled as exponential [17,18]. Drop-in is optionally modelled as a gamma distribution following Puch-Solis [19]. In addition, STRmix™ employs a per allele stutter model, the parameters of which are based on empirical data [16,20,21].

Posterior distributions of mass parameters are sampled from using Markov chain Monte Carlo (MCMC). In general, MCMC is a numerical method used, in this case, to approximate an integral (typically multi-dimensional) of the observed data across all parameters. MCMC methods sample from the posterior distribution of the desired integral. It does so by using Markov chains that

Table 1
A summary of the scientific principles, the STRmix™ version in which they were introduced and their publications.

Algorithms, scientific principles and methods	Version introduced	Reference
Allele and stutter peak height variability as separate constants within the MCMC	V2.0	[14]
Peak height variability as random variables within the MCMC	V2.3	[31]
Model for calibrating laboratory peak height variability	V2.0	[31]
Application of a Gaussian random walk to the MCMC process	V2.3	Described within this paper
Modelling of back stutter by regressing stutter ratio against allelic designation	V2.0	[16,20,32,33]
Modelling of back stutter by regressing stutter ratio against *LUS*	V2.3	[16,20,21,33]
Modelling of forward stutter	V2.4	[34]
Modelling of allelic drop-in using a simple exponential or uniform distribution	V2.0	[14]
Modelling of allelic drop-in using a Gamma distribution	V2.3	[19]
Modelling of degradation and dropout	V2.0	[17]
Modelling of the uncertainties in the allele frequencies using the HPD	V2.0	[30]
Modelling of the uncertainties in the MCMC	V2.3	[29,30,35]
Database searching of mixed DNA profiles	V2.0	[28]
Familial searching of mixed DNA profiles	V2.3	[26]
Relatives as alternate contributors under the defence proposition	V2.3	[26]
Modelling expected stutter peak heights in saturated data	V2.3	[34]
Taking into account the 'factor of two' in *LR* calculations	V2.3	[36]
Model for incorporating prior beliefs in mixture proportions	V2.3	[37]

have the posterior distribution as their equilibrium distribution. These chains 'walk' around in a memoryless fashion using an acceptance-rejection criterion to determine whether to take a step or not. At each step that the chain accepts the integrand value, it is counted towards the integral. At each step that the chain rejects the integrand value at that proposed point, the current point is counted towards the integral. The rejection-acceptance rule used within STRmix™ is called the Metropolis-Hastings algorithm [22,23]. The chain will then propose new steps in its search for a state that provides a reasonably high contribution to the integral until it finds a state which it will accept and move to. The statistical algorithms within STRmix™ are described in Taylor et al. [14].

STRmix™ does not use the reference profiles during profile deconvolution unless a reference from a known contributor is available (for example the complainant's DNA on their intimate samples collected as part of an investigation into a sexual assault). Where a reference profile is available from a person of interest (POI) a likelihood ratio may be calculated. It is the ratio of the probability of the observed crime stain (O) given each of two competing hypotheses, H_p and H_d, and given all the available information, I. Mathematically, we express this as:

$$LR = \frac{\Pr(O|H_p, I)}{\Pr(O|H_d, I)}$$

The likelihood ratio is calculated in STRmix™ incorporating values for F_{ST} (theta) using the subpopulation model of Balding and Nichols in 1994 [24], referred to as recommendation 4.2 in the 1996 National Research Council report (NRCII) [25]. As a continuous extension to the classic incorportation of a theta value (which is typically a fixed value) STRmix™ can consider a distribution for theta. Propositions within STRmix™ are flexible. The defence proposition aligns with exclusion of the person of interest and typically considers an unknown, unrelated individual within a selected population. Where appropriate, alternate propositions are calculated under the defence propositions such as a sibling, parent, child or cousin of the person of interest [26]. Additionally STRmix™ can provide an *LR* based on the unifying theory. This is where rather than specifying either an unrelated individual or a nominated relative (sibling, parent etc.) under the defence propositon, all members of the population, including possible relatives of the POI can be considered by taking into account their prior probabilities based on population properties.

If one or more contributors is known to be present (i.e. conceded by both parties) then this information can be provided to STRmix™ at the deconvolution stage in order to assist in the deconvolution of the remaining questioned contributors. This assumption of a known contributor is then carried forward to the *LR* calculation. If a reference profile is not available from a person of interest, the profile may be compared directly with a database of known individuals [28] to identify investigative leads.

STRmix™ uses the highest posterior density (HPD) method for calculating an *LR* distribution, from which a quantile can then be chosen in order to report a bound of the probability density distribution [29,30]. Within STRmix™ versions 2.3 onwards, the variability due to MCMC, the sampling variation inherent in generating allele frequency databases and the variability in F_{ST} (theta) can be estimated.

1.2. Guideline 3.2 Sensitivity and specificity studies

With respect to interpretation methods, sensitivity is defined as the ability of the software to reliably resolve the DNA profile of known contributors within a mixed DNA profile for a range of starting DNA template. The log(*LR*) for known contributors (H_p true) should be high and should trend to 0 as less information is present within the profile. Information includes the amount of DNA from the contributor of interest, conditioning profiles (for example the victim's profile on intimate samples), PCR replicates and decreasing numbers of contributors. Specificity is defined as ability of the software to reliably exclude known non contributors (H_d true) within a mixed DNA profile for a range of starting DNA template. The *LR* should trend upwards to neutral as less information is present within the profile. This is shown diagrammatically in Fig. 1.

Specificity and sensitivity within STRmix™ were tested by calculating the *LR* for a number of GlobalFiler™ mixtures for both

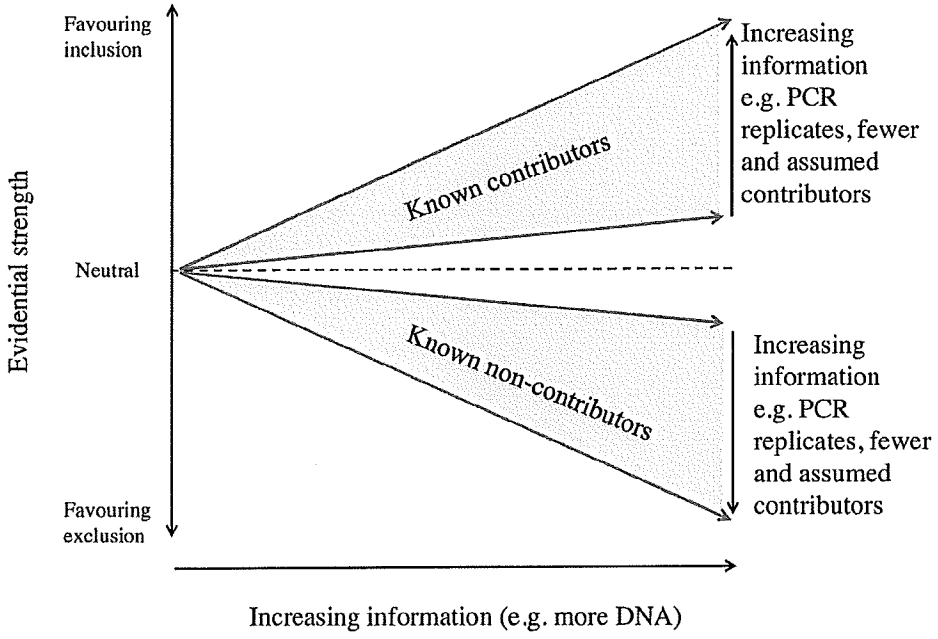


Fig. 1. A diagram showing the desired performance of a method of mixture interpretation.

Table 2
A summary of the experimental set up.

Sample	Mixture proportions for contributor				Total DNA added to PCR (pg)
	One	Two	Three	Four	
1–3	0.50	0.50	–	–	400,200,50
4–6	0.33	0.67	–	–	
7–9	0.20	0.80	–	–	
10–11	0.17	0.83	–	–	
13–15	0.09	0.91	–	–	
16–18	0.33	0.33	0.33	–	
19–21	0.50	0.33	0.17	–	
22–26	0.25	0.25	0.25	0.25	400,200,50,20,10
27–31	0.40	0.30	0.20	0.10	

known contributors and known non-contributors [38]. Two, three and four contributor mixtures were constructed in varying proportions and amplified with varying amounts of template DNA as described in Table 2.

Each sample was amplified in triplicate giving a total of 93 samples. Profiles were interpreted using STRmix™ v1.08 and *LR*s calculated for the known contributors and 186 non contributors. The propositions considered were:

Hp: The DNA originated from the person of interest and *N*-1 unknown contributors

Hd: The DNA originated from *N* unknown individuals

Where *N* was the number of contributors within the profile.

The plots of $\log_{10}(LR)$ versus DNA in the PCR (pg) produced for these comparisons are reproduced in Figs. 2–6 . The *LR*s produced from comparisons to known contributors (sensitivity tests) are signified by a blue point and those produced from comparisons to known non-contributors (specificity tests) are signified by a red point. A minimum value for $\log_{10}(LR)$ of −30 was used, and any *LR*s obtained that fell below this were given the value of −30. The lines on figures are given only as a visual indication of trends in the scattered results. The polygons seen give a visual indication of the spread of the *LR*s.

The plots in Figs. 2–6[1] clearly demonstrate the sensitivity of STRmix™ for these mixtures by inspection of the spread of blue points. They show the range of expected *LR* values for contributors given the amount of input DNA (guideline 3.2.1.2). Type I errors (incorrect rejection of a true hypothesis) are clearly identified as blue points *below* the horizontal line of $\log_{10}(LR)=0$. As expected, this is dependent on the amount of DNA per contributor and the number of contributors to a profile (guideline 3.2.1.1).

The plots also demonstrate the specificity of STRmix™ by inspection of the red points. The per contributor amount for H_d true contributors was taken as the average of the known contributors (guideline 3.2.2.2). Type II errors (failure to reject a false hypothesis) are clearly identified as reds points *above* the horizontal line of $\log_{10}(LR)=0$. As for sensitivity tests, this depends on the amount of DNA within the profile and number of contributors (guideline 3.2.2.1). A series of much larger simulations (over 100 million *LR*s in total) exploring the specificity of STRmix™ and comparing it to theoretical expectations was carried out in [39]. This work found close alignment with expected and observed specificity from STRmix™ results.

The *LR* distributions for H_p true and H_d true are very well separated at high template for two contributor mixtures. As the number of contributors increased and the template lowered the two distributions converged on $\log_{10}(LR)=0$. At high template STRmix™ correctly and reliably gave a high *LR* for true contributors and a low *LR* for false contributors. At low template or high contributor number STRmix™ correctly and reliably reported that the analysis of the sample tends towards uninformative or inconclusive.

There are some arguments [1–3] that a single point estimate of the *LR* as given in Figs. 2–6 is actually the best and most theoretically sound estimate to give if the goal was an even handed and probabilistic treatment of uncertainty. In DNA profile interpretation we typically deliberately give an underestimate. In our own casework we predicate this with the words "at least" by which we mean that the number reported is either below or very near the bottom of the plausible range. Our experience suggests that this is done because of the desire by the courts and forensic scientists to avoid overstating the evidence. Over time the avoidance of overstatement has changed into what is probably a very considerable and deliberate understatement. This has been facilitated, we believe, because DNA can afford this understatement given the magnitude of our likelihood ratios.

Sensitivity and specificity studies however have a scientific component to them and it may be desirable to use the best estimate available for these. If these studies are used to formulate decisions such as assigning terms to a verbal scale then it should be noted that they refer to the point estimate and not the lower bound. This has an additional and possibly undesirable consequence that if the verbal scale is calibrated from the sensitivity and specificity plots and then this scale is applied to the lower bound, the scale itself now possesses an element of conservativeness.

There is no specific SWGDAM guideline regarding error rate but it is one of the Daubert standards regarding the admissibility of expert evidence in the US [40], with acknowledgement that these guiding factors are neither exclusionary nor mandatory [41]. With respect to forensic DNA evidence, the concept of error rates and false inclusions[2] are similar and often confused. False inclusions would come under the specificity guideline of SWGDAM (guideline 3.2).

Our preferred procedure when using STRmix™ is that the analyst assesses whether a person of interest is excluded prior to either their assessment of the results of software calculations or interpretation of the profile using the software at all. Following this procedure, STRmix™ is being continually checked against human expectations and hence is being continually validated.

The number of *LR*s > 1 is largely determined by the sample. Factors include the number of contributors and template. Considerable research has been undertaken that allows informed statements to be made about the false inclusion rate for any given sample [14,28,38,42].

The *LR* is an assessment of the weight of evidence. It is developed by considering two propositions: one aligned with the prosecution and an alternative. *LR*s > 1 support the prosecution proposition and those lower than one support the alternative.

To highlight the matter, consider that we make up a DNA mixture and hence we know the donors. Consider that this mixture is made from Smith and Brown. If we test the proposition that it contains Smith we expect a high *LR*. Suppose the *LR* is a billion. Is this correct? It is larger than one and as such that part is correct, but is a billion too large or too small or just right? The problem is

[1] Reprinted from Forensic Science International: Genetics, Volume 11, Duncan Taylor, Using continuous DNA interpretation methods to revisit likelihood ratio behaviour. Forensic Science International: Genetics, Pages 144-53, Copyright 2014, with permission from Elsevier.

[2] Note that the terms 'false inclusion' and 'false exclusion', whilst commonly used, imply an error has occurred, when in reality the probability has been assigned as expected in accordance with theory. A better term would be 'support for a false proposition'; however we retain the terms 'false inclusion/exclusion' for general understanding.

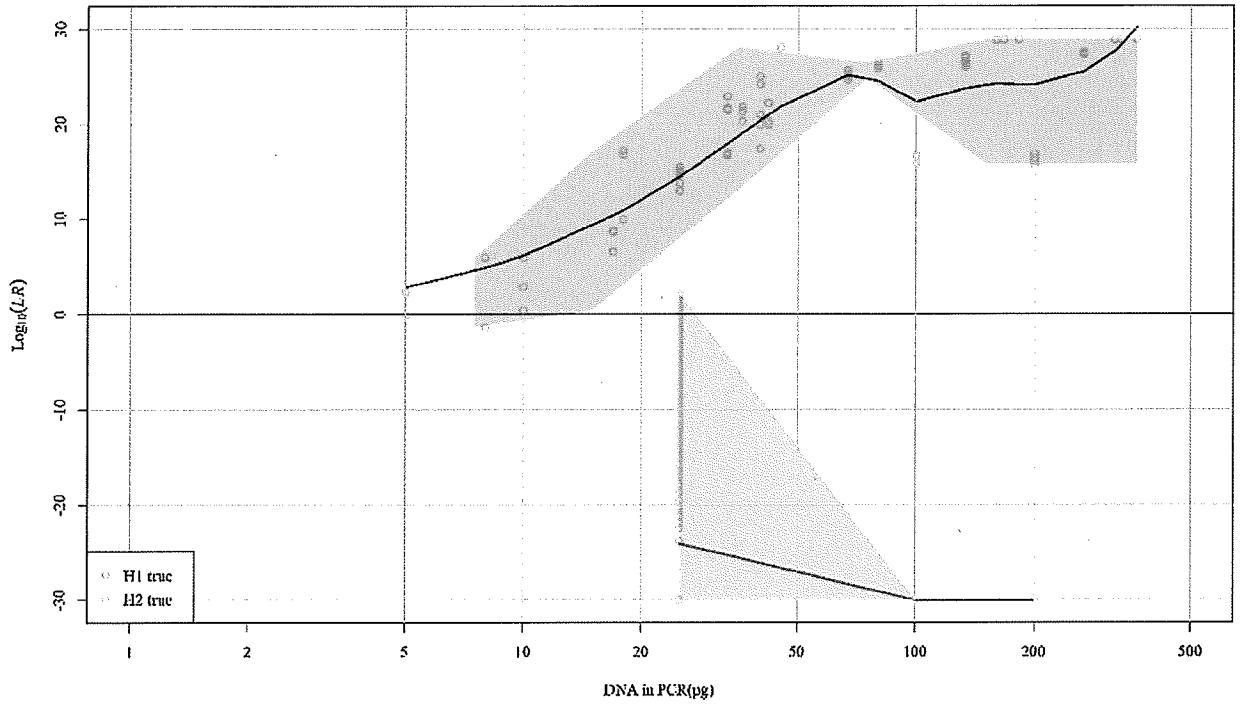


Fig. 2. *LR*s produced for two person mixtures.

that we do not have the 'true answer' and this cannot be obtained by any method.

False exclusions or false inclusions need to be interpreted in an *LR* framework. A false exclusion most nearly corresponds with an *LR* markedly less than one when H_p is true. A false inclusion most nearly corresponds with an *LR* markedly greater than one when H_d is true. *LR*s near one are best described as uninformative and this may be the correct indication of the value of the profile even for comparisons with true or false donors if the information present in the profile is limited.

When we consider a possible error rate for STRmix™ this must be balanced against the error rate for the entire DNA analysis

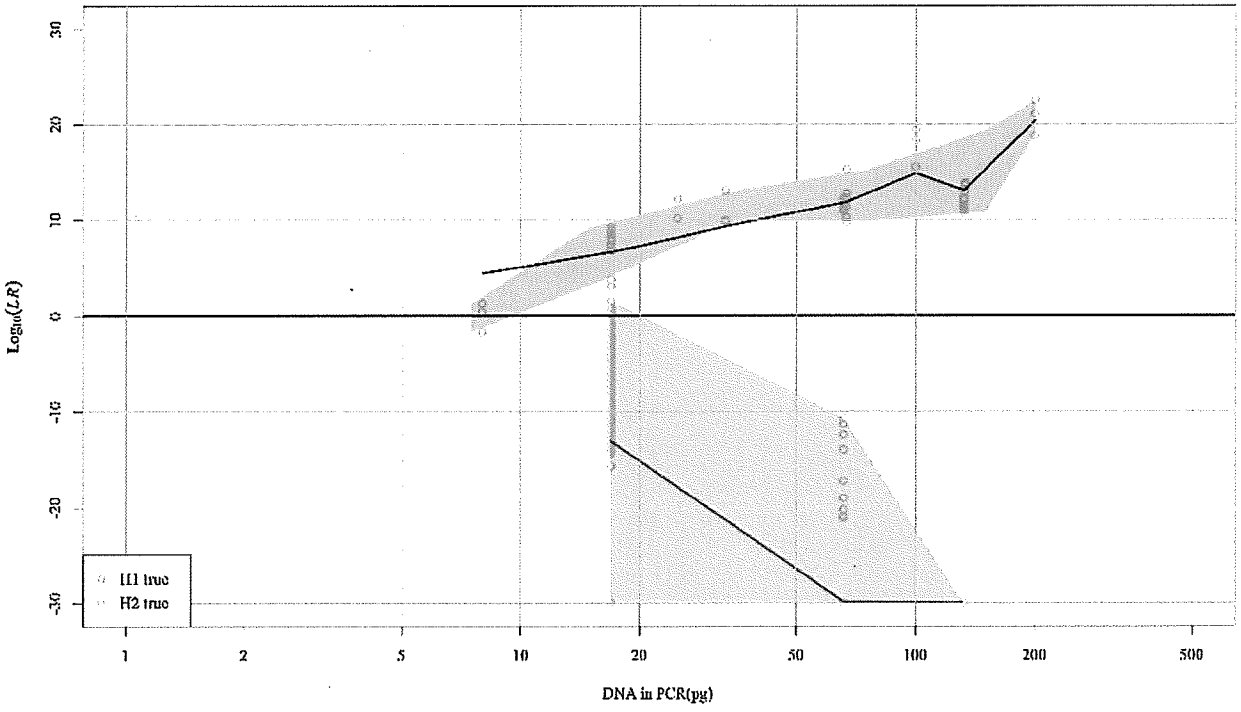


Fig. 3. *LR*s produced for three person mixtures.

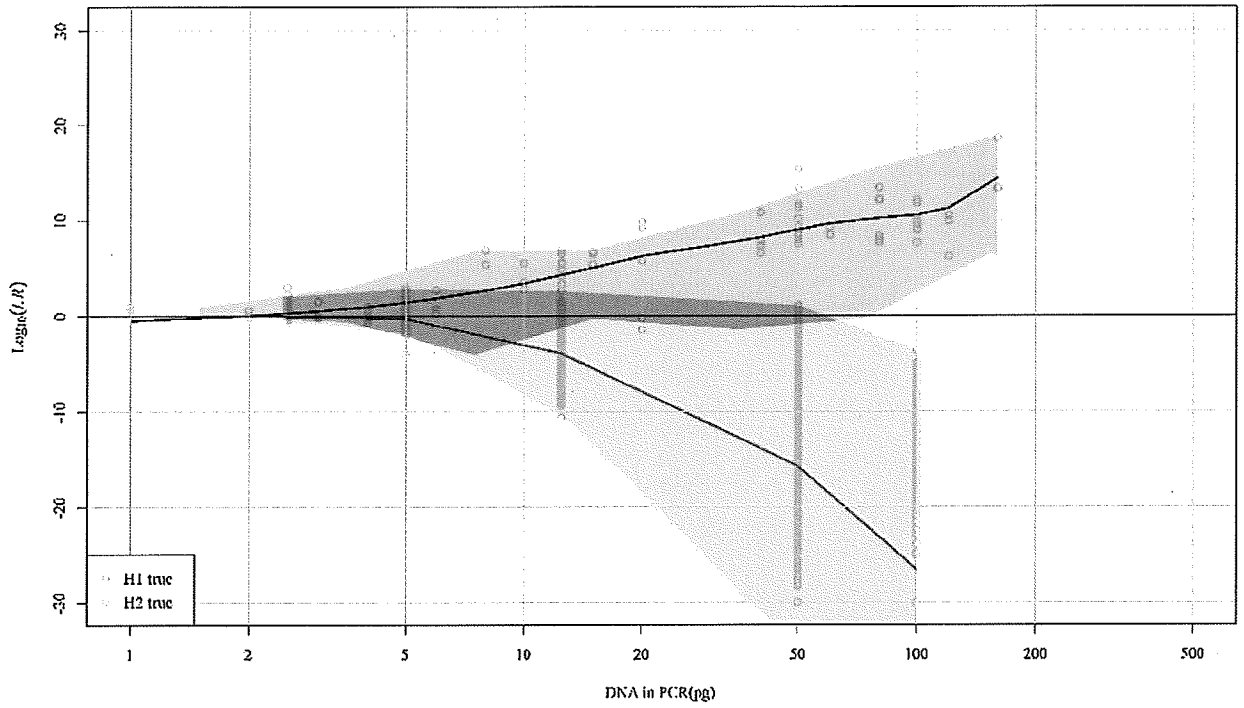


Fig. 4. *LR*s produced for four person mixtures.

process which can cause false inclusions and exclusions independent of the program. A false inclusion occurs when:

- A non-donor has the correct alleles by chance, in total or in large part, to explain the mixture.

It is very improbable that operator error (such as the inclusion of artifacts) or false information about a known contributor would cause a false inclusion.

The rate of false inclusion is increased in situations where the true DNA donor is a close relative of the POI.[3] Higher order mixtures, say four contributors, increase the chance of false inclusions. Depending on the type of profile and proportion of DNA corresponding to the POI, replication and the correct use of known contributors can reduce the chance of false inclusions (refer Figs. 5 and 6). In addition, more loci used in the analysis will reduce the chance of false inclusion.

A false exclusion occurs when:

- The PCR reaction runs sufficiently poorly that the peak or stutter heights give misleading information, or
- A non-contributor is assumed to be present, or
- There is an operator error, notably inclusion of an artifact in the peak information used by STRmix™ at interpretation. An artifactual peak that has been retained within the input file will become part of the information used by STRmix™ to build genotype combinations. This will result in genotype combinations containing the artifact which will not align with the "true" genotypes of contributors to the profile. If the POI aligns with one of these altered (false) genotypes, this might result in a false exclusion.

[3] Exploratory experimental work (ongoing) undertaken in conjunction with USACIL and the FBI suggests that STRmix™ can handle most of these situations.

There are a number of factors within STRmix™ under the control of the operator or the laboratory that affect errors. Most significantly are the two variance terms. If these are set too low they increase false exclusions. Set too high they increase false inclusions. These variances are set during a laboratory's internal validation by modelling the observed variation in allelic and stutter peak heights within a set of single source profiles of varying quality [31]. There are a number of diagnostics output by STRmix™ that allow a human check of the results including the genotypic weights ($p(O|S_j)$), the posterior mean of the variance terms and summary statistics of the MCMC (discussed later).

False inclusions and false exclusions may occur as a result of a combination of specific software, multiplex and operator factors. These are measurable. The most significant factors affecting them are the number of contributors, the number of known contributors, template levels, and the multiplex used. These factors are wrapped up in the *LR* in a way that the chance of producing an *LR* equal to or larger than the one in any particular case (LR_{case}) is less than $1/LR_{case}$. This relationship has been tested in trials of over 120 million cases of simulated false contributors and has always held [39].

The fraction of false donors exceeding LR_{case} has been termed the *p*-value [43–45] and it has been convincingly argued that they do not replace the *LR* [46]. Nor is the *p*-value a direct measure of the false inclusion rate since an *LR* for a false donor less than LR_{case} but still much larger than one would be considered a false inclusion.

We have no realistic way of measuring the false exclusion rate except to say that we have no undiagnosed instances of false exclusion.

The pink data within Figs. 2–6 are the $\log_{10}(LR)$ values for non-donors. Any red data points above the line support H_p and may therefore be considered false inclusions. These data, which are towards the low template end, are slightly above the $\log_{10}(LR)=0$ line, and are usually likelihood ratios between 1 and 1000 ($\log_{10}(LR)$ 0–3). We term these low grade false inclusions since the *LR*s are low and near neutrality or only slightly to the

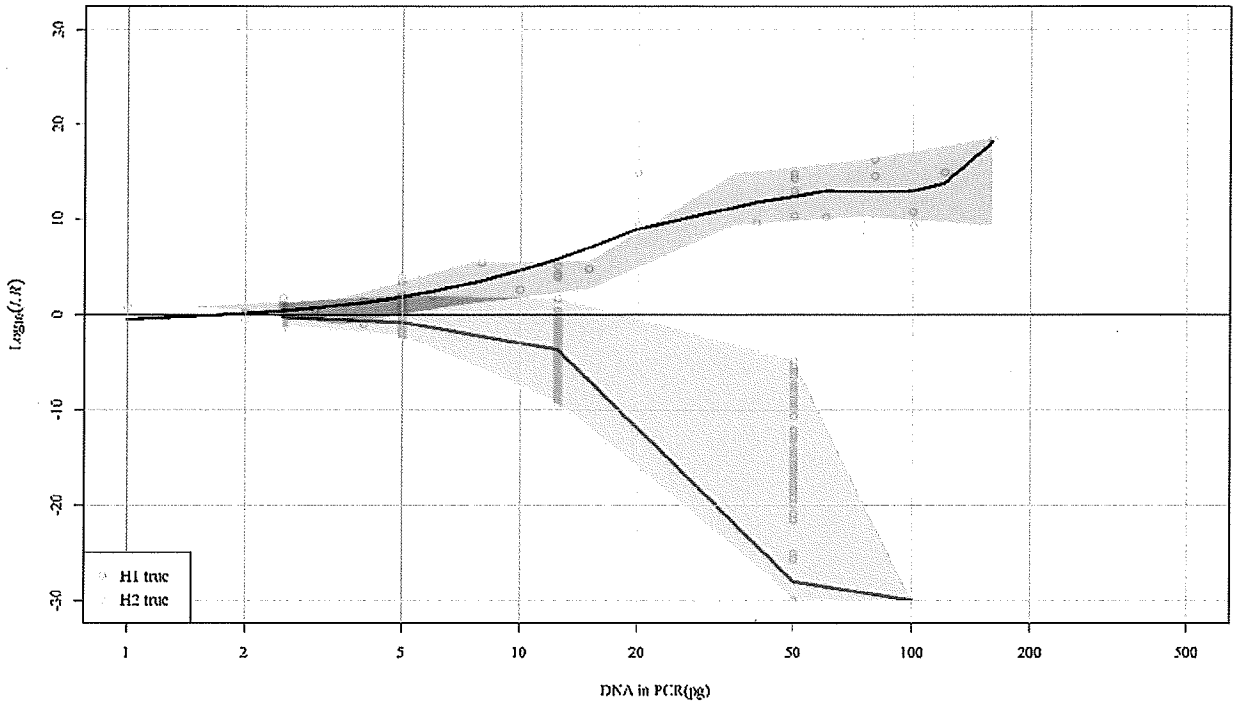


Fig. 5. *LR*s produced for four person mixtures using three replicate amplifications.

inclusionary side. They occur when the false donor has the correct alleles for inclusion and hence they are a property of DNA rather than a consequence of the software not performing. There are no modelling improvements that could ever be made which will eliminate all *LR*s that falsely favour inclusion. This is because the phenomenon causing these results is not a modelling phenomenon, but is due to the available biological data. With any interpretation method there is a modelling component (including probability of dropout and drop-in) that will affect the magnitude of the *LR*, and this could mean the difference between a false inclusion and correct exclusion for a particular non-donor.

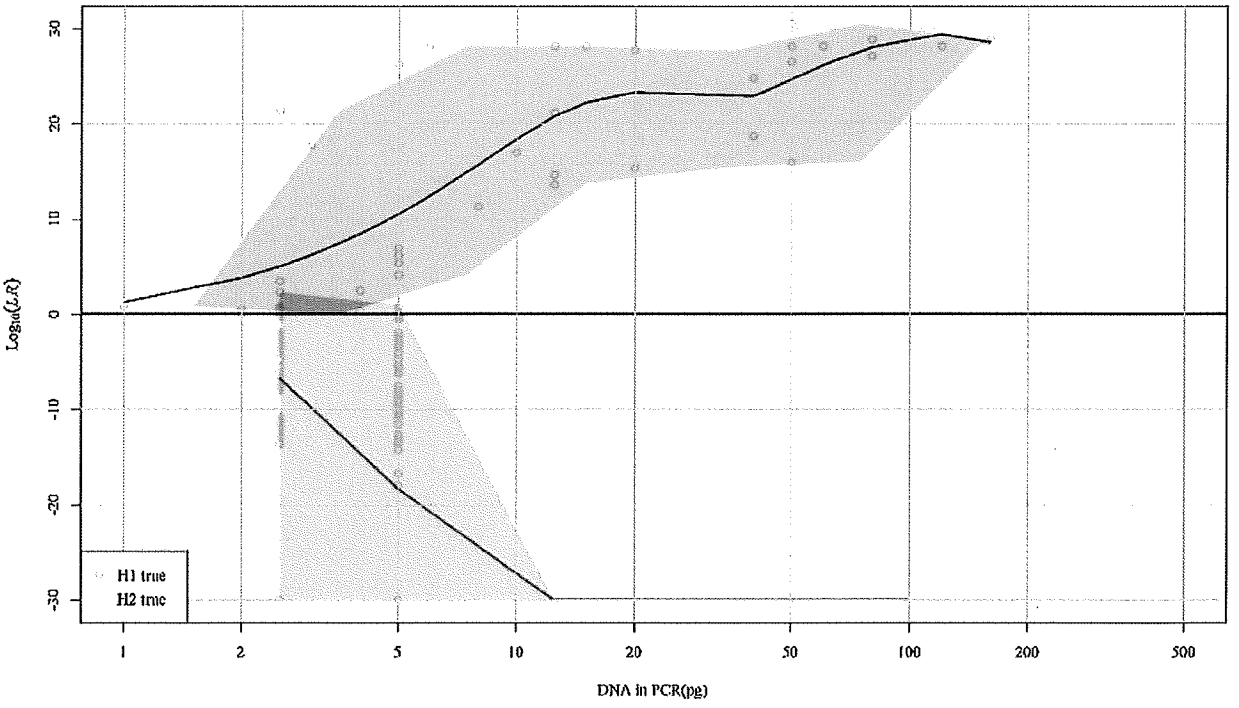


Fig. 6. *LR*s produced for four person mixtures using three replicate amplifications and assuming three out of the four known contributors in each analysis.

1.2.1. Uncertainty in the number of contributors

The determination of the effect of incorrectly assigning the number of contributors to a profile on the interpretation is not explicitly a requirement of developmental validation within the SWGDAM guidelines however this is something the STRmix™ development team has explored. The true number of contributors to a profile is always unknown. Analysts are likely to add contributors in the presence of an artifact, high stutter, or forward stutter peak. The assumption of one fewer contributor than that actually present may be made when contributors are at very low levels, are affected by peak masking and are dropping out (or visible below the analytical threshold), and in profiles where DNA is from individuals with similar profiles at the same concentrations.

The effect of the uncertainty in the number of contributors within STRmix™ has previously been reported for a number of profiles with N and $N+1$ assumed contributors, where N is the known number of contributors [28,42]. The inclusion of an additional contributor beyond that present in the profile had the effect of lowering the *LR* for trace contributors within the profile. STRmix™ adds the additional (unseen) profile at trace levels which interacts with any known trace contribution, diffusing the genotype weights and lowering the *LR*. There was no significant effect on the *LR* of the major or minor contributor within the profiles.

Separately, the effect of underestimating the number of contributors to a profile (N versus N-1) has been investigated. Assigning the number of contributors as N-1 (where N is the known number of contributors) may result in an exclusionary *LR* for a known contributor. This occurs as STRmix™ is more likely to favour an incorrect genotype as it had to account for profiling information that does not explain the data accurately.

1.3. Guideline 3.2.3. Precision

STRmix™ assigns a relative weight to the probability of the epg given each possible genotype combination at a locus. These weights are determined by Markov chain Monte Carlo (MCMC) methods. The results of no two analyses will be completely the same using a stochastic system like MCMC. This is a phenomenon that is relatively new to forensic DNA interpretation, which up until this point has always had the luxury of, at least theoretically, completely reproducible interpretation results. The reproducibility of *LR*s calculated using STRmix™ has previously been explored by Bright et al. [35,48].

The main cause of high variability within STRmix™ is non-convergence with the MCMC. Strictly, Markov chains do not converge. They explore the sampling space forever until they are told to stop. What we mean when we say Markov chains have reached convergence is that all chains are sampling from, and remain in, the 'same' high probability space.

Non-convergence is caused by the MCMC chains not being run for a sufficient number of accepts. The MCMC process starts with a number of iterations termed the 'burn-in'. Accepted genotypes from the burn-in process are not counted as they are likely to start at a low probability location. At the completion of burn-in the MCMC progresses to post burn-in. STRmix™ is set to run for a user defined number of burn-in and post burn-in accepts. STRmix™ uses accepts as a method of controlling how long the MCMC runs rather than total iterations. The reason for this is that by ensuring a defined number of accepts is obtained there is some degree of automatic scaling, whereby more complicated problems (with lower acceptance rates) will automatically run for more iterations, without the need for user intervention.

Non-convergence can be diagnosed using the Gelman-Rubin statistic [49,50]. A high Gelman-Rubin statistic in conjunction with other diagnostics may be an indication of non-convergence. The solution to non-convergence is to run the problem for longer, i.e. for more MCMC accepts. We typically multiply the number of burn-in and post burn-in accepts by 10.

Putting aside non-convergence, there will always exist a level of MCMC run to run variability. This is simply due to the fact that the analysis is based on random number generation to function, which as the name suggests, is random. Ideally this variability in some output value is small in comparison to the size of the value itself and hence its impact on interpretation is minimal, and in some instances can be taken into account. Variation in *LR*s produced from STRmix™ analyses will depend on both the sample and the run parameters. Sample specific factors that affect precision include:

1. Number of contributors to a DNA profile
2. Quality/intensity of the DNA profile
3. Number of replicates available for analysis
4. The probability of the observed data given the genotype of the POI as a contributor (commonly referred to as the 'fit' of the POI)
5. The amount of STR information available in the profile.

STRmix™ run specific parameters that affect precision include

1. Number of iterations the MCMC has run
2. The number of Markov chains used
3. The step size of the Markov chain (termed the random walk standard deviation, RWSD).

The RWSD is a metaparameter that describes the standard deviation of the normal distributions from which the step size for each continuous parameter is drawn. We describe this metaparameter in more detail below. The effect of these run specific parameters on the variability of the *LR* is discussed in detail below.

1.3.1. Number of MCMC chains and accepts

Increasing the number of either accepts or moves and adjusting the step size (the RWSD) can reduce but not totally remove the variation. There is, however, an associated runtime cost. Hence a trade-off between reproducibility and runtime must be struck.

The variation in the calculated *LR* due to sample factors and run specific parameters in STRmix™ has been explored for a number of different profiles with varying numbers of contributors and quality. Eight profiles were generated 'in silico'. These included one, two, three and four contributor profiles, in various template (high and low level) and proportions, in the GlobalFiler™ kit

Table 3
Summary of run parameters (chains, burn-in and post burn-in accepts) undertaken to interpret the sixteen profiles in order to explore the precision of STRmix™.

Set	Chains	Burn-in accepts	Post burn-in accepts
1	4	50,000	150,000
2	4	500,000	200,000
3	4	50,000	2,000,000
4	4	500,000	2,000,000
5	8	50,000	400,000
6	8	500,000	400,000
7	8	50,000	4,000,000
8	8	500,000	4,000,000
9	16	50,000	800,000
10	16	500,000	800,000
11	16	50,000	8,000,000
12	16	500,000	8,000,000
13	20	50,000	1,000,000
14	20	500,000	1,000,000
15	20	50,000	10,000,000
16	20	500,000	10,000,000

configuration. Each profile was interpreted in STRmix™ v2.3.07 ten times giving 80 runs in a batch. For each batch, a different combination of number of chains, burn-in and post burn-in accepts were trialled. In total, sixteen different chain/iteration combinations were tested generating data for over 1200 profile deconvolutions. The data was analysed to determine which chain/iteration combination resulted in the best reproducibility whilst also considering the impact on run time. A summary of the number of chain and accepts combinations considered is provided in Table 3.

A summary of the point estimate and 1st percentile (taking into account sampling variation in allele proportions and weights) of the distribution of $\log_{10}(LR)$ value (called the $\log_{10}(LR)$ and $\log_{10}$(HPD) respectively) for each of the ten replicates is provided in Appendix A (ordered by run parameter set) and Appendix B (ordered by profile). In addition summary statistics including the Gelman-Rubin diagnostic and posterior means of the allele and stutter variance constants are provided.

Inspection of the results in Appendix A and B show that as the profile is interpreted using more Markov chains and higher numbers of accepts, STRmix™ analyses are more likely to converge to the same parameter values, resulting in more reproducible *LR*s. The number of chains, total number of burn-in and post burn-in accepts and number of contributors all had an effect on run times. Consequently some interpretations were not completed after reviewing the wider results.

The *LR* for the two GlobalFiler™ single source profiles under all run configuration was identical. Due to the peak heights of these

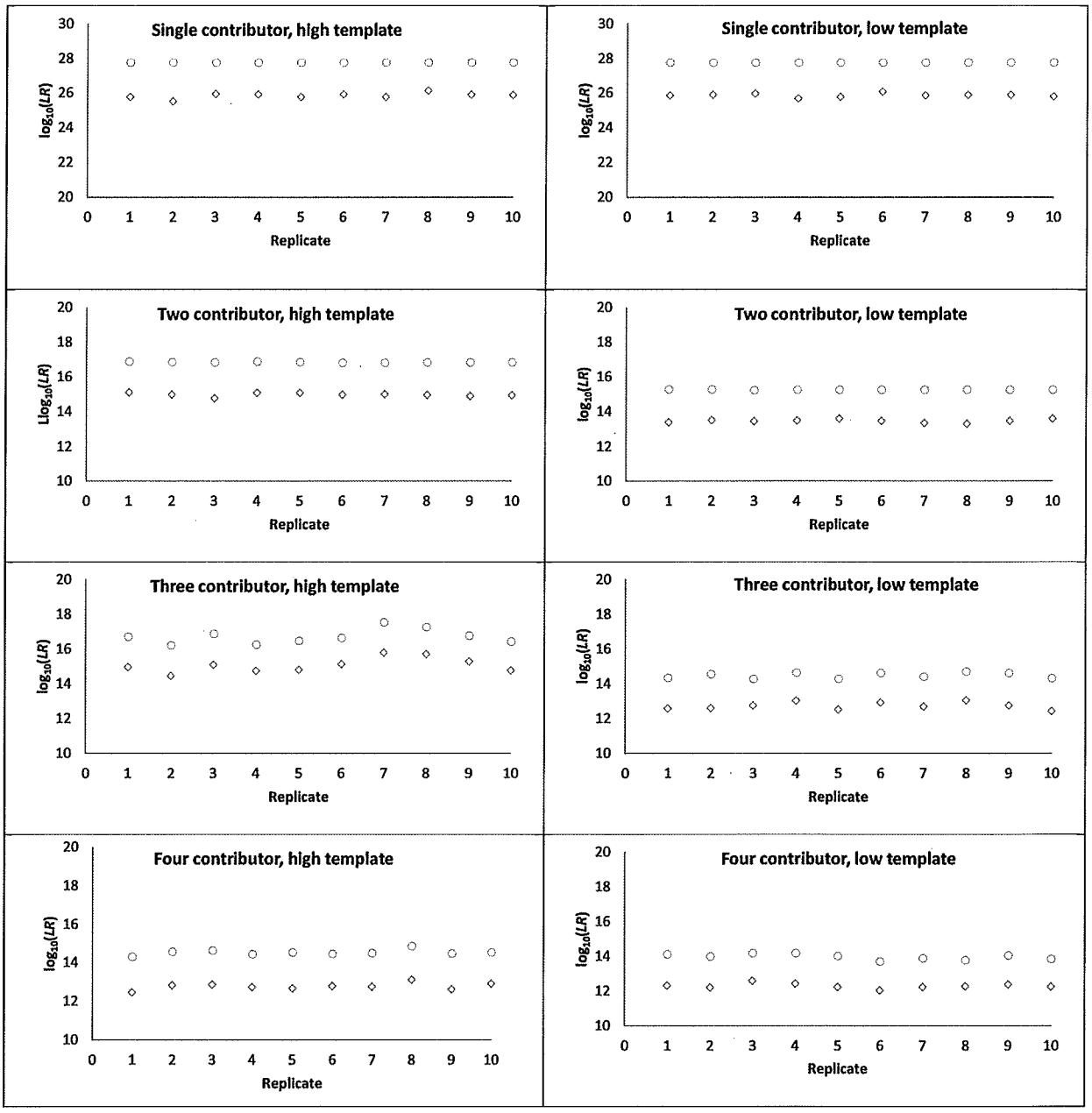


Fig. 7. $\log_{10}(LR)$ (○) and $\log_{10}$(HPD) (◇) of ten replicate interpretations of different GlobalFiler™ profiles, interpreted using eight chains with 50,000 burn-in accepts and 400,000 post burn-in MCMC accepts.

profiles dropout was not considered, resulting in a single genotype combination at each locus with weights equalling one. This was the expected result. The two person mixtures all gave *LR*s within one order of magnitude across all run configurations. There was an increase in observed *LR* variability within the three and four person mixtures with lower numbers of chains and lower total iterations.

A summary of the distribution of the $\log_{10}(LR)$ and $\log_{10}$(HPD) for ten replicates of the eight GlobalFiler™ profiles using eight chains with 50,000 burn-in accepts and 400,000 post burn-in MCMC accepts is provided in Fig. 7.

1.3.2. Random walk standard deviation

At each iteration, the MCMC will have a particular set of values stored that describe the profile. When proposing new values for the next MCMC iteration the values will be chosen close to the current set of values. The distance of the step-size is based on a normal distribution with a mean set to the current value and a variance that dictates step-size. This is known as a Gaussian random walk. In a Gaussian walk the size of the step for any given variable is sampled randomly from $\sim N(0, sd^2)$. The size of sd^2 is dependent on the parameter and is tuned by the RWSD. Setting the RWSD too high will result in the values for the mass parameters that are used to describe the profile differing significantly between steps. This will allow the Markov chain to explore much more posterior topography but will result in many rejected iterations, where parameters have been chosen that do not describe the profile adequately, resulting in longer run times. It may also have the effect of requiring additional iterations to ensure fine scale posterior topography is adequately explored. A RWSD that is set too small will mean the larger scale topography may not be explored sufficiently resulting in a decrease in precision and, potentially, accuracy. While this suggests that values for RWSD which are either too high or too low can have determimental outcomes, in practise the MCMC can accommodate a broad range of values with little negative effect, but some potentially positive. A demonstration of the effect of varying the RWSD on the $\log_{10}(LR)$ for the four contributor high and low template GloabFiler™ profile is given in Fig. 8. The profile was interpreted ten times each using three different values for the RWSD: 0.01, 0.005 and 0.0001. Interpretations were undertaken using eight chains with 50,000 burn-in accepts and 400,000 post burn-in MCMC accepts within STRmix™ version 2.4.02.

Inspection of Fig. 8 shows that reproducible *LR*s (within one order of magnitude) were generated using both 0.01 and 0.005. The run times using a RWSD of 0.005 were significantly less however than when using 0.01. The *LR*s assigned using a RWSD of 0.0001 were highly variable indicating STRmix™ had not likely explored the probability space sufficiently. On balance the RWSD value of 0.005 afforded a reproducible *LR* with a low run time.

We have demonstrated that at least 50,000 burn-in and 400,000 post burn-in accepts across eight chains and a RWSD of 0.005 are suitable MCMC run parameters leading to reproducible *LR*s (within one order of magnitude) for many different types of profile. These settings are likely to be excessive for many one, two and some three person profiles. They will be sufficient for the remaining three and most four person profiles. Decreasing the number of accepts may mean that STRmix™ has not converged and, even with convergence, more variability is expected. Increasing the number of accepts has been shown to help with reproducibility for more complex profiles and will certainly mean higher run times. A summary of the approximate run times for different profiles interpreted using STRmix™ v2.4.02 on a laptop (Windows 7 64 bit, Intel Core i7-5600U CPU, 2.6 GHz, 16 GB RAM) are given in Table 4.

In calculating the *LR*, the numerator is the weighted sum of the probability of fewer genotype sets than the denominator. In many cases the numerator may have only one term. Since the denominator is the weighted sum across the probability of many genotype sets it has a stability to variation in the *LR*. However the numerator of the *LR* is more sensitive and this effect is at its greatest when the weight for the numerator genotype set(s) is low. This is most obvious for profiles where the inclusion of a POI requires an improbable peak height variability (observed as large heterozygote balances or dropout) i.e. where the fit of the POI to the profile is poor, or when the inclusion of the POI requires one or more drop-in events to have occurred (which will also increase *LR* variability due to allele proportion uncertainty).

We have demonstrated that higher order mixtures and profiles with low template and/or poor quality lead to a decrease in precision (replication in *LR* across replicates runs). As a general guide, we have observed that if the overall *LR* is greater than 1 and one or more of the locus *LR* values are less than or equal to 1, the POI is likely to have a poor fitting genotype to the observed data at these loci. In these cases the MCMC can be run at 10 or more times the default number of accepts and/or by increasing the RWSD in order to ensure improved precision.

In general, using the default settings as described above, when comparing a POI who is a good fit to the observed profile the difference between the smallest and largest *LR* is small in relation to the size of the *LR*. For profiles where an unlikely stochastic effect has occurred, or the POI is a poor fit to the profile then the difference between the largest and smallest *LR* may be higher but again small in relation to the *LR*. In the 1200 dataset described above (Appendix B) the largest differences between the smallest and largest log(*LR*) using the recommended run settings was 1.3 fold. For profiles where an unlikely stochastic effect occurred, or the POI was a poor fit to the profile then the difference in log(*LR*) values can be above one. These situations can be minimised or eliminated via policies that suggest increasing iterations based on the profile data.

1.3.3. Reproducibility

Reproducibility is often stated as one of the main principles of the scientific method. A value is reproducible if there is a high degree of agreement between *LR*s run on the same input in different locations by different people. Reproducibility is one component of the precision of a measurement or test method. The other component is repeatability which is the degree of agreement of *LR*s on the same input by the same observer in the same laboratory.

Reproducibility is not intended to mean "exactly the same". Reproducibility means that the results are very similar within the limits of measurement or they lead to the same conclusion. In any real world application we must accept measurements to a degree of resolution and models of a limited level of complexity, or we must accept that the property we are measuring has a degree of variability. A level of uncertainty can exist in a measurement (or model) and yet the measurement can still be informative. In fact science and statistics rely on this fact.

If the same or a different operator interpreted the same input file using STRmix™ with the same random number seed[4] they would obtain exactly the same answer. So why then do we not set

[4] No computer code can actually produce a truly random number. When you tell a computer to generate a sequence of random numbers it draws upon an algorithm that generates what looks like (to humans) as being random, but will eventually start repeating itself. If a computer was told to generate a set of 1000 random numbers twice then it would generate two lists of 1000 seeming random looking numbers, but the lists would be identical. The way to get around this is by providing the algorithm with a random starting value (or 'seed').

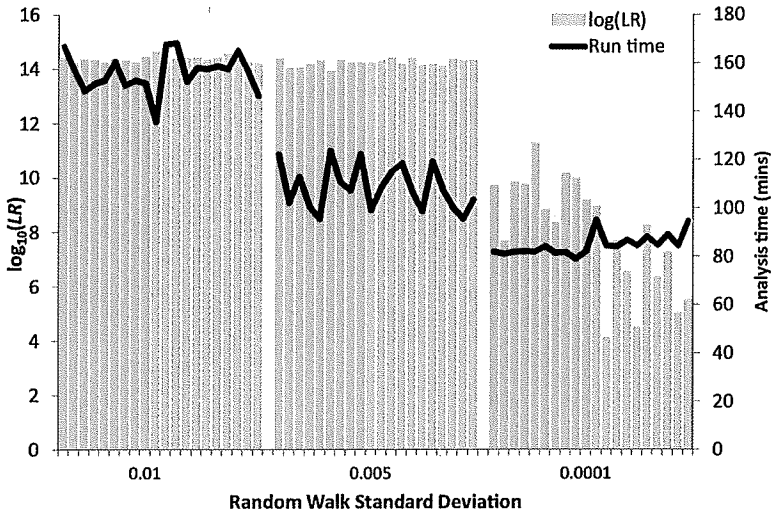


Fig. 8. $\text{Log}_{10}(LR)$ of ten replicate interpretations of the high template (blue bars) and low template (grey bars) four person GlobalFiler™ profiles, interpreted using eight chains with 50,000 burn-in accepts and 400,000 post burn-in accepts and varying RWSD. Runtime (in min) is indicated by the black lines, which correspond to the right hand vertical axis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Approximate time taken to complete interpretation of various GlobalFiler™ profile types within STRmix™ (hours:minutes:seconds), 8 chains with 50,000 burn-in and 400,000 post burn-in MCMC accepts, and RWSD of 0.005.

Number of contributors	High template profile	Low template profile
Single contributor	0:00:12	0:00:12
Two contributors	0:00:34	0:01:13
Three contributors	0:16:52	0:16:42
Four contributors	1:53:37	1:42:50

the seed and obtain exactly the same answer each time? Strangely this is dishonest repeatability. It would give a false impression of perfect precision. We prefer to give a true measure of our precision.

For very simple situations we can manually calculate the value of the *LR* from the mixture deconvolution part of the software. For the remaining situations, which comprise the vast majority of situations, we can predict limits and patterns but not exact values (for example by referring to plots such as those in Figs. 2–6). If we retain the concept of a correct, but unknowable, answer, and we plot the output from STRmix™ against these limits the patterns can be assessed to draw conclusion about the function of the STRmix™ models.

1.4. Guideline 3.2.4. Case-type samples

The mixtures described in section 3.2.3 above (Precision) include a range of profile types typically encountered in casework. These profiles include single source and mixed DNA profiles containing up to four contributors generated for both Identifiler™ and GlobalFiler™ profiles. In addition, the developmental validation of STRmix™ involved the testing of a number of profiles generated using other kits and different capillary electrophoresis instruments (3130 and 3500) including ProfilerPlus®, PowerPlex® 21, Fusion, MiniFiler™, SGMPlus™ (profiles amplified at 34 cycles) and NGM Select™ (data not shown). Back stutter is explicitly modelled in all versions of STRmix™ and version 2.4 introduces to modelling of forward stutter. The profiles included contributors with shared alleles. STRmix™ models the variability of single peaks. The variance of this model is determined by directly modelling laboratory data. This is undertaken within STRmix™ using the Model Maker function.

1.4.1. Mock samples versus casework

Three experiments have previously been reported comparing the use of mock case samples and casework samples, or single source and mixed DNA profiles, to form interpretation policies [31,51,52]. None of the studies found any obvious difference between these sets. This may be the expectation from theory. Peak height is approximately linearly proportional to the number of template molecules sampled. The standard deviation in that peak height is proportional to the square root of the number of template molecules [53,54].

If we posit that casework has degradation and inhibition effects not modelled with mock samples then we need to see how that would affect the peak heights and their variability. Degradation effectively reduces template from the starting extract but whatever number of quality template molecules survive this number is still the primary explanatory variable for peak height and relative variation. Therefore if 50% of the template was degraded we would expect this to behave similarly to a mock sample with half the template.

The effect of inhibition is more difficult to predict. Inhibitors may bind to the single stranded DNA or to the polymerases or any other co-factor. If they are simply removing template from the reaction then they would act the same as degradation. In any case what we tend to observe is that a whole locus or sets of loci amplify poorly and all peaks are lowered [55]. We could easily see how the relative variability might remain the same. STRmix™ explicitly models locus specific amplification efficiencies (LSAE). The LSAE model reflects the observation that even after template DNA amount, degradation and variation in peak height within loci are modelled, the peak heights between loci are still more variable than predicted, resulting in poorer amplification of some loci possibly due to inhibition. The variance of the LSAE model is determined by directly modelling laboratory data (see [31]). LSAE values for each STRmix™ interpretation appear within the results. We can demonstrate the relationship of LSAE values to average peak heights (APH) via a simple plot. The LSAE values should mimic the average peak heights of the locus if degradation is minimal, otherwise you will see a trend across sets of loci within dye colours according to molecular weight. This is demonstrated for one single source Identifiler™ profile in Fig. 9. The differences in APH and LSAE in this figure are due to overall profile degradation which is modelled separately.

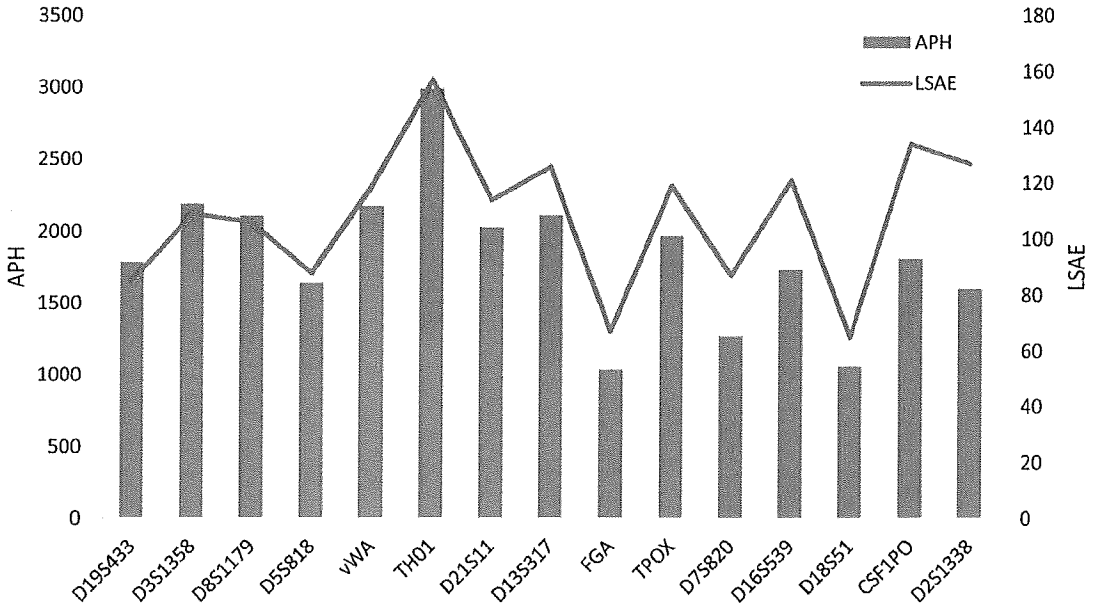


Fig. 9. Plot of APH (bars) and LSAE value (line) for each locus ordered by molecular weight for a single source Identifiler™ profile.

We have described above the theoretical expectations from the interpretation of inhibited and degraded profiles using STRmix™. Separately, we have interpreted a number of DNA profiles derived from various mock crime samples such as cigarette butts, bloodstains on wood, touched items and worn clothing. Inspection of the diagnostics from these STRmix™ interpretations including degradation and LSAE values align with our expectations (data not shown).

1.5. Guideline 3.2.6. Accuracy

There is a subset of profiles where the expected answer may be replicated relatively easily by hand. By comparing the software output with the expected answer, the performance and limitations of the software may be examined. An understanding of the models behind the methods is essential for this process. Examples of where we can predict the answer include single source profiles, mixtures where the profile of a major contributor is unambiguous (major/minor) and mixtures of two contributors in equal proportions (balanced). STRmix™ has been shown to give the expected result in each case [48].

Functionality has been installed within STRmix™ to facilitate validation and performance checks. This includes the extended output and set seed functions. The extended output contains all of the parameters and calculated probabilities for each iteration within a run. The 'set seed' function turns off the random processes within STRmix™ and allows direct comparison of runs within and between different versions of the software. STRmix™ is built in two separate parts that communicate via a text file. The first part runs the MCMC, the second the *LR* calculation. Hence, in some version releases it is possible to test one part using an old output from the other part variously using the set seed or checks of the extended output to allow the direct comparison of outputs and lessen the validation load.

The following functionality and outputs from STRmix™ were verified by hand as part of the developmental validation tasks for each commercial version:

1. Expected allele and stutter heights given mass parameters
2. Expected peak heights of drop or 'Q' alleles given mass parameters
3. Probabilities given expected and observed peak heights and varying analytical thresholds
4. Locus specific amplification efficiency calculations
5. Summation of probabilities for each allele in a locus and across a profile
6. Summation of probabilities across multiple replicate profiles
7. Informed priors on mixture proportion
8. *LR* values where there are no assumed contributors
9. *LR* values for propositions with assumed contributors
10. *LR* values with varying theta values
11. Relatives calculations (where a relative is considered as an alternate contributor under H_d)
12. Sampling from the Beta distributions for theta
13. *LR* stratified point estimates
14. *LR* highest posterior density (HPD) interval values
15. Gaussian walk
16. Gelman-Rubin statistic, ESS, weight resampling
17. Drop-in function
18. Database search functionalities
19. Model maker.

The comparison of expected heights, probability and *LR* values was conducted in MS Excel or by comparison to results generated in the r_{HPD} package written by Professor James Curran in R [56].

The likelihood ratios calculated using STRmix™ have been compared to two probabilistic genotyping methods employing semi-continuous models and two binary methods of profile interpretation [48,57]. Where a profile was able to be fully resolved or for single source profiles where dropout was not a consideration (weight, w_i, equals one at each locus) the *LR* between STRmix™ and the semi-continuous methods were comparable where they were using the same population genetics model. For mixed DNA profiles, generally STRmix™ resulted in higher *LR*s for ground truth known trials as continuous models use more of the profiling information (for example peak height information) compared with semi-continuous and binary interpretation methods.

2. Conclusion

Within this paper we describe the exercises undertaken as part of STRmix™ developmental validation following the SWGDAM guidelines for the validation of probabilistic genotyping software [1]. This work demonstrates that STRmix™ is suitable for its intended use for the interpretation of single source and mixed DNA profiles including profiles of a complex and low level nature.

A number of different parameters within STRmix™ that are known to affect *LR* reproducibility were investigated. We have interpreted over 1200 profiles and conclude that at least 50,000 burn-in and 400,000 post burn-in accepts across eight Markov chains and a RWSD of 0.005 are suitable STRmix™ run parameters leading to reproducible *LR*s (within one order of magnitude) for many different types of profile.

Having undertaken both internal and developmental validations following the SWGDAM guidelines we find them a good template within which to work. Recommendations 3.2.5 (control samples) and 3.2.6.2 (analysis of raw data files) are not applicable to STRmix™.

Acknowledgements

The authors would like to thank Johanna Veth (ESR), John Simich (Erie County, NY), Shawn Monpetit (San Diego, CA), Bjorn Sutherland (ESR) and two anonymous reviewers for their helpful comments that greatly improved this paper.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2016.05.007 .

References

[1] Scientific Working Group on DNA Analysis Methods (SWGDAM). Guidelines for the Validation of Probabilistic Genotyping Systems. 2015.
[2] B. Budowle, A.J. Onorato, T.F. Callagham, A.D. Manna, A.M. Gross, R.A. Guerreri, et al., Mixture Interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework, J. Forensic Sci. 54 (2009) 810–821.
[3] P. Gill, J. Buckleton, Commentary on: Budowle B, Onorato AJ, Callaghan TF, della Manna A, Gross AM, Guerrieri RA, Luttman JC, McClure DL, Mixture interpretation: Defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework [J. Forensic Sci. 2009, 54 (4), 810–821], J. Forensic Sci. 55 (2010) 265–268.
[4] J.S. Buckleton, C.M. Triggs, S.J. Walsh, DNA Evidence, CRC Press, Boca Raton, Florida, 2004.
[5] I.E. Dror, D. Charlton, A.E. Peron, Contextual information renders experts vulnerable to making erroneous identifications, Forensic Sci. Int. 156 (2006) 74–78.
[6] W.C. Thompson, Painting the target around the matching profile: the Texas sharpshooter fallacy in forensic DNA interpretation Law, Law Probab. Risk 8 (2009) 257–276.
[7] Scientific Working Group on DNA Analysis Methods (SWGDAM). SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories. 2010.
[8] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, et al., DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2006) 90–101.
[9] H. Haned, Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics, Forensic Sci. Int. Genet. 5 (2011) 265–268.
[10] H. Haned, P. Gill, Analysis of complex DNA mixtures using the Forensim package, Forensic Sci. Int. Genet. Suppl. Ser. 3 (2011) e79–e80.
[11] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Sci. Int. Genet. 4 (2009) 1–10.
[12] K. Lohmueller, N. Rudin, Calculating the weight of evidence in low-template forensic DNA casework, J. Forensic Sci. 58 (s1) (2013) s234–59.
[13] A.A. Mitchell, J. Tamariz, K. O'Connell, N. Ducasse, Z. Budimlija, M. Prinz, et al., Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in, Forensic Sci. Int. Genet. 6 (2012) 749–761.
[14] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int. Genet. 7 (2013) 516–528.
[15] FBI Quality Assurance Standards for Forensic DNA Testing Laboratories. 2011.
[16] J.-A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Sci. Int. Genet. 7 (2013) 296–304.
[17] J.-A. Bright, D. Taylor, C. J.M, J.S. Buckleton, Degradation of forensic DNA profiles, Aust. J. Forensic Sci. 45 (2013) 445–449.
[18] J. Buckleton, H. Kelly, J.-A. Bright, D. Taylor, T. Tvedebrink, J.M. Curran, Utilising allelic dropout probabilities estimated by logistic regression in casework, Forensic Sci. Int. Genet. 9 (2014) 9–11.
[19] R. Puch-Solis, A dropin peak height model, Forensic Sci. Int. Genet. 11 (2014) 80–84.
[20] C. Brookes, J.-A. Bright, S. Harbison, J. Buckleton, Characterising stutter in forensic STR multiplexes, Forensic Sci. Int. Genet. 6 (2012) 58–63.
[21] J.-A. Bright, J.M. Curran, J.S. Buckleton, Investigation into the performance of different models for predicting stutter, Forensic Sci. Int. Genet. 7 (2013) 422–427.
[22] W.K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57 (1970) 97–109.
[23] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equations of state calculations by fast computing machines, J. Chem. Phys. 21 (1953) 1087–1091.
[24] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (1994) 125–140.
[25] National Research Council Report: The Evaluation of Forensic DNA Evidence, National Academy Press, Washington DC, 1996.
[26] D. Taylor, J.-A. Bright, J. Buckleton, Considering relatives when assessing the evidential strength of mixed DNA profiles, Forensic Sci. Int. Genet. 13 (2014) 259–263.
[28] J.-A. Bright, D. Taylor, J. Curran, J. Buckleton, Searching mixed DNA profiles directly against profile databases, Forensic Sci. Int. Genet. 9 (2014) 102–110.
[29] D. Taylor, J.A. Bright, J. Buckleton, J. Curran, An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations, Forensic Sci. Int. Genet. 11 (2014) 56–63.
[30] C.M. Triggs, J.M. Curran, The sensitivity of the Bayesian HPD method to the choice of prior, Sci. Justice 46 (2006) 169–178.
[31] D. Taylor, J. Buckleton, J.-A. Bright, Factors affecting peak height variability for short tandem repeat data, Forensic Sci. Int. Genet. 21 (2016) 126–133.
[32] J.-A. Bright, J.M. Curran, Investigation into stutter ratio variability between different laboratories, Forensic Sci. Int. Genet. 13 (2014) 79–81.
[33] H. Kelly, J.-A. Bright, J.S. Buckleton, J.M. Curran, Identifying and modelling the drivers of stutter in forensic DNA profiles, Aust. J. Forensic Sci. 46 (2013) 194–203.
[34] D. Taylor, J.-A. Bright, C. McGoven, C. Hefford, T. Kalafut, J. Buckleton, Validating multiplexes for use in conjunction with modern interpretation strategies, Forensic Sci. Int. Genet. 20 (2016) 6–19.
[35] J.-A. Bright, K.E. Stevenson, J.M. Curran, J.S. Buckleton, The variability in likelihood ratios due to different mechanisms, Forensic Sci. Int. Genet. 14 (2015) 187–190.
[36] D. Taylor, J.A. Bright, J. Buckleton, The 'factor of two' issue in mixed DNA profiles, J. Theor. Biol. 363 (2014) 300–306.
[37] D. Taylor, J. Buckleton, J.-A. Bright, Does the use of probabilistic genotyping change the way we should view sub-threshold data? Aust. J. Forensic Sci. (2015), doi:http://dx.doi.org/10.1080/00450618.2015.1122082.
[38] D. Taylor, Using continuous DNA interpretation methods to revisit likelihood ratio behaviour, Forensic Sci. Int. Genet. 11 (2014) 144–153.
[39] D. Taylor, J. Buckleton, I. Evett, Testing likelihood ratios produced from complex DNA profiles, Forensic Sci. Int. Genet. 16 (2015) 165–171.
[40] Daubert et al. v Merrell Dow Pharmaceuticals Inc., 509 US 579 (1993). 1993.
[41] Kumho Tire Co. Ltd et al. v. Carmichael et al. In: Court USS, editor. 526 US 1371999.
[42] J.-A. Bright, J.M. Curran, J.S. Buckleton, The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation, Forensic Sci. Int. Genet. 12 (2014) 208–214.
[43] G. Dørum, Ø. Bleka, P. Gill, H. Haned, L. Snipen, S. Sæbø, et al., Exact computation of the distribution of likelihood ratios with forensic applications, Forensic Sci. Int. Genet. 9 (2014) 93–101.
[44] P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, Forensic Sci. Int. Genet. 7 (2013) 251–263.
[45] H. Haned, G. Dorum, E. Egeland, P. Gill, On the meaning of the likelihood ratio: is a large number always an indication of strength of evidence? 25th Congress of the International Society for Forensic Genetics, Melbourne, Australia, 2013.
[46] M. Kruijver, R. Meester, K. Slooten, p-Values should not be used for evaluating the strength of DNA evidence, Forensic Sci. Int. Genet. 16 (2016) 226–231.
[48] J.-A. Bright, I.W. Evett, D. Taylor, J.M. Curran, J. Buckleton, A series of recommended tests when validating probabilistic DNA profile interpretation software, Forensic Sci. Int. Genet. 14 (2015) 125–131.
[49] A. Gelman, D.B. Rubin, Inference from iterative simulation using multiple sequences, Stat. Sci. 7 (1992) 457–511.
[50] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, Chapman & Hall, New York, 1995.
[51] J.-A. Bright, K. McManus, S. Harbison, P. Gill, J. Buckleton, A comparison of stochastic variation in mixed and unmixed casework and synthetic samples, Forensic Sci. Int. Genet. 6 (2012) 180–184.

[52] J.-A. Bright, J. Turkington, J. Buckleton, Examination of the variability in mixed DNA profile parameters for the Identifiler(TM) multiplex, Forensic Sci. Int. Genet. 4 (2009) 111–114.
[53] P. Gill, J. Curran, K. Elliot, A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci, Nucleic Acids Res. 33 (2005) 632–643.
[54] J. Weusten, J. Herbergs, A stochastic model of the processes in PCR based amplification of STR DNA in forensic applications, Forensic Sci. Int. Genet. 6 (2012) 17–25.
[55] J.-A. Bright, S. Cockerton, S. Harbison, A. Russell, O. Samson, K. Stevenson, The effect of cleaning agents on the ability to obtain DNA profiles using the Identifiler™ and PowerPlex® Y multiplex kits, J. Forensic Sci. 56 (2011) 181–185.
[56] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2004.
[57] T.W. Bille, S.M. Weitz, M.D. Coble, J.S. Buckleton, J.-A. Bright, Comparison of the performance of different models for the interpretation of low level mixed DNA profiles, Electrophoresis 35 (2014) 3125–3133.

Forensic Science International: Genetics 29 (2017) 126–144

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Internal validation of STRmix™ for the interpretation of single source and mixed DNA profiles

Tamyra R. Moretti[a,*], Rebecca S. Just[a], Susannah C. Kehl[b], Leah E. Willis[a], John S. Buckleton[c,d], Jo-Anne Bright[c], Duncan A. Taylor[e,f], Anthony J. Onorato[a]

[a] DNA Support Unit, Federal Bureau of Investigation Laboratory, 2501 Investigation Parkway, Quantico, VA 22135, USA
[b] Biometrics Analysis Section, Federal Bureau of Investigation Laboratory, 2501 Investigation Parkway, Quantico, VA 22135, USA
[c] Institute of Environmental Science and Research, Private Bag 92021, Auckland 1025, New Zealand
[d] National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899, USA
[e] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[f] School of Biological Sciences, Flinders University, GPO Box 2100 Adelaide, SA, 5001 Australia

ARTICLE INFO

Article history:
Received 28 July 2016
Received in revised form 15 March 2017
Accepted 3 April 2017
Available online 5 April 2017

Keywords:
STRs
DNA Mixtures
Probabilistic Genotyping
Likelihood Ratios

ABSTRACT

The interpretation of DNA evidence can entail analysis of challenging STR typing results. Genotypes inferred from low quality or quantity specimens, or mixed DNA samples originating from multiple contributors, can result in weak or inconclusive match probabilities when a binary interpretation method and necessary thresholds (such as a stochastic threshold) are employed. Probabilistic genotyping approaches, such as fully continuous methods that incorporate empirically determined biological parameter models, enable usage of more of the profile information and reduce subjectivity in interpretation. As a result, software-based probabilistic analyses tend to produce more consistent and more informative results regarding potential contributors to DNA evidence. Studies to assess and internally validate the probabilistic genotyping software STRmix™ for casework usage at the Federal Bureau of Investigation Laboratory were conducted using lab-specific parameters and more than 300 single-source and mixed contributor profiles. Simulated forensic specimens, including constructed mixtures that included DNA from two to five donors across a broad range of template amounts and contributor proportions, were used to examine the sensitivity and specificity of the system via more than 60,000 tests comparing hundreds of known contributors and non-contributors to the specimens. Conditioned analyses, concurrent interpretation of amplification replicates, and application of an incorrect contributor number were also performed to further investigate software performance and probe the limitations of the system. In addition, the results from manual and probabilistic interpretation of both prepared and evidentiary mixtures were compared.

The findings support that STRmix™ is sufficiently robust for implementation in forensic laboratories, offering numerous advantages over historical methods of DNA profile analysis and greater statistical power for the estimation of evidentiary weight, and can be used reliably in human identification testing. With few exceptions, likelihood ratio results reflected intuitively correct estimates of the weight of the genotype possibilities and known contributor genotypes. This comprehensive evaluation provides a model in accordance with SWGDAM recommendations for internal validation of a probabilistic genotyping system for DNA evidence interpretation

1. Introduction

As the sensitivity of forensic DNA typing procedures has improved with the development of better DNA extraction and amplification chemistries and detection instrumentation, more DNA profiles originating from the DNA of two or more individuals are being encountered in forensic casework. The complexity of profile interpretation increases with each additional contributor to

* Corresponding author.
E-mail addresses: Tamyra.Moretti@ic.fbi.gov, trmoretti828@gmail.com (T.R. Moretti).

http://dx.doi.org/10.1016/j.fsigen.2017.04.004
1872-4973/

a mixture, particularly if the DNA contribution is low and therefore subject to stochastic effects (e.g., allele dropout and greater heterozygous peak height variance). Binary decision making can be applied to the interpretation of mixed profiles and has historically been used in many aspects of the analysis of DNA for human identification purposes. This approach has provided an easily applied means of addressing biological phenomena exhibited in PCR-based typing results at short tandem repeat (STR) loci [1–3]. Two primary outcomes are considered in a binary interpretation method. For example, (a) a peak observed in an electropherogram at an expected stutter position is interpreted as either stutter or an allelic peak based on relative height, (b) two allelic peaks are interpreted as having originated from the same or different individuals depending on whether they fall within height variance expectations for heterozygous alleles, and (c) an allele is either used or not used to estimate evidential weight based on whether its height meets an empirically determined stochastic threshold [4].

Such "either-or" determinations, however, can be difficult to make given the characteristics of STR mixture results. The primary criterion used in STR interpretation is peak amplitude, relative to the size and position of the peak in the electropherogram. Yet, the sharing of an allele with that of another contributor and/or with a stutter product renders peak height information less meaningful. Furthermore, locus-specific amplification efficiencies and DNA degradation, which can vary in degree among contributors in a mixture, impact relative peak heights. Also, an allelic component of peaks that qualify as stutter cannot be ruled out when alleles from a minor contributor(s) are in the same general height range as stutter peaks [2]. Together with the possibility of allele dropout, the intricacies of mixture analysis create scientific uncertainty in the determination of possible contributor genotypes and can complicate manual interpretation of mixed DNA profiles.

The use of safeguards (such as a stochastic threshold) was recommended by the Scientific Working Group on DNA Analysis Methods (SWGDAM) to mitigate the uncertainty inherent to binary interpretation of single source, mixed-source and low-level typing results [5]. These safeguards, if applied correctly, tend to limit the usage of profile information and thereby typically lead to more common profile probability estimates, as well as more inconclusive conclusions.

Statistical software programs that incorporate probabilistic interpretation models overcome these limitations and fully utilize the available DNA typing information [6–9]. Probabilistic genotyping refers to the use of software and computer algorithms to apply biological modeling, statistical theory, and probability distributions to infer the probability of the profile from single source and mixed DNA typing results given different contributor genotypes [10]. The software weighs potential genotypic solutions for a mixture by utilizing more DNA typing information (e.g., peak height, allelic designation and molecular weight) and accounting for uncertainty in random variables within the model, such as peak heights (e.g., via peak height variance parameters and probabilities of allelic dropout and drop-in, rather than a stochastic or dropout threshold). Likelihood ratios (*LRs*) are generated to express the weight of the DNA evidence given two user-defined propositions. Probabilistic genotyping software has been demonstrated to reduce subjectivity in the interpretation of DNA typing results and, compared to binary interpretation methods, is a more powerful tool supporting the inclusion of contributors to a DNA sample and the exclusion of non-contributors [11]. Despite the effectual incorporation of higher level interpretation features, though, probabilistic software programs are not Expert Systems as defined under the National DNA Index System (NDIS) Procedures [12]. The DNA typing data and probabilistic genotyping results require human interpretation and review in accordance with the *Quality Assurance Standards for Forensic DNA Testing Laboratories* [13].

The fundamental onus on the forensic laboratory with regard to the analysis of DNA mixtures is to seek to remain current with technological developments and relevant issues and to ensure the reliability of its procedures and usage in casework by properly and thoroughly validating any new method prior to use. The interpretation of complex mixtures in particular requires that the laboratory design and execute thorough, targeted experimental studies as part of its internal validation, recognize limitations revealed through the results, and use the results of validation studies to develop detailed, reliable procedures that can be applied uniformly and consistently among analysts. SWGDAM provides guidelines and the *Quality Assurance Standards for Forensic DNA Testing Laboratories* provide quality assurance requirements for validation [13,14].

We outline here the internal validation of STRmix™ [6,15] at the FBI Laboratory in accordance with *SWGDAM Guidelines for the Validation of Probabilistic Genotyping Systems* [10]. STRmix™ is software that employs a continuous model for DNA profile interpretation and genotype determination based on a Markov Chain Monte Carlo (MCMC) sampling method. Using weights assigned to the resultant genotypes or genotype sets, STRmix™ calculates *LRs*, which are the probability of the DNA evidence under two opposing hypotheses referred to as H_1 and H_2. The terms H_1 and H_2 are used in lieu of "Prosecution hypothesis" (H_p) and "Defense hypothesis" (H_d), respectively, given that they are assigned by the scientist, usually without consultation with legal representatives.

A *LR* greater than 1 provides support for a specified person of interest as a contributor to the DNA evidence (H_1), whereas an *LR* less than 1 provides support that the person of interest is not a contributor (H_2). An *LR* of 1 provides no greater support for either proposition. We describe suitable experiments using single source samples and a breadth of mixed DNA samples to meet the recommendations and requirements for internal validation and detail additional testing conducted at the FBI Laboratory to aid in procedural and policy development.

2. Methods

All single source and mixed DNA profiles were generated in-house using DNA samples (collected and typed with informed consent) that were amplified for 27 cycles using the Applied Biosystems AmpFlSTR® Identifiler® Plus PCR Amplification Kit (Thermo Fisher Scientific, Waltham, MA), followed by detection on a 3130xl Genetic Analyzer (Thermo Fisher Scientific). 3130xl data were subsequently analyzed using Applied Biosystems GeneMapper® ID-X version 1.3 (Thermo Fisher Scientific). Protocols and analysis settings (including an analytical threshold of 50 relative fluorescent units, or rfu) previously validated by the FBI Laboratory for casework usage were used to prepare DNA samples, generate DNA typing results and perform preliminary interpretations prior to STRmix™ analysis. Given that STRmix™ version 2.3 models back stutter of one repeat unit, such peaks were retained in all input files for questioned profiles. All other artifact peak labels were removed following FBI Laboratory guidelines, including forward stutter, which is not modeled by the software versions tested.

Laboratory-specific STRmix™ parameters were established using more than 1400 single source Identifiler® Plus profiles of variable quantity and quality (Table S1). Some DNA extracts used for parameter setting were artificially degraded during 90 or 180 seconds of UV irradiation using a Spectrolinker™ XL-1000 UV Crosslinker (Spectronics Corporation, Westbury, NY), with the samples placed, caps open, 2.5 inches from the ultraviolet light

source. Degradation was confirmed following amplification of the irradiated extracts (serially diluted for amplification of 1 ng–0.03 ng template DNA) and demonstration of complete locus dropout, particularly of higher molecular weight alleles, at one or more loci. Per allele stutter ratio (SR) expectations were determined by regressing SR against allele designation and SR against the longest uninterrupted stretch (LUS) of repeats within an allele for some compound or complex repeats (e.g., TH01 9.3 allele has a LUS of 6 repeat units [16]). The maximum allowable stutter ratio was set arbitrarily high at 0.3, or 30% (this parameter is used in the initial assessment of potential alleles, not in stutter modeling, only for run time pick up). For saturated data (peaks generated from high template amounts and/or over-amplification that saturate the camera within the genetic analyzer), an alternate model is invoked within STRmix™ since the relationship of DNA template to peak height is no longer linear. Specifically, the height of a given stutter peak is not determined from the observed (saturated) allele but from an expected allele height based on the proposed model. Based on empirical 3030xl data, a peak height upper limit of 7000 rfu was established for saturation.

The ModelMaker function within STRmix™ uses a MCMC system to analyze a set of laboratory data of known origin in order to determine the distribution of peak height variability specific to the laboratory [17]. This process is used to establish a distribution of expected values for allele, stutter and locus specific amplification variances that are used by STRmix™ in the analysis of data [15,17,18]. By assessing over 700 single source profiles of varying quantity and quality (Table S1) using ModelMaker, peak height variance constant prior distributions for allelic peaks [c^2, Γ(4.2818,1.0671) with a mode of 4.219] and for stutter peaks [k^2, Γ(9.1442,1.1239) with a mode of 7.528] and a mean locus-specific amplification efficiency variance (0.0113) were determined.

The peak height variance constant is explained using the allelic example (the other distributions have similar forms): the distribution for allelic peaks has a mode at 4.219, and 95% of values for allele variance fall between 1.3 and 9.9. These values depict the way that variance changes with peak height (high at low template and low at high template). There is a relationship between peak variance and heterozygote balance (*Hb*) [19,20]. To show this, $\log_{10}(Hb)$ was plotted against average peak height (APH, based on rfu values of alleles at heterozygote loci), and the expected 95% bounds were calculated at $\pm\sqrt{2} \times 1.96 \times \sqrt{\frac{c^2}{APH}}$, where $c^2 = 4.219$. Such a graph (Fig. 1) allows for assessment of the parameters: with 97.6% of the *Hb* data falling within the 95% bounds, allelic variance was demonstrated to be sufficiently optimized.

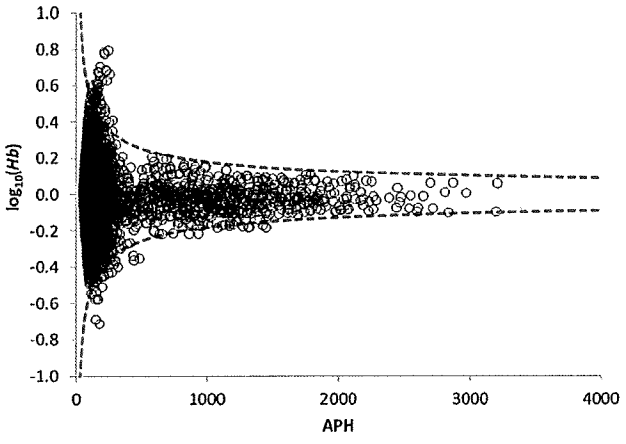


Fig. 1. Plot of $\log_{10}(Hb)$ versus APH for 4125 heterozygote loci from 709 single source ModelMaker profiles. The dashed lines represent the expected 95% bounds and encapsulate 97.6% of all data points.

The drop-in rate was set to zero since no allelic peaks ≥50 rfu were detected at the 16 Identifiler® Plus loci following 27 cycles of amplification of 500 reagent blanks extracted using the EZ1® Advanced XL (QIAGEN Sciences, Inc., Gaithersburg, MD), nor in any amplified DNA sample throughout the study.

To confirm proper calculation of the *LR* by STRmix™, the *LR* for two single source profiles and two two-person mixtures (57 individual loci), where weights determined by STRmix™ equaled both one and less than one for the known contributor profile, were calculated "manually" within Excel. Loci included both heterozygous and homozygous examples, and calculations were undertaken with $\theta = 0.01$ and $\theta = 0$. Setting θ to zero returns the product rule, where *LR* equals:

$2p_ip_j$ for heterozygous loci ($i \neq j$)

p_i^2 for homozygous loci.

When θ > 0, the Balding and Nichols formulae [21] (or equations 4.10 from NRC II [22]) are applied. For single source profiles:

$$\frac{2[\theta + (1-\theta)p_i]\left[\theta + (1-\theta)p_j\right]}{(1+\theta)(1+2\theta)} \quad \text{for heterozygous loci} \tag{1}$$

$$\frac{[3\theta + (1-\theta)p_i][2\theta + (1-\theta)p_i]}{(1+\theta)(1+2\theta)} \quad \text{for homozygous loci} \tag{2}$$

where p_i is the allele frequency for allele i, p_j the allele frequency for allele j and θ is the F_{ST} value. The allele frequencies used within equations 1 and 2 are posterior mean frequencies. These are calculated using the following equation:

$$\frac{x_i + \frac{1}{k}}{N_a + 1} \tag{3}$$

where x_i is the number of observations of allele i in a database, N_a is the number of alleles at that locus in the database and k is the number of allele designations with non-zero observations in the database at the locus which the allele, whose probability is being calculated, resides.

Data used for internal validation studies included DNA typing results from serially diluted single source samples (0.031–1 ng) amplified in duplicate. Additionally, a total of 290 two, three, four and five-person mixture profiles, prepared using DNA from thirteen contributors with varying individual template amounts (0.006–3.2 ng) and total template amounts (0.019–4 ng), were created in a range of contributor ratios (Table S2). Two DNA extracts used for mixture preparation were artificially degraded by UV irradiation as described. Some contributor samples were selected based on alleles shared with other contributors or unique to a single contributor (obligate). The resulting profiles variously exhibited inter- and intra-locus peak height variation, complete profile recovery, allele and locus dropout, DNA degradation, additive effects of allele or stutter peak sharing and peak saturation (off-scale peaks). APH for a given contributor was calculated as the average of the heights of all obligate alleles, with undetected obligate alleles assigned a peak height of either 0 or 25 rfu, as noted.

Adjudicated case studies entailed 30 evidentiary mixtures that had been previously developed using Identifiler® Plus and reported as originating from two to five individuals. STRmix™ analyses included the reference DNA profiles of one to four subjects of investigation.

DNA typing results from all single source, mixed and forensic specimens were exported from GeneMapper® ID-X and imported

into STRmix™, where they were interpreted using the established laboratory-specific parameters in STRmix™ version 2.3.06 (http://strmix.esr.cri.nz/). A subset of five-person mixtures was examined using STRmix™ version 2.3.06. Subsequently, the complete set of five-person mixtures was interpreted using STRmix™ version 2.4.02. STRmix™ analyses of single source profiles were conducted according to the propositions:
H_1: The DNA originated from the person of interest
H_2: The DNA originated from an unknown, unrelated individual

For mixtures of N contributors, STRmix™ analyses were conducted according to the propositions:
H_1: The DNA originated from the person of interest and N-1 unknown, unrelated individual(s)
H_2: The DNA originated from N unknown, unrelated individuals

Obligate alleles, template amount and APH were evaluated relative to the STRmix™ results.
Conditional STRmix™ analyses of some mixtures were performed assuming the presence of DNA from a known individual, according to the propositions:
H_1: The DNA originated from the assumed individual, the person of interest and N-2 unknown individual(s)
H_2: The DNA originated from the assumed individual and N-1 unknown individual(s)

Also, for some mixtures, multiple contributors were assessed concurrently in STRmix™, according to the proposition:
H_1: The DNA originated from person of interest 1, person of interest 2 and N-2 unknown individual(s)
H_2: The DNA originated from N unknown individuals.

Where indicated for H_2-true tests, two-hundred non-contributor profiles, which were artificially constructed in Excel™ by randomly sampling alleles from the FBI's U.S. Caucasian allele frequency database [23–26] based on their observed frequency, were analyzed in STRmix™ as persons of interest.

For one, two and three-contributor profiles, STRmix™ analyses were also performed assuming $N+1$ contributors. To assess $N-1$ contributors in a manner that would not result in an exclusion outright (i.e., as would five alleles per locus under the assumption of two contributors), some two-person mixture results were modified to simulate a third contributor with no new alleles. This was done by manipulating the input files directly in Excel™, as follows: in order to avoid creating a 5th allele, a 'child' of the two contributors was constructed by adding 50 rfu to the peaks selected to be shared by parent and child. Two additional mixtures were thus created by increasing peak heights by 100 rfu and 200 rfu. With this approach, a virtual child, present as a trace or minor contributor, represented a third contributor in a mixture that could be interpreted as a two-person mixture.

The default MCMC number of accepts (100,000 burn-in and 400,000 post burn-in) were used to assign weights, which were used in the calculation of *LR*s in STRmix™ using the population genetic model described in Balding and Nichols [21], referencing equations 4.10 in NRC II [22] to correct for population substructure. Statistical calculations were based on allele frequencies from the FBI U.S. Caucasian database following STRmix™ execution in either (a) the standard analysis mode with a θ point estimate of 0.01 or (b) the Database Search mode.

The Database Search function, which produces a total "investigative" *LR* that does not include θ in the calculation, was used to facilitate rapid consideration of a large number of non-contributor propositions (i.e., specificity testing), as well as known contributor propositions (i.e., sensitivity testing) for the mixtures summarized in Table 1 [27]. Some mixtures were interpreted or reinterpreted using the standard mixture analysis mode to develop a lower-bound highest posterior density (HPD) *LR*, in addition to a total "evaluative" *LR*, with the HPD interval set to 99.0% and the NI calculation [28] enabled. For sensitivity and specificity assessment, equations derived from a best fit regression analysis of the investigative *LR* and evaluative HPD *LR* results of the same mixtures (Fig. S1) were applied to the investigative *LR* values to derive HPD *LR* estimates. Based on these estimates, tests that generated a HPD $LR<1$ for H_1-true propositions and a HPD $LR>1$ for H_2-true propositions were re-analyzed in STRmix™ for verification purposes. Given that the equation derived for four-person mixtures was suitable for application to the five-person mixtures for purposes of establishing HPD *LR* estimates, re-analyses in STRmix™ to develop HPD *LR*s were not performed for five-person mixtures due to computational constraints.

Precision was evaluated through five repeated interpretations of one, two, three and four-person typing results, with both the minor and major contributor considered the person of interest in H_1. The effect on repeatability by increasing the total number of MCMC accepts from 500,000 to 600,000 and 700,000 and the Random Walk Standard Deviation (RWSD) from 0.005 to 0.01 and 0.02 was evaluated.

STRmix™ analyses of the adjudicated case mixtures were performed in standard analysis mode using the U.S. Caucasian database ($\theta=0.01$) or, where match probabilities had been reported for Native Americans, a Navajo database [25,26] ($\theta=0.03$) to generate total *LR*s and HPD *LR*s. For some assessments, the reciprocal of HPD *LR*s between 0 and 1 was calculated.

For both the evidentiary mixtures and a subset of the prepared mixtures, the results of STRmix™ and manual analyses of the same data were compared to evaluate general consistency of the results. Manual profile interpretations and calculation of random match probabilities (RMPs) and combined probabilities of inclusion (CPIs) were performed in accordance with FBI Laboratory standard operating procedures, including usage of a stochastic threshold (200 rfu).

3. Results and Discussion

3.1. Verification of model performance, accuracy and precision

For a small subset of profiles, the *LR* is evident without calculation or can be estimated easily as described in Bright *et al.* [29]. These include single source profiles where the genotype at

Table 1
Summary of mixtures and propositions tested in STRmix™.

Number of contributors	Contributor template range	Total mixture template range	Contributor ratio range	Number of mixtures interpreted	Number of H_1-true propositions tested	Number of H_2-true propositions tested
2	0.006 to 0.9 ng	0.019 to 1 ng	10:1 to 1:1	105	202	22,504
3	0.021 to 1 ng	0.38 to 3 ng	16:1:1 to 1:1:1	64	192	13,620
4	0.05 to 3.2 ng	1 to 4 ng	16:1:1:1 to 1:1:1:1	84	336	17,808
5	0.016 to 1.25 ng	0.25 to 2 ng	10:1:1:2:2 to 1:1:1:1:1	24	120	5,256

each locus is unambiguous, and hence the weight for the correct genotype combination is expected to be 1. As an initial verification of software performance, manually calculated *LR*s for individual loci in a single source sample were identical to the corresponding *LR*s produced by STRmix™, and STRmix™ reported correct genotypes for the known contributor.

As an additional verification of model performance, using a serially-diluted single source sample, *LR*s based on STRmix™ analyses were demonstrated to decrease with template amount (1–0.03 ng) (Fig. S2). As expected, the *LR* progressed from the maximum value for the full profile (attained at ≥0.25 ng) towards log(*LR*) = 0 due to allele dropout as DNA template decreased [11]. For profiles exhibiting higher levels of dropout, simultaneous analysis of amplification replicates in STRmix™ resulted in a higher *LR*. The mass parameter *t* (template, or DNA amount) in the STRmix™ output declined steadily with decreasing peak heights, as expected (Table 2).

Weights generated by STRmix™ were assessed as a measure of the deconvolution process and model performance. Any contributor genotypes deconvoluted manually by a skilled analyst should exhibit intuitively correct, high weights and thereby indicate proper modeling. The most weight is expected to be assigned to genotype sets that correspond to the genotypes of the DNA donors. The result of such an assignment of weight is high levels of support for the inclusion of DNA contributors and exclusion of non-contributors when assessed using *LR*s (depending on profile quality) [11]. Counterintuitive weights are an indication of poor biological modeling or incorrect tuning of the models. Inspection of the STRmix™ output, including weights over the range of single source and mixed DNA profiles, showed the anticipated response to relative template amounts, with lower weights for genotypes that exhibited allelic dropout (Fig. S2).

Mixture proportions were assessed as a final check of model performance, using two-person mixtures constructed in the ratios 1:10, 1:5, 1:3 and 1:1. Mixture proportions obtained from the STRmix™ output (1:10.1, 1:4.0, 1:1.9, 1:1.0) were similar to the targeted proportions. The minor differences from the expected values may be attributable to variability of quantitative PCR results and/or pipetting. A plot of log(*LR*) for each mixture type considering both the major and minor contributors is provided in Fig. S3. The maximum potential log(*LR*s) based on a single source, full profile for each contributor are plotted as horizontal dashed lines. As expected, the *LR* calculated for the major contributor trended from the maximum potential *LR* at 10:1 downward, with the lowest *LR* produced for the 1:1 mixture. The decrease in *LR* occurs where peak heights of major and minor alleles begin to fall within heterozygous peak height variance expectations (here, less than 1:3). The *LR* for the minor contributor did not reach the maximum potential *LR* for a single source profile. At 1:10, some alleles from the minor contributor may be dropped, masked by major contributor alleles, or obscured by stutter peaks from the major contributor. At 1:5, an increase in *LR* for the minor contributor was observed, perhaps as the distinction between minor contributor and major contributor stutter peaks is greater and allele sharing is more evident from assessment of peak heights. At 1:3 and 1:1, the *LR* decreased as the minor and major contributor alleles were less distinct. These data demonstrate that as an analyst's ability to manually deconvolute a mixture decreases, the weights assigned to genotype sets also decrease and are reflected in lower *LR*s.

Table 2
Summary of single source dilution series interpreted in STRmix™ demonstrating the decrease in *LR* with APH and template quantity, *t*.

Input DNA, ng	log(*LR*)	log(HPD)	APH, rfu	*t*
1	18.38	17.71	1366	1975
	18.38	18.11	1241	1935
0.5	18.38	18.12	612	929
	18.38	18.01	618	861
0.25	18.38	18.06	325	484
	18.38	18.01	226	302
0.125	16.99	16.48	138	190
	14.58	14.15	151	194
0.063	9.02	8.82	107	115
	11.70	11.37	101	99
0.031	6.12	5.89	82	61
	4.66	4.35	68	42

For the mixed typing results, *LR*s for both the major and minor contributors varied within one order of magnitude (comparing the minimum and maximum *LR*) across five repeated interpretations, as expected due to MCMC sampling [30]. For mixtures with similar donor contributions (1:2 and 1:1 ratios), greater *LR* variability was occasionally encountered but typically still fell within two orders of magnitude. While increasing the number of MCMC iterations and RSWD might be expected to improve repeatability, no consistent benefit was observed in such trials relative to the minor contributors tested. However, given the conservatism inherent in the use of NRC II equations 4.10 and θ point estimates in STRmix™, along with a HPD interval set to 99.0%, the observed variability in *LR* is within acceptable levels [31–34].

3.2. Sensitivity and specificity studies

Sensitivity of a probabilistic genotyping system refers to the ability of the software to reliably support the presence of a contributor's DNA within the DNA typing results. Sensitivity studies demonstrate the propensity of the system to return support for H_2 for a H_1-true test (i.e., the presence of a true contributor's DNA in the profile is not supported) [10]. It should be noted, however, that failure to detect alleles and/or return support for the presence of a low-level known contributor do not necessarily constitute an error in the analytical process or probabilistic genotyping system. In general, the *LR* for a true contributor should be high but trend to 1 as less typing information to aid interpretation is available and as the number of contributors increases. Information that aids interpretation includes the detection of more alleles from a given contributor, a conditioning profile (e.g., from the donor of an intimate sample) and replicate amplification results from a DNA sample.

Specificity of a probabilistic genotyping system refers to the ability to reliably support the absence of a non-contributor's DNA within the DNA typing results. Specificity studies demonstrate the propensity of the system to return support for H_1 for a H_2-true test [10].

For any mixture study, the proportion of analyses that provide support for a true or false hypothesis (H_1 or H_2) is dependent on the design and quantities of the mixtures tested and should not be used as an indication of error expectations. In this study, the specificity and sensitivity of STRmix™ were tested using a total of 277 two, three, four and five-person mixtures (Table 1). These mixtures exhibited the range of features typically encountered in forensic casework and included many challenging specimens (e.g., with degraded DNA and/or multiple contributors of similar low-level template quantities), specifically to identify potential limitations of the software.

The Database Search mode of STRmix™ was initially used to efficiently execute more than 60,000 comparisons to a 'database' of the true mixture contributor profiles and 200 non-contributor profiles, for a total of 855 H_1-true tests and 59,188 H_2-true tests. The known number of contributors to these mixtures was used for these analyses. Generally, at high template levels STRmix™

returned high *LR*s for true contributors and *LR*s of zero for non-contributors (Figs. S4 through S7).

As contributor template amount decreased and/or contributor number increased, *LR*s trended to 1. The *LR*s for 255 mixtures were converted to HPD *LR* estimates according to the equations in Fig. S1. The HPD probability interval provides a one-sided, lower bound point value based on a *LR* distribution that reflects both population sampling effects and MCMC variability. The HPD *LR* can thus be interpreted and reported in a manner similar to a confidence interval. All propositions generating an HPD *LR* estimate that provided support for the false proposition were re-analyzed for assessment of a STRmix™-generated HPD *LR*. The estimated and

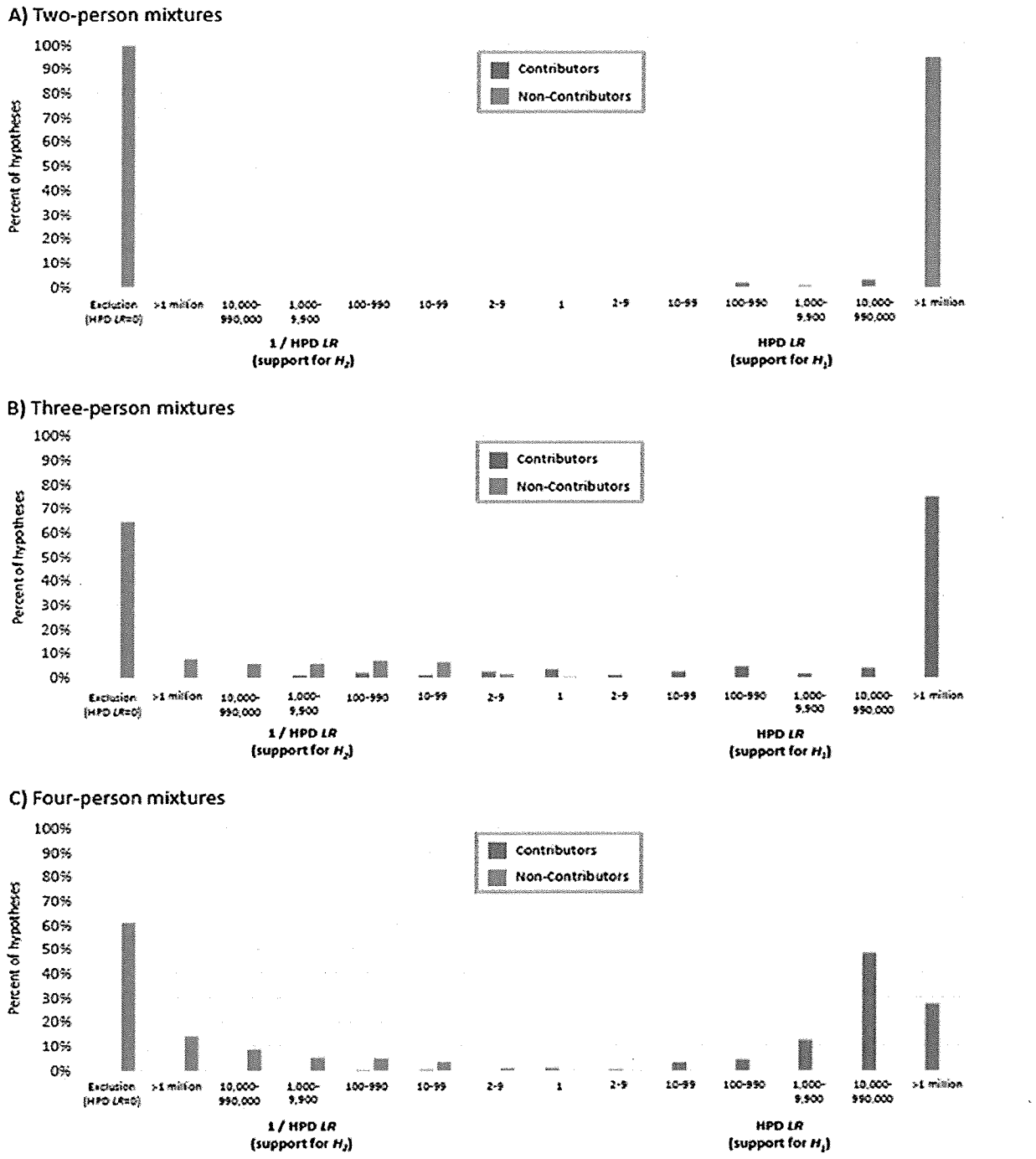


Fig. 2. Sensitivity and specificity of STRmix™ interpretation of two, three, four and five person mixtures. The plots display estimated 99.0% one-sided lower-bound HPD *LR*s for true contributors and known non-contributors proposed as contributors to (A) two-person, (B) three-person, (C) four-person, and (D) five-person mixed DNA profiles. The number of mixtures and propositions tested is provided in Table 1; results for tests using poor quality profiles, as discussed, are not plotted. HPD *LR*s greater than 1 indicate support for the H_1 (contributor) hypothesis, whereas HPD *LR*s less than 1 (here, converted from decimals to positive integers by taking the reciprocal) indicate support for the H_2 (non-contributor) hypothesis. Here, sensitivity is indicated by the percentage of H_1-true propositions (blue bars) within the blue-shaded area (HPD $LR > 1$) versus the percentage in the orange-shaded area (HPD $LR < 1$). Similarly, specificity is indicated by the percentage of H_2-true propositions (orange bars) within the orange-shaded area compared to the blue-shaded area. In total, the histograms in panels A–D represent more than 800 known contributor (H_1-true) propositions, and more than 50,000 non-contributor (H_2-true) propositions. Panel E shows the improvement in specificity for five-person mixtures when interpretations are conditioned on a known contributor.

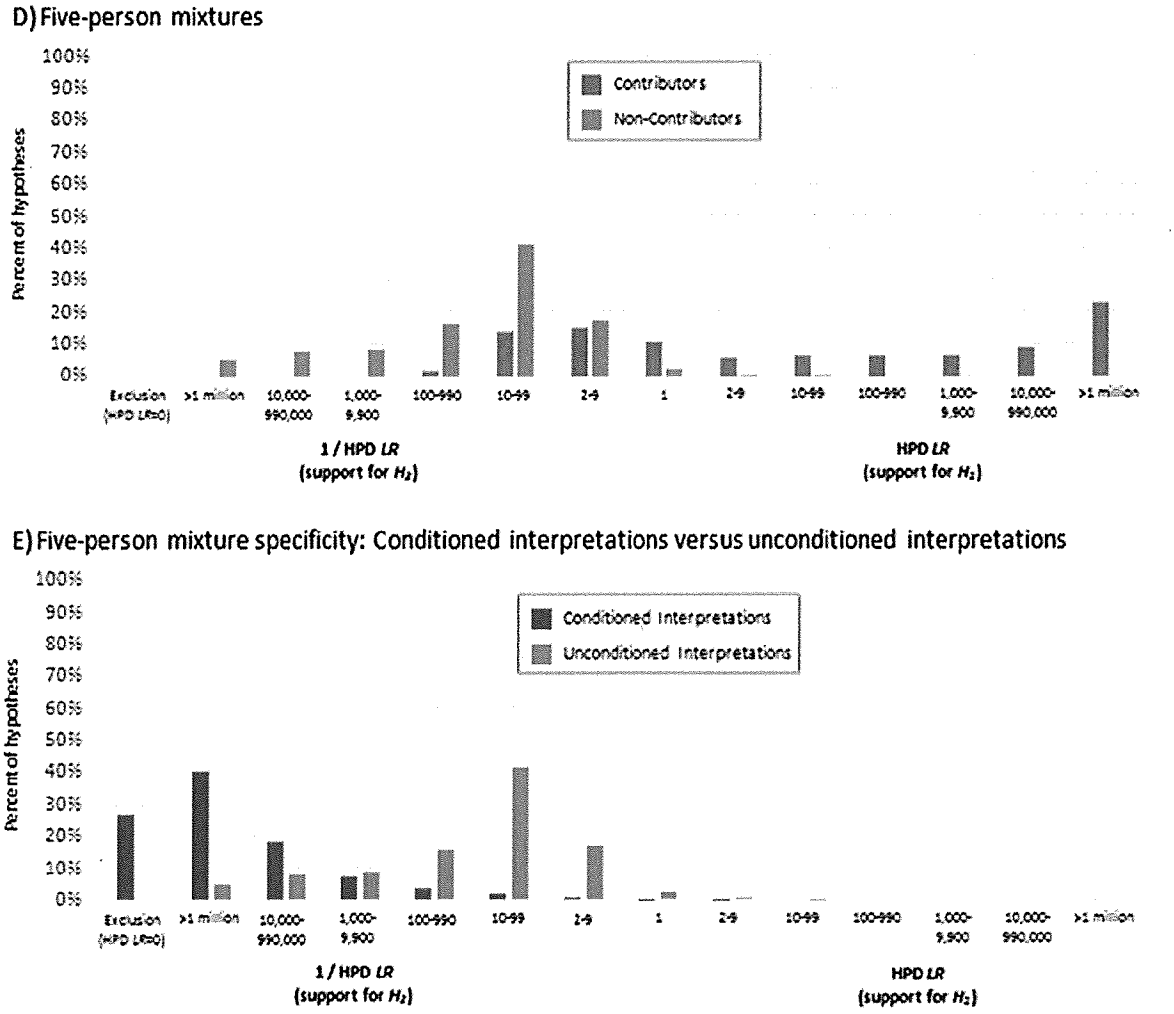


Fig. 2. (Continued)

calculated HPD *LR*s were found to be different by only 0.16 orders of magnitude on average. Overall, the HPD *LR*s indicated very high sensitivity and specificity of STRmix™ analysis (Fig. 2).

For two-person mixtures (Figs. 2 and 3) spanning ratios of 1:1–1:20 and donor contributions of 0.006 ng to 0.90 ng, all H_1-true propositions ($N = 166$) for both contributors resulted in HPD *LR*s > 100. The majority of HPD *LR*s (95%) exceeded 1 million. Nearly all H_2-true propositions for the two-person mixtures produced exclusions (HPD $LR = 0$). All non-zero *LR*s for H_2-true propositions correctly provided support for H_2, with the exception of two that incorrectly provided support for H_1 (discussed below).

For three-person mixtures (Figs. 2 and 3) spanning ratios of 1:1:1–1:1:16 and donor contributions of 0.02 ng to 1 ng, 87% of H_1-true propositions ($N = 192$ total tests) resulted in HPD *LR*s > 1. As with two-person mixtures, the majority of HPD *LR*s (73%) exceeded 1 million. A small portion of H_1-true propositions for the three-person mixtures resulted in HPD *LR*s < 1. These findings occurred when little or no indication of the known contributor was observed in the STR profile due to DNA levels typically <50 pg, allele sharing and dropout; in fact, only two H_1-true propositions produced *LR*s < 1 when the APH and number of obligate alleles for the minor contributors were >0. Five false exclusions were returned for true contributor tests due to poor profile quality (unresolved alleles differing in size by 1 bp; discussed below). As compared to two-person mixtures, fewer of the non-contributor propositions (65%) ($N = 13{,}620$ total tests) resulted in exclusions, though most HPD *LR*s (>99%) were less than 1 and correctly indicated support for H_2. Incorrect H_1-support resulted from 26 H_2-true propositions for the three-person mixtures (discussed below). Given that these analyses were conducted using the known number of contributors to the mixtures, and because several of the incorrect support results occurred with an undetected known minor contributor (APH and number of obligate alleles for the minor contributor equaling 0), the mixtures producing false H_1-support were recalculated in STRmix™ using the observed number of contributors. Based on this re-analysis, 7 of 13,620 non-contributor propositions generated *LR*s > 1 (2–27) (Table 3).

For four-person mixtures (Figs. 2 and 3) spanning ratios of 1:1:1:1 to 1:1:1:16 and donor contributions of 0.05 ng to 3.2 ng, 96% of H_1-true propositions ($N = 336$) resulted in HPD *LR*s > 1, and 27% exceeded 1 million. For H_1-true propositions, one mixture and its replicate amplification produced false-H_2 support for a known minor contributor that was present in both mixtures at 50 pg. Five false exclusions were returned for true contributor tests due to poor profile quality ("saturated" allelic peaks derived from 4 ng contributor template inputs; discussed below). All H_2-true

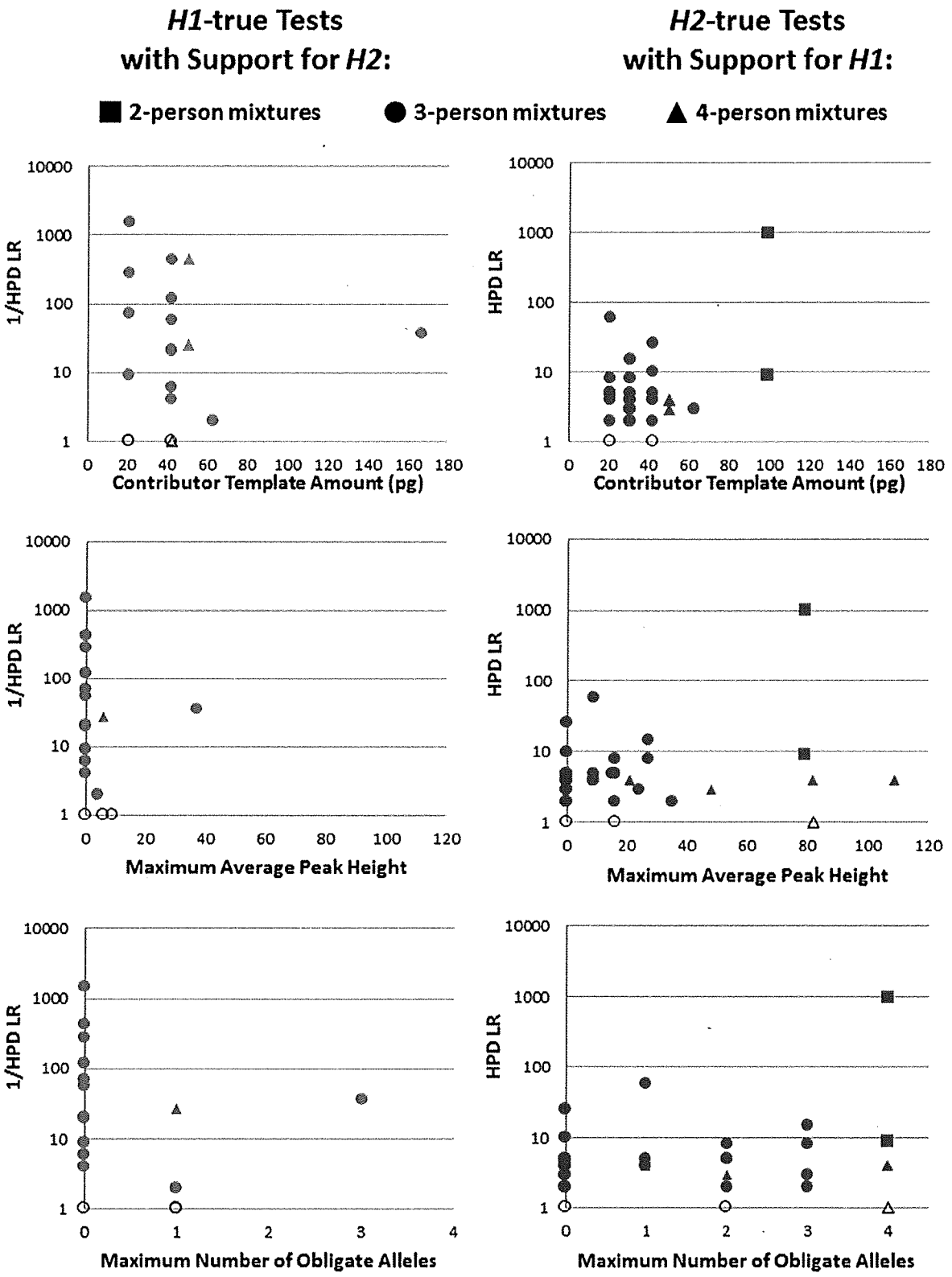


Fig. 3. HPD *LRs* that provided incorrect support for H_1 or H_2. For the contributor and non-contributor tests for which estimated HPD *LRs* indicated incorrect support for H_1 or H_2, respectively, actual HPD *LRs* were calculated using the standard mixture analysis mode and using the known number of contributors to the mixture. The number of mixtures and propositions tested is provided in Table 1. In the H_1-true plots, average peak heights (APH) are based on the rfu values for all obligate (unshared) alleles for the contributor tested, with dropout of an obligate allele captured as rfu = 0. As the known number of contributors was used for the interpretations, rather than the number of contributors that would be inferred via visual review of the electropherogram, in some instances no obligate alleles were detected for a minor contributor, and thus APH also equals zero. All x-axis values for the non-contributor tests are based on the highest value for any minor contributor to the mixture. The open markers in the plots indicate HPD *LR* = 1 (inconclusive) results.

Table 3
HPD *LR* results that failed to support the correct hypothesis.

(A) H_1-True Hypotheses

# of Contributors	Contributor Ratio	Mixture Identifier	Person of Interest	Contributor Template (pg)	Obligate Alleles	Contributor APH	1/HPD *LR*	Inconclusive (HPD *LR* = 1)
3	16:1:1	C.1	Known: J	21	1	9		1
		C.2	Known: J	21	0	0		1
3	8:1:1	D.1	Known: F	42	0	0	20	
		D.1	Known: J	42	0	0	120	
		D.2	Known: F	42	0	0	6	
		D.2	Known: J	42	0	0	21	
3	16:1:1	E.1	Known: J	21	0	0	9	
		E.2	Known: J	21	0	0	280	
3	2:1:1	F.1	Known: B	167	3	37	36	
3	8:1:1	G.1	Known: B	63	1	4	2	
3	8:1:1	H.2	Known: B	42	0	0	57	
		H.2	Known: F	42	0	0	430	
3	16:1:1	J.1	Known: B	21	0	0	71	
		J.2	Known: F	21	0	0	1500	
3	8:1:1	K.1	Known: B	42	1	6		1
		K.2	Known: B	42	0	0	4	
4	16:1:1	O.1	Known: A	50	1	6	26	
		O.2	Known: A	50	0	0	460	

(B) H_2-True Hypotheses

# of Contributors	Contributor Ratio	Mixture Identifier[a]	Non-Contributor Identifier[b]	Contributor Template (pg)	Obligate Minor Alleles	Maximum Minor Contributor APH	HPD LR	Inconclusive (HPD *LR* = 1)
2	10:01	A.1	Rand: 148	100	4	79	980	
		A.1	Rand: 179	100	4	79	9	
3	16:01:01	C.1	Rand: 19	21	1	9	27	
		C.1	Rand: 157	21	1	9	20	
		C.1	Rand: 62	21	1	9		1
3	16:01:01	I.1	Rand: 12	31	0	0	10	
		I.1	Rand: 129	31	0	0	7	
		I.1	Known: Q	31	0	0	2	
		I.1	Known: T	31	0	0	2	
		I.2	Rand: 199	31	0	0	8	
		I.2	Rand: 154	31	0	0		1

STRmix results are shown for all (A) H_1-true hypotheses that returned support for H_2 or HPD *LR* = 1, followed by (B) H_2-true hypotheses that returned support for H_2 or HPD *LR* = 1.

[a] Each mixture is designated with a letter (e.g., C), with replicate amplifications of the mixed DNA extract designated as .1 and .2.

[b] The mixtures were analyzed in STRmix™ with the specified POI.

propositions with the exception of 5 produced *LRs* < 1 (discussed below). Two of the incorrect H_1-support results occurred with the APH and number of obligate alleles for the minor contributor equaling 0. After re-analysis using the observed rather than the known number of contributors, none of the 17,808 non-contributor propositions generated *LRs* > 1 (Table 3).

For five-person mixtures (Figs. 2 and 3) spanning ratios of 1:1:1:1:1–1:1:2:2:10 and donor contributions of 0.02 ng to 1 ng, 58% of H_1-true propositions (N = 120) resulted in HPD *LRs* > 1, and 23% exceeded 1 million. This lowered sensitivity is expected given the high number of contributors per sample and very low DNA inputs for some contributors. Yet results from the H_2-true tests indicated that specificity is still high with five-person mixtures: 97% of HPD *LRs* were <1 (N = 5256 total tests). Specificity was further improved when the STRmix™ interpretations were conditioned on one known contributor. For these analyses, >99% of H_1-true propositions resulted in HPD *LRs* < 1 (N = 3200), and only four resulted in incorrect H_1 support.

Overall across 60,000 STRmix™ analyses, the majority of HPD *LRs* indicating false H_2-support for known contributor propositions ranged from 2 to 2,200 and occurred with 3 or more contributor mixtures that generally exhibited 0 or 1 obligate minor alleles. False H_1-support for known non-contributors was typically demonstrated by *LRs* < 10 (though the highest *LR* was 980) for mixtures of three or more contributors exhibiting 0 to 4 obligate minor alleles. To further explore the results demonstrated with low level DNA contributions, STRmix™ Database Search results for known contributor testing (H_1-true) and non-contributor testing (H_2-true) for the three and four-person mixtures are provided in Fig. 4, a plot of log(*LR*) versus average peak height of the individual contributors for profiles where individual contributions were less than 100 pg. While the HPD *LRs* calculated for these same tests (discussed and tallied above as, for example, incorrect H_1 support for three and four-person mixtures) are lower for true contributors than the *LRs* shown in Fig. 4 (and in some instances returned correct support whereas the Database Search returned incorrect support), the results indicate that STRmix™ was able to correctly separate out true from false contributors even at low levels of DNA. Some low level false positive results were observed, as expected given the complexity of mixtures (i.e., number of contributors and

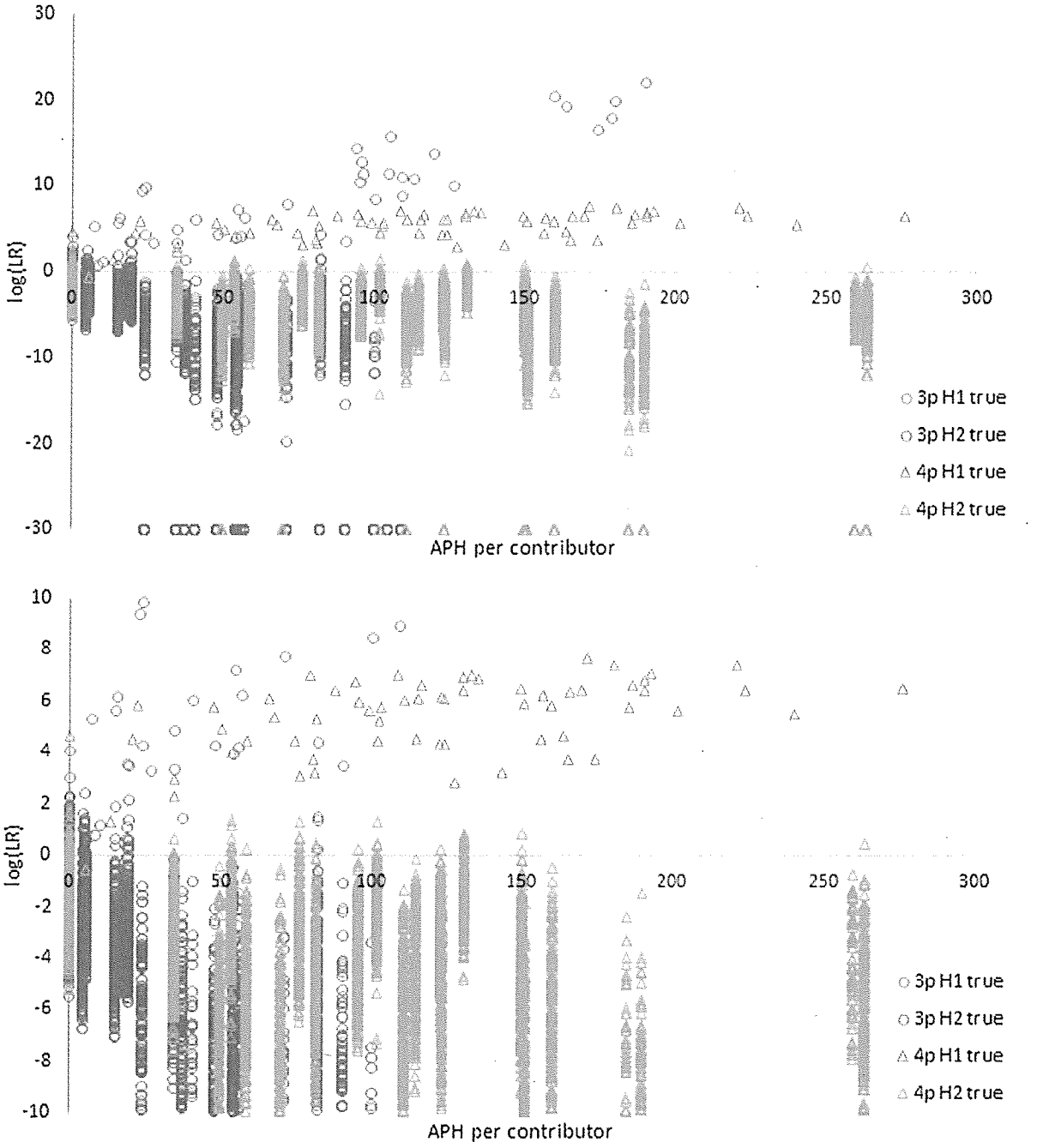


Fig. 4. Plot of log(*LR*) versus average peak height (APH) per contributor for three and four person mixtures where individual contributions are < 100 pg DNA. In the second pane the y-axis has been truncated at +/– log(LR) = 10 in order to better see the results.

low level DNA contributions). To further assess the utility of STRmix™ for discerning minor contributors, the range of HPD *LR* estimates for all minor contributors to the two and three-person mixtures was considered with respect to differing contributor ratios, DNA inputs and, for the two-person mixtures, DNA degradation (Fig. 5). As would be expected due to profile quality, comparisons to two-person mixtures in which one or both DNA extracts were degraded resulted in both lower and more variable HPD *LR* values than when no degradation was present. However, aside from the lowest DNA input tested for the degraded extracts set (0.006 ng), all other minor contributor HPD *LR* estimates exceeded 100,000. A few general trends were also apparent from the three-person mixture plots. As expected, the HPD *LR*s were overall more variable than was observed with the two-person mixtures, and the values generally decreased with decreasing DNA template amounts. But, within three broad template categories representing all minor contributor DNA amounts greater than 0.05 ng (colored in red, blue and purple in Fig. 5), the minor contributor HPD *LR*s increased as the gap between the major to minor contributor input increase. At template amounts below

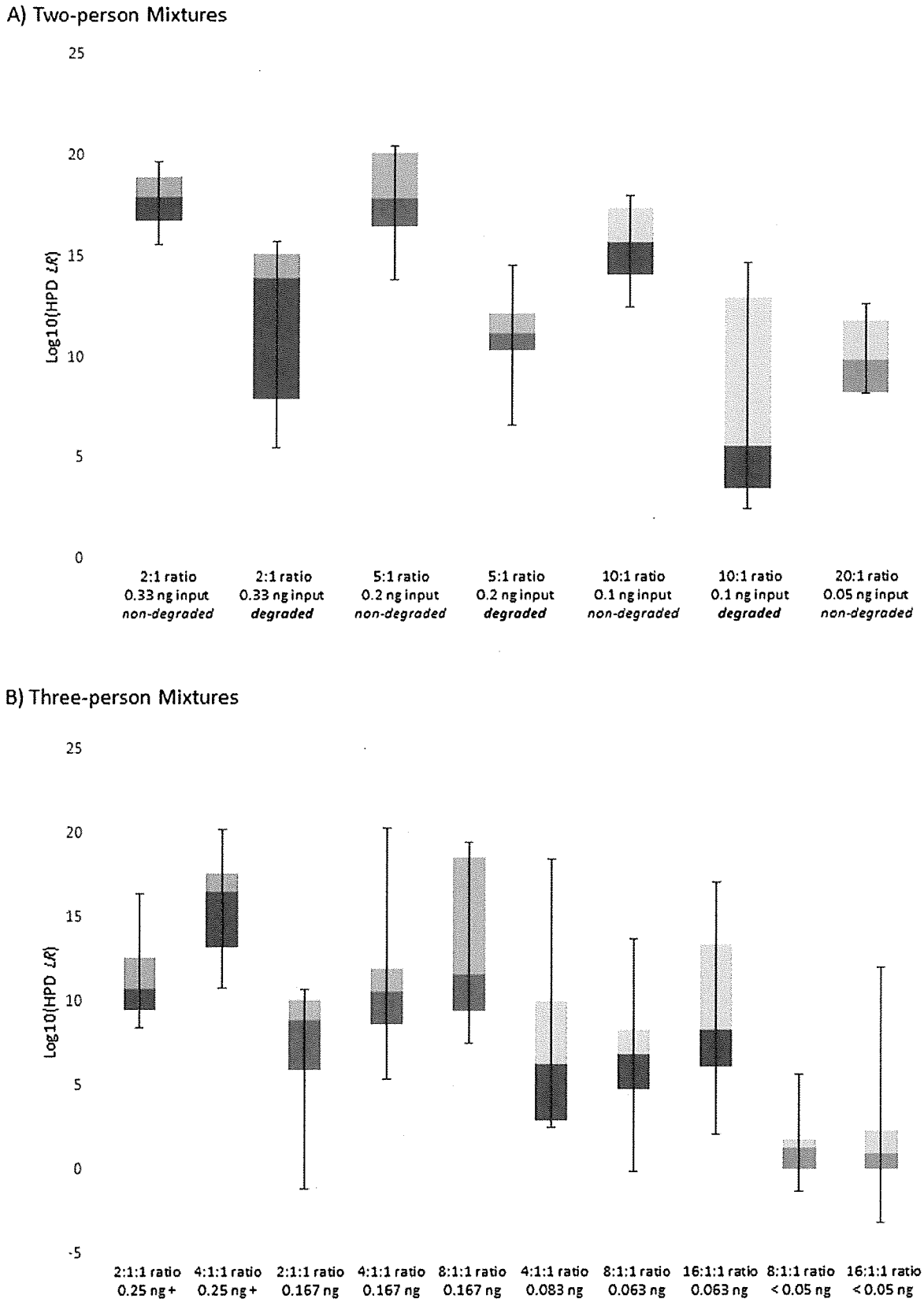


Fig. 5. Range of minor contributor HPD *LRs*. Box and whisker representation of the HPD *LR* estimates for known minor contributors to two-person (panel A) and three-person (panel B) mixtures considering DNA input of the minor contributor, major:minor contributor ratio, and (in the case of the two-person mixtures) DNA degradation of one or both contributor extracts by ultraviolet irradiation.

0.05 ng, most minor contributor comparisons resulted in HPD *LR* estimates indicating some support for H_1, but approximately 25% produced values less than 1.

There were two notable exceptions to the general conclusions from the sensitivity and specificity testing of STRmix™:

(1) Two non-contributors provided HPD *LRs* of 9 and 980 for a single 1:10 mixture (Fig. S8, panel A) developed from two DNA extracts that were each artificially degraded by ultraviolet irradiation. The mixture had a minor contribution of 100 pg. At several loci, the mixture displayed complete dropout of all alleles or all minor contributor alleles. Comparison of the known minor contributor to the mixture resulted in a low HPD *LR* estimate of 320, correctly in support H_1. For the minor contributor (APH = 79 rfu), STRmix™ deconvolution assigned only 5 alleles with >99% probability. Both non-contributor profiles included all 5 of these alleles, as well as several undetectable alleles shared with the major contributor. The absence of all other undetected alleles could be reasonably attributed to dropout. As for alleles that could not be accounted for by the non-contributors, for one of the comparisons (HPD *LR* = 980), a single additional peak was modeled as stutter (9%); for the other comparison (HPD *LR* = 9), two additional peaks were modeled as stutter (9% and 15%). Accordingly, the probabilistic interpretation relative to these non-contributors makes intuitive sense given the mixture results. Under a binary interpretation approach, there is no basis for exclusion of these non-contributors; however, because all minor contributor alleles are below a stochastic threshold (200 rfu), the mixture is not suitable for statistical analysis (i.e., CPI calculation) and would be reported as inconclusive.

A replicate amplification of the same 1:10 mixture was considered with respect to the non-contributor comparison that had produced the HPD *LR* = 980 result. Deconvolution of the duplicate amplification (Fig. S8, panel B) assigned to the minor contributor with >99% probability an allele (not detected in the first amplification) that was inconsistent with the non-contributor, resulting in a HPD *LR* = 0. When the PCR replicates were analyzed simultaneously in STRmix™, the non-contributor was also unambiguously excluded as a potential contributor to the mixture (HPD *LR* = 0). Moreover, the concurrent consideration of both mixed profiles with the true minor contributor as the hypothesized person of interest resulted in a HPD *LR* higher by 2.5 orders of magnitude than when either STR profile was interpreted alone. As a point of reference, repeated STRmix™ analyses of the same mixed profile are generally expected to produce *LRs* with a maximum of a 10-fold (one order of magnitude) difference between the highest and lowest values [30]. Thus, *LR* differences

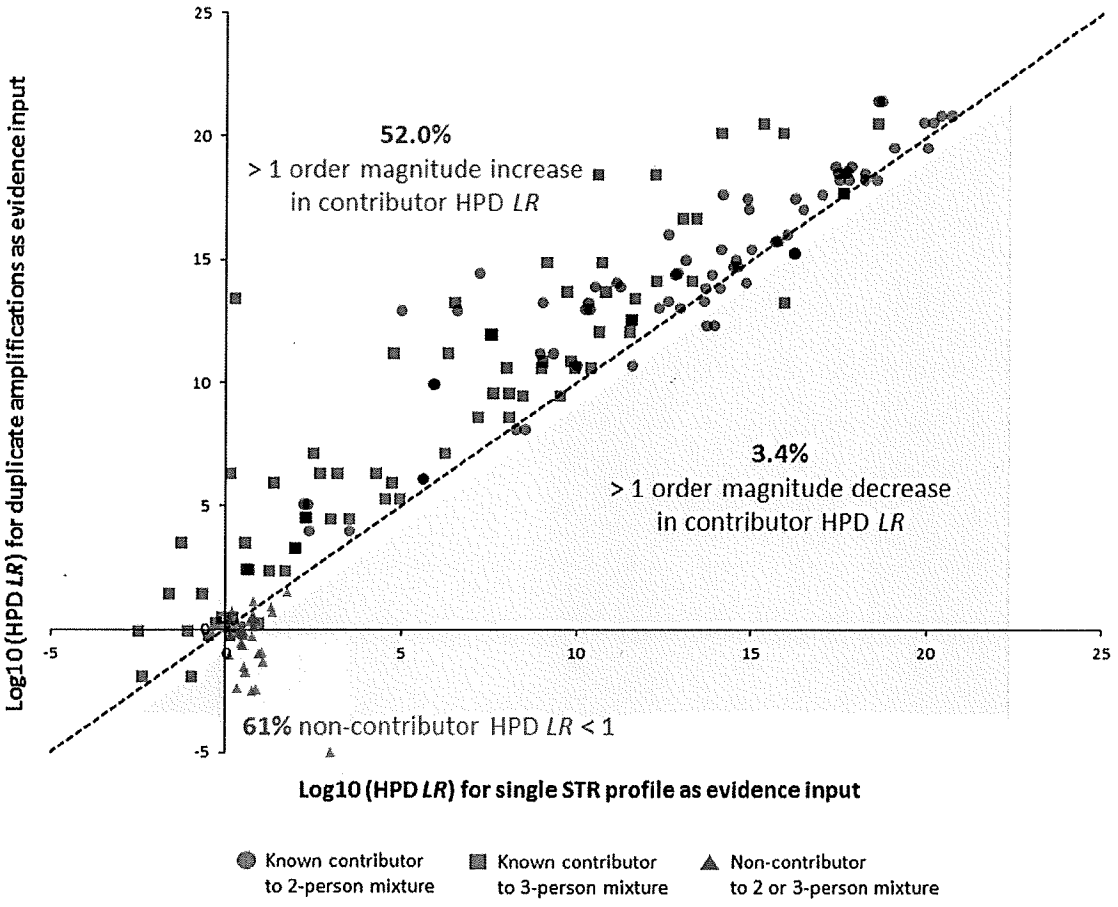


Fig. 6. HPD *LRs* resulting from interpretation of single STR profiles versus PCR replicates. Two-person and three-person mixtures were interpreted with respect to known minor contributors and non-contributors to assess the impact of considering PCR replicates. The data are plotted on a log10 scale, with HPD *LR* = 0 plotted as −5. The dashed line represents where the HPD *LRs* for the single profile (x-axis) and PCR replicates (y-axis) interpretations are equal, and the blue and green-shaded areas represent an increase or decrease (respectively) of one or more orders of magnitude. Only non-contributor propositions that produced any degree of support for inclusion (HPD *LR* > 1) for the single mixed profile interpretations were examined, thus all non-contributor data points (red triangles) are found to the right of the y-axis. The pink-shaded area highlights the non-contributor PCR replicates analyses that produced HPD *LRs* < 1 (support for exclusion).

exceeding one order of magnitude are greater than what would be expected due to variations in the statistical sampling process (MCMC) alone.

(2) One known minor contributor to a 2:1:1 mixture, with template input of 167 pg, resulted in a reciprocal HPD *LR* of 36 in support of H_2. The STRmix™ deconvolution indicated a 1:1:1 contributor ratio, which may in part have been due to incomplete resolution of a shared TH01 allele 9.3 and the minor contributor's TH01 allele 10. The inferred contributor ratio of 1:1:1 resulted in low weights for genotype sets with allelic dropout. However, only half of the minor contributor's obligate alleles were detected, and the APH for the contributor was 37 rfu. As a result, loci with allelic dropout produced *LRs* < 1, resulting in an overall HPD *LR* that incorrectly supported H_2.

The *LRs* from sensitivity and specificity testing in STRmix™ tend towards one as the information in the profile declines, usually correlating with lower template amounts (Figs. 2 and 4 and S4 through S7). Where profiles exhibit stochastic effects and allele dropout, particularly at very low template where few or no obligate alleles for a given contributor are detected, the *LR* for false contributors (as well as true contributors) tends to spread slightly above and below one. Given probabilistic modeling within the stochastic range, *LRs* >1 are expected for some non-contributors. In validation testing, failure to demonstrate false support would indicate that the system is either not functioning properly or has not been queried with sufficiently challenging specimens. In fact, a high *LR* for a simulated non-contributor may even result from a high template single source profile, since simulation of a large number of non-contributor genotypes will eventually produce one that matches the profile. In general, however, the results demonstrate the accuracy of support (i.e., inclusionary or exclusionary), with a much greater probability of excluding a true contributor (6.1% of H_1-true tests) than of including a non-contributor (0.1% of H_2-true tests).

3.3. Concurrent STRmix™ analysis of replicate amplification results

Analyses of duplicate amplifications of two and three-person mixtures ($N = 56$) yielded improvements in the *LR* results. Considering known minor contributors across a broad range of DNA template quantities and contributor ratios, 52% of concurrent analyses of PCR replicates produced a HPD *LR* at least one order of magnitude higher than when an amplification result was analyzed singly, and approximately one third of the time the HPD *LR* increased by two or more orders of magnitude (Fig. 6). Notably, over a range of a HPD *LRs* from10 to 100 million for a single amplification, with the addition of amplification replicate results, the HPD *LR* was never reduced. In fact, 77% of HPD *LRs* increased by

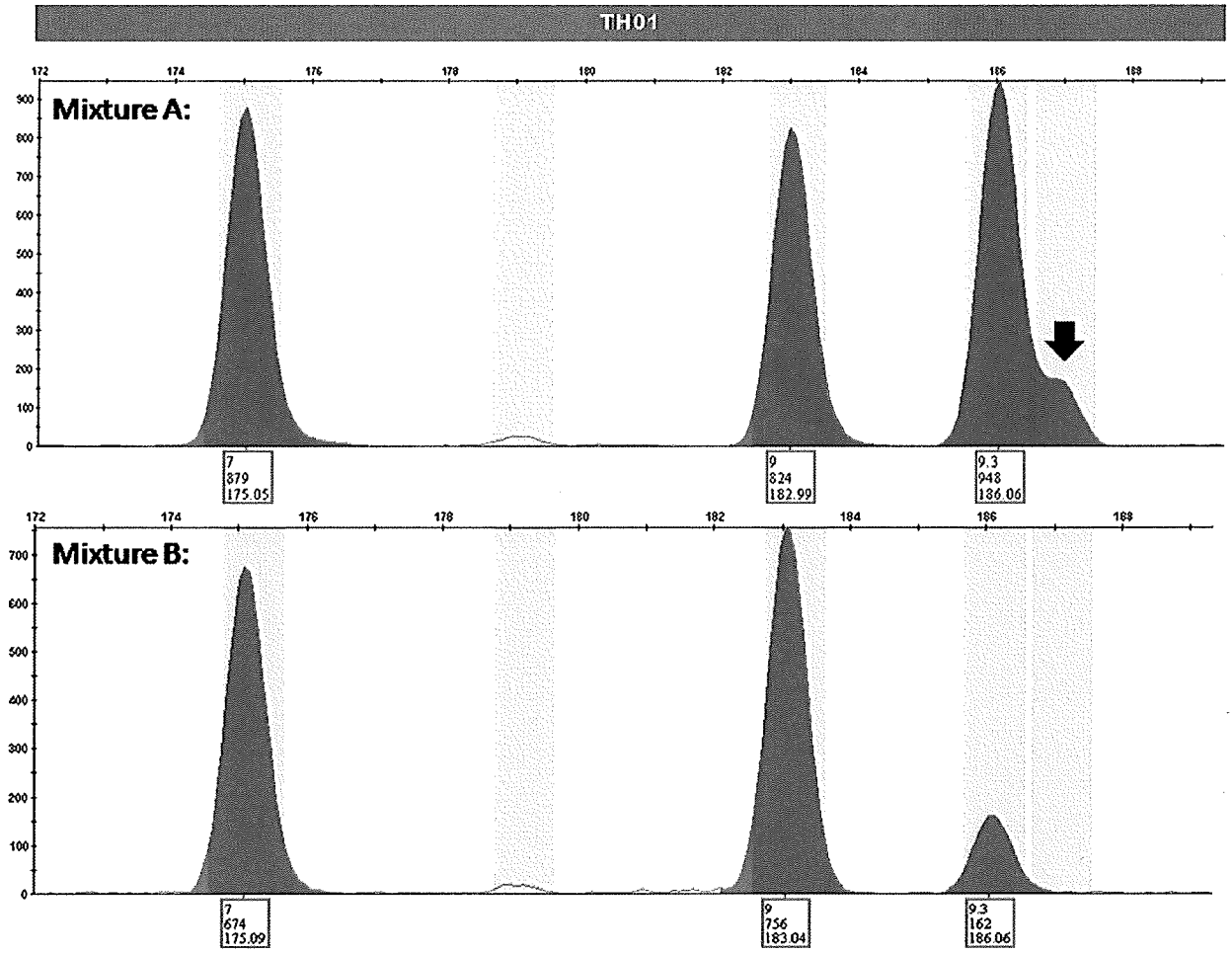


Fig. 7. Typing results at the TH01 locus generating false exclusion results (*LR* = 0) by STRmix™. An unresolved allele 10 (arrow) is apparent for the minor contributor to Mixture A (the multi-locus electropherogram is provided as Fig. S9). Mixture B (Fig. S10) exhibits all alleles at TH01 for its two contributors. STRmix™ unambiguously determined the major contributor genotype to be 7,9, but did not consider 9.3,9.3 as a possible minor contributor type. A *LR* = 0 was returned for the known minor contributor, 9.3,9.3.

more than one order of magnitude (on average by 2.8 orders of magnitude). In addition, 33 non-contributor propositions that had produced some degree of support for inclusion (HPD *LR* > 1) were reanalyzed with the addition of a duplicate amplification. In 61% of these analyses, concurrent interpretation of the PCR replicates resulted in a HPD *LR* less than 1 (Fig. 6). Overall, these data demonstrate that the inclusion of more information in the form of replicate amplifications tends to improve both sensitivity and specificity. However, in practice, replicate amplifications (e.g., with the same template amount) are not routinely performed.

3.3.1. False exclusions

In limited instances, STRmix™ returned readily recognizable false exclusions (*LR* = 0) for known contributors. These had two distinct causes: profile quality and mixture deconvolution issues. False exclusions due to poor profile quality included (a) saturated peaks and (b) electrophoretic failure to resolve major and minor alleles differing in size by 1 bp.

In the present study, false exclusion occasionally occurred in saturated three and four-person mixtures for which the major contribution was 2 ng or more. Additionally, some H_2-true propositions for saturated profiles resulted in inconclusive results (HPD *LRs* = 1). While some saturated peaks may have a nominal effect on *LRs* and weights in some STRmix™ analyses, it is advisable to reprocess the sample (e.g., inject for less time for capillary electrophoresis), given that no useful quantitative information is associated with such peaks and there is a greater potential for elevated stutter, electrophoretic artifacts resulting from amplification of high template amounts and false exclusion with saturated data.

False exclusions also occurred in a three-person mixture presenting with a 3:1:1 ratio. STRmix™ unambiguously determined the major contributor genotype at all loci except for TH01, which exhibited three alleles (Fig. 7, Mixture A; the electropherogram is provided as Fig. S9). At this locus the genotypes of the three known contributors are 7,9 (major contributor), 9.3,9.3 and 9.3,10. On inspection of the electropherogram (Fig. 7), the 10 allele corresponding to the third contributor appears to have been unresolved from the 9.3 during capillary electrophoresis (the 10 allele was thus not 'called' by GeneMapper ID-X and therefore not subsequently analyzed by STRmix™). Based on the modeling of template and degradation for the third contributor, STRmix™ did not consider dropout at TH01 (of, say, the unresolved allele 10), and therefore a *LR* = 0 was returned for this locus only (Table 4, Mixture A). As a general indicator of a potential problem with the data/analysis, review of the STRmix™ output data showed intuitively correct results for all loci except for TH01. Five repeat reinterpretations of the profile with the unresolved allele returned LR = 0 or a very low LR providing incorrect H_2 support (approximately 10^{-4}), as expected given the flawed input data. A replicate amplification of the sample resolved the minor allele 10 and provided a *LR* of 2×10^9. If the unresolved/uncalled peak were not evident in the mixture, given the height of the peaks at this locus, manual interpretation of the profile would also have resulted in an exclusion. However, given the apparent presence of the unresolved allele and the evident interpretation issue, re-injection or potentially re-amplification of the sample is warranted to improve the input data for STRmix™ analysis. Should such repeat processing not provide conclusive typing results, setting the software to ignore a problematic locus is appropriate and, in this instance, correctly produced a non-zero total *LR* as expected based on truth data. With respect to these false exclusions occurring with saturated data and unresolved alleles, STRmix™ performed as expected given the profile quality.

Four instances of a mixture deconvolution problem also presented as an exclusion of a known contributor due to a *LR* = 0 at a single locus (TH01), with the remaining loci producing non-zero *LRs* (Table 4, Mixture B). All four instances occurred with comparison of the minor contributor to mixtures constructed from the same two individuals exhibiting the types 7,9 (major contributor) and 9.3,9.3 (minor contributor) (Fig. 7, Mixture B; the electropherogram is provided as Fig. S10). STRmix™ unambiguously assigned the major contributor type at TH01 (weighted 1.000). However, a weight was not assigned to a genotype combination that, based on the electropherogram, one would reasonably consider a possible minor contributor type: 9.3,9.3. Review of STRmix™ results file in each of these instances indicated that the only genotypes considered were 9,9.3 (weighted 0.919), 7,9.3 (weighted 0.080) and Q,9.3 (weighted 0.001), where Q represents an undetected allele. *LR* = 0 was therefore returned for the known minor contributor. This phenomenon is an infrequent result of the statistical sampling process (MCMC) and occurs when the probability space that includes the true genotype is not sampled. To investigate the *LR* = 0 result, the four mixtures were each deconvoluted ten or more times in STRmix™, and most of the time the repeated analysis (which proceeded from a different

Table 4
LR results for two different mixtures, A and B, that incorrectly returned exclusionary results (underlined) at a single locus.

	Mixture A	Mixture A omit TH01	Mixture A replicate amplification	Mixture B	Mixture B replicate STRmix™
	LRs for individual loci				
D8S1179	2.90	2.94	2.11	9.93	9.88
D21S11	25.12	21.76	17.65	79.87	77.69
D7S820	3.41	3.51	3.01	32.88	31.41
CSF1PO	2.91	3.04	3.75	4.71	4.52
D3S1358	1.96	1.99	6.22	9.63	9.63
TH01	0.00	–	11.66	0.00	1.69
D13S317	4.45	4.52	3.93	9.45	9.42
D16S539	2.81	2.48	1.57	2.77	2.87
D2S1338	7.96	8.93	7.55	5.31	4.84
D19S433	1.79	1.67	1.52	10.79	9.93
vWA	2.51	2.41	2.36	2.62	2.65
TPOX	2.16	1.99	1.85	0.83	0.81
D18S51	30.25	27.83	19.92	2.92	3.01
D5S818	2.14	2.16	2.31	20.88	20.59
FGA	6.25	5.41	4.69	11.65	11.17
	LRs for the multi-locus profile				
LR total	0	3.55E+08	2.03E+09	0	3.44E+12
Factor of N! *LR*	0	1.85E+08	9.68E+08	0	1.72E+12
HPD *LR*	0	3.72E+07	3.20E+08	0	8.10E+11

random starting seed for the MCMC) produced a non-zero *LR* for the affected locus. In fact, the *LR* = 0 result could sometimes only be replicated by setting the MCMC starting seed to the value that produced the initial false exclusion. Repeated deconvolutions of the profiles were also performed with an increased number of MCMC accepts (up to 5 million) and with a larger random walk standard deviation (RWSD; up to 0.02), but at least one *LR* = 0 result was observed even with these changes to the STRmix™ parameters (data not shown). The false exclusion is likely due to a larger than expected variability in peak heights at this locus that was atypical for the dataset used in establishment of variance parameters. In casework one should attempt to remedy any issue stemming from amplification or electrophoretic phenomena rather than change the interpretation parameters in STRmix™. However, these four particular instances of erroneous STRmix™ results stemmed from the software, not the typing results. Such inconsistency of the electropherogram and STRmix™ results indicates a need for repeating the STRmix™ analysis.

In the examples presented, a weighting of >0.99 was returned for nearly all but the problematic locus. All other loci returning *LRs* > 1 while a single locus has a *LR* = 0 result is a clear indicator that careful review of the typing results and the genotype weights for all loci is merited. Consideration by a skilled analyst of the DNA typing results and all STRmix™ results data is critical in identifying the source of error in the analysis. In all such instances in our study, both the presence of a potential false exclusion and its cause were readily identified by examination of the profile and the STRmix™ results output, with results appearing inconsistent with scientific expectations based on manual review and comparison of the profiles and, in these instances from validation studies, truth data.

3.4. Alternative propositions

Although reference samples for more than one individual may be provided to a laboratory for comparison to evidentiary profiles, the relevant question in the typical legal context relates to a single individual (e.g., complainant, defendant 1 or defendant 2). Nonetheless, STRmix™ analyses were conducted to assess the effects of testing known contributors concurrently, as well as contributors and non-contributors. When two known contributors were assessed concurrently (i.e., for a three-person mixture, H_1 = contributor 1 + contributor 2 + unknown contributor; H_2 = three unknown contributors), the *LR* was additive, approximating the combined *LRs* of testing the contributors individually. However, when one of the two contributors thus assessed was a non-contributor (i.e., for a three-person mixture, H_1 = contributor + non-contributor + unknown contributor; H_2 = three unknown contributors), different outcomes were observed. When a non-contributor tested individually returned a *LR* of zero, the *LR* for concurrent testing with a known contributor was zero. This result is a correct assessment for both individuals considered together but does not appropriately represent the presence of the known contributor in the mixture. When the non-contributor tested singly was not excluded but returned support for H_2 ($0 < LR < 1$), the results of concurrent testing with a known contributor varied: (a) *LR* of zero, or (b) an additive *LR* > 1. The latter indicates incorrect support for H_1 given that one of the individuals tested concurrently is a non-contributor.

Where appropriate, mixture interpretation can be conditioned on the assumption that the DNA of a given individual (e.g., donor of a vaginal swab) is present in the sample. The use of such information has been shown to increase the *LR* for H_1-true propositions and reduce the *LR* for H_2-true propositions [11]. A total of 94 two, three, four and five-person mixtures that had been analyzed in STRmix™ with no contributor assumed were reinterpreted in STRmix™ to test the effect of conditioning the analysis on a known contributor to the mixture (Fig. S11). For all analyses, a known contributor to the mixture was tested as a person of interest (H_1-true). In general, conditioning the analysis improved the *LR* when the DNA input amounts for the assumed contributor and the person of interest were similar [e.g., the minor contribution to the mixture is at least 50% (red data points) as compared to at most 20% (green data points). As an exception, conditioning the analysis of 1:1 mixtures produced no substantial difference in *LR* when the DNA of one of two contributors was degraded, since the differences in peak heights due to degradation enabled resolution of the two genotypes. In addition, *LRs* increased when both the assumed contributor and person of interest were minor contributors, with the magnitude of the effect decreasing as the number of individuals in the mixture increased. For example, for three-person mixtures with a minor person of interest, the *LR* increased by 4.2 orders of magnitude upon conditioning on another minor contributor, but for four-person mixtures, the benefit was reduced to 2.6 orders of magnitude. By contrast, if the person of interest was a minor contributor, the *LR* rarely improved when the analysis was conditioned on a definitive major contributor, and vice versa (green data points). This result is intuitive: conditioning on a clear major contributor, for example, does not typically improve resolution of the minor component(s).

For the five-person mixtures, the same general trends in the data were most apparent when "trace" contributors (in this instance, defined as individuals with DNA inputs of 0.18 ng or less who also represented less than 10% of the total DNA load for the mixture) were distinguished from non-trace contributors. In these mixtures, *LRs* increased to the greatest degree and most consistently when the assumed contributor and person of interest were both (a) trace contributors, or (b) non-trace contributors. Conditioned analyses of 1:1:1:1:1 mixtures, as well as analyses conditioned on a trace contributor in which a non-trace conditioner was the person of interest (or vice versa), increased the *LR* by less than one order of magnitude on average.

3.5. Incorrect number of contributors

Within STRmix™ the number of contributors to a DNA profile must be assigned prior to analysis. The true number of contributors to an evidence profile, however, is unknown. Uncertainty in determination of the number of contributors may increase due to artifacts, stutter percentages that exceed expectation, allele dropout and, particularly with higher contributor numbers, allele sharing. Given a contributor number of *N* and the assumption of an additional contributor ($N+1$), STRmix™ adds the additional (unseen) contributor at trace levels which, when considered with the true trace contributor, diffuses the genotype weights and can either return a false exclusion or lower the *LR* of a true contributor [27,35]. The *LR* of the major contributor is not appreciably different when $N+1$ contributors are assigned.

In the present study, the effect on the *LR* of assuming an incorrect number of contributors was tested by both increasing ($N+1$) and decreasing (N-1) the number. For the $N+1$ tests, 27 total one, two and three-person profiles were interpreted as originating from two, three and four individuals, respectively. The *LR* was calculated using the Database Search function for both true contributors and 200 non-contributors, which were converted to HPD *LR* estimates as previously described.

For true contributors (H_1-true), the majority of HPD *LRs* under the assumptions of *N* and $N+1$ contributors were similar (within one order of magnitude); for 13% of the analyses, the HPD *LRs* decreased by more than one order of magnitude (Fig. S12). With regard to non-contributors (H_2-true), 89.6% were excluded (HPD *LR* = 0) under assumption of the correct number of contributors,

and the remainder (excepting the false H_1-support instances noted above) returned HPD *LR*s < 1. Under the incorrect assumption of an additional contributor, only 5.3% of non-contributors were excluded outright, though overall 94.3% returned HPD *LR*s < 1.; 4.0% of such analyses were inconclusive (HPD *LR* = 1), and only 1.7% of H_2-true tests analyzed with $N+1$ contributors returned incorrect support for H_1. The vast majority (91.9%) of results with incorrect H_2 were HPD *LR*s ≤ 10; only one result out of 5,716 $N+1$ analyses was >100 (HPD *LR* = 126).

STRmix™ generates a *LR* = 0 if contributor number is underestimated since any "extra" allele cannot be accounted for under the assigned number of contributors. Therefore, as a means of examining the impact of assuming too few contributors without returning an exclusion outright, three mixtures were artificially created from a two-person mixture (1:5 contributor ratio) by adding an additional contributor without adding any new (exclusionary) alleles. The "third" contributor was constructed as if a child of the two true contributors, sharing alleles at all loci, by increasing the rfu values of the alleles by approximately 50 rfu, 100 rfu or 200 rfu per mixture. Each artificially constructed three-person profile was analyzed as a two-person mixture and compared with the true contributors and 200 non-contributors. The resulting *LR*s for the major or minor contributor were not affected by the addition of a third contributor at any of the three average peak heights. All non-contributors resulted in exclusions (*LR* = 0)

3.6. Manual and probabilistic interpretation of the same mixed profiles

To assess general consistency between manual interpretation and STRmix™ analysis, a set of mixtures prepared over a range of contributor ratios and DNA template amounts was analyzed using both methods. Where a person of interest was not excluded as a possible contributor to a mixture based on manual interpretation, STRmix™ analysis demonstrated support for H_1, with HPD *LR*s ranging from 8.7×10^8 to 1.8×10^{19} (Table S3). Where loci or entire profiles were manually disqualified for CPI calculation following application of a stochastic threshold of 200 rfu, STRmix™ results varied: HPD *LR*s for true contributors ranged from 2,200 to 250 billion in support of H_1, and for non-contributors were 0 (exclusion), 850 trillion in support of H_2 and 2 in support of H_1. The latter result occurred in a three-person mixture noted above as providing incorrect support for H_1.

A comparative examination of mixtures from 30 authentic forensic specimens was also performed (Fig. 8). For reference samples included by manual interpretation as potential major

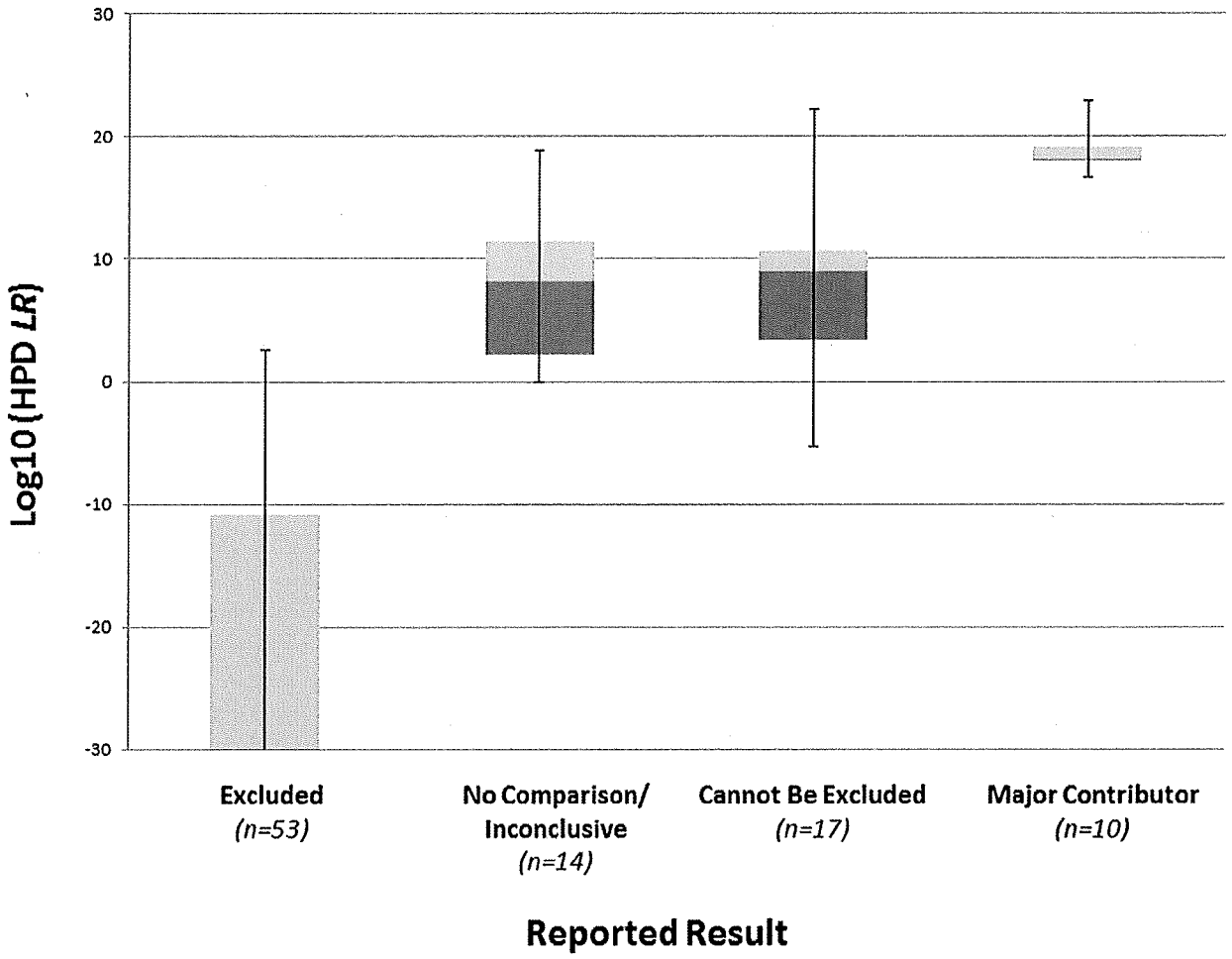


Fig. 8. Manual versus probabilistic interpretation of mixed profiles from evidentiary specimens. Box and whisker representation of the HPD *LR*s calculated for 94 total reference genotype comparisons to Identifiler® Plus profiles developed from 30 authentic forensic specimens from adjudicated FBI cases. The manual interpretation results (x-axis) were categorized based on how the person of interest comparison was reported: excluded as a contributor to the DNA evidence, inconclusive (did not meet standards for match comparison), cannot be excluded (potential contributor to the mixed DNA profile reported with a CPI statistic), or major contributor (deduced single-source profile reported with a RMP statistic). 99.0% lower-bound HPD *LR*s (y-axis) are plotted using a log scale, with HPD *LR* = 0 plotted as −30. All 30 evidentiary profiles indicated mixtures of DNA from two or more individuals.

contributors and previously reported with a RMP statistic, the STRmix™ results produced HPD *LR*s in excess of 1×10^{16}. Considering 53 manual exclusions: in 51 instances the probabilistic interpretation also supported exclusion, in one instance the HPD *LR* was 1 (denoting an inconclusive result), and in the last instance the HPD *LR* was 410 in support of inclusion. In this last instance, the probabilistic interpretation was conducted assuming four contributors, and STRmix™ indicated the reference as one of three trace contributors, comprising just 5% of the DNA load. Examination of the evidence electropherogram and reference genotype (Fig. S13) revealed a peak disqualified as stutter in the manual interpretation that was modeled by STRmix™ as an allele for a trace contributor that exhibited dropout at multiple loci.

Among the 94 total hypotheses tested for the evidentiary specimens, reference samples were manually designated as potential contributors ("cannot be excluded") in 17 instances (involving ten distinct mixtures; Table 5). These comparisons were previously reported with CPI statistics, which ranged from a low of 1 in 1 to a high of 1 in 30,000. Where the statistical estimate was 1 in 1, the percentage of the population that would be included as possible contributors (75%–98%) was also reported (Table 5). For STRmix™ analyses of these ten mixtures, nine were analyzed as originating from four persons and one was analyzed as originating from three persons. For all potential contributors previously reported with a CPI of 1 in 2 or greater, the STRmix™ results supported inclusion, with HPD *LR*s ranging from 2,600 to 16 sextillion. Considering the six reference comparisons for which the reported CPI was 1 in 1, the probabilistic interpretations produced HPD *LR*s less than 1 (denoting support for H_2) in four instances. This is not a surprising result given the complexity of the mixtures and the weak statistical support calculated for the manual inclusions. For two of these four instances, both involving the same mixture (C1Q10, compared to reference samples C1K10 and C1K18), the total and HPD *LR* values from the STRmix™ output indicated a far larger HPD interval than is typical (Table 5). Given the approximate 4:4:1:1 contributor ratio for this mixture, it seems likely that the similarities in DNA loads may have resulted in MCMC uncertainty, creating the wide HPD interval, which in turn accounts for the strength of the support for H_2. The remaining two CPI results of 1 in 1 provided support for H_1.

For the evidentiary mixtures that were deemed inconclusive by manual interpretation, STRmix™ produced wide-ranging HPD *LR*s, as with the prepared mixtures, from 1 to greater than 1×10^{18} in support of H_1 (Fig. 8). These results indicate that a fully continuous probabilistic method enabling usage of more profile information and modeling features in STRmix™ yields more refined conclusions for some mixed DNA profiles as compared to a binary interpretation method.

4. Conclusions

The internal validation studies described herein involved the examination of more than 300 autosomal STR profiles, derived from one to five contributors and representing a wide range of contributor ratios and DNA template amounts. The probabilistic interpretations using laboratory-specific parameters totaled more than 800 known contributor propositions, nearly 60,000 non-contributor tests, and nearly 100 reference sample comparisons to mixed profiles developed from authentic forensic specimens. Overall, the study results demonstrate that STRmix™ software

Table 5
Manual interpretation statistics compared to STRmix™ results for casework comparisons reported as "cannot be excluded".

		Manual Interpretation		STRmix Interpretation		
Forensic Sample Identifier	Reference Sample Identifier	Reported CPI	Percentage of Population Included	Assigned Number of Contributors	HPD *LR* or 1/HPD *LR**	*Orders of magnitude difference between total LR and HPD LR*
C1Q10	C1K10	1 in 1	75%	4	210,000*	*5.1*
	C1K14	1 in 1	75%	4	85 million	*2.8*
	C1K18	1 in 1	75%	4	2,000*	*8.2*
C1Q14	C1K6	1 in 4	n/a	4	8.1 billion	*0.7*
C1Q15	C1K6	1 in 18	n/a	4	27 million	*0.6*
	C1K10	1 in 18	n/a	4	51 billion	*0.8*
	C1K14	1 in 18	n/a	4	4.4 billion	*0.5*
C1Q19	C1K14	1 in 33	n/a	3	1.1 billion	*0.4*
C1Q22	C1K14	1 in 1	98%	4	2*	*0.7*
C1Q23	C1K14	1 in 1	98%	4	370 trillion	*0.8*
	C1K18	1 in 1	98%	4	2*	*0.6*
C1Q26	C1K6	1 in 9	n/a	4	140 million	*0.7*
	C1K10	1 in 9	n/a	4	52 billion	*0.6*
C1Q39	C1K14	1 in 2	46%	4	1.6 trillion	*1.0*
C1Q40	C1K14	1 in 2	64%	4	16 sextillion	*0.8*
C3Q2	C3K10	1 in 30,000	n/a	4	2,600	*0.3*
	C3K11	1 in 30,000	n/a	4	230 trillion	*0.4*

Combined probabilities of inclusion (CPIs), population percentages and *LR*s are based on U.S. Caucasian or Navajo population sample allele frequencies and theta values of 0.01 or 0.03 (respectively), as appropriate based on the case details. Population percentages were included in the FBI Laboratory Report of Examination for all reported populations if the CPI statistic was 1 in 1 for any of the reported populations.

* Asterisks denote conversion of HPD *LR*s that were less than 1 to a positive integer (1/HPD *LR*) to convey the degree of support for the H_2 hypothesis on the same scale as HPD *LR*s >1 results. All HPD *LR* and 1/HPD *LR* values were truncated to two significant figures.

n/a = not applicable (percentage of population included is only provided for "1 in 1" CPIs)

performed as expected. With very few exceptions, genotype weights were intuitively correct, and the statistical results were consistent with scientific expectations. Across multiple studies, the data showed that as the informative content of a profile increased such as with higher DNA template amounts, greater disparity in contributor ratios, and simultaneous consideration of PCR replicates, *LR*s increased for true contributors and decreased for known non-contributors.

When a 99.0% one-sided lower-bound HPD *LR* value was used to assess the STRmix™ results for the two, three, four and five-person prepared mixtures, the software proved to be appropriately sensitive and specific. Across more than 60,000 tests, 93.4% of true contributors produced HPD *LR*s supporting inclusion, and greater than 99.9% of non-contributors resulted in HPD *LR*s supporting exclusion. Specificity with five-person mixtures was further increased by conditioning the interpretation on a known contributor. In all cases where non-contributor comparisons generated HPD *LR*s > 1, the results were consistent with scientific expectations given the mixture quality and complexity (e.g., degradation, allelic dropout) and the number of contributors. A few exclusionary results (*LR* = 0) for known contributors occurred due to poor profile quality (e.g., inadequate capillary electrophoresis resolution) and when MCMC sampling failed to identify the true genotype combination for a single locus. In all instances, the cause of the unexpected results could be deduced upon review of the STRmix™ output files in relation to the mixture electropherogram and resolved by a repeat STRmix™ analysis with or without modifications (according to error type). Taken together, the results of the various studies (a) aptly demonstrate the reliability of the STRmix™ software in terms of sensitivity and specificity when laboratory-specific parameters are employed for analyses and (b) underscore the importance of analyst review of both the DNA typing and probabilistic genotyping results.

These studies establish that STRmix™ version 2.3.06 is fit for purpose for the interpretation and statistical assessment of single source profiles and mixtures originating from two, three, four and five individuals. To convey the statistical weight and aid comprehension of the reported statistical results, which in this study ranged from *LR*s of 0 to approximately 10^{27}, the *LR* may be accompanied by a qualitative description of the degree of support for the H_1 or H_2 hypothesis [10]. The FBI Laboratory reports the HPD *LR* for Identifiler® Plus typing results with a verbal expression of evidential strength as recommended by the European Network of Forensic Science Institutes (ENFSI) [36] and founded by the Association of Forensic Science Providers [37]; HPD *LR*s of 0 are reported as exclusions, and HPD *LR*s of 1 are reported as uninformative.

The implementation of a fully continuous probabilistic genotyping system on December 1, 2015 represents a major step forward in the interpretation of autosomal STR data at the FBI Laboratory. As evidenced by the comparative examinations of prepared mixtures and evidentiary profiles from prior FBI cases, the conclusions derived from the results of probabilistic genotyping can be expected to align with properly applied historical methods. The probabilistic approach used by STRmix™ greatly increases the information that can be used to deconvolute mixtures and estimate evidentiary weight, showing distinct advantages with mixtures with three or more individuals and low-level contributors. Our analysis of findings supports that STRmix™ reliably applies suitable biological modeling and statistical methods, is sufficiently robust for usage with forensic-type specimens and, as a probabilistic genotyping system, represents a vital advancement in the field of human identification testing.

Acknowledgements

The authors would like to thank the many individuals who provided scientific or technical support for this work, including: Jerrilyn Conway, Jade Gray, Jeremy Fletcher and Baxter Cohen of the DNA Casework Unit, FBI Laboratory; Jill Smerick and Jodi Irwin of the DNA Support Unit, FBI Laboratory; Laura Russell, Catherine McGovern and Stuart Cooper of the Institute of Environmental Science and Research; Jeffrey Monaghan of Robotech Science, Inc.; and Luigi Armogida of NicheVision. This work was supported in part by Award No. 2014-DN-BX-K028, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. This work was also supported in part through the FBI's Visiting Scientist Program, an educational opportunity administered by the Oak Ridge Institute for Science and Education (ORISE). The opinions and assertions presented herein are those of the authors and should not be construed as official or as reflecting the views of the U.S. Department of Justice, the U.S. Department of Commerce, the U.S. Department of Energy or the U.S. Government. Certain commercial equipment, instruments, materials, suppliers and software are identified to specify experimental procedures and foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, the Federal Bureau of Investigation or any branch of the U.S. Government, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2017.04.004.

References

[1] T. Moretti, A.L. Baumstark, B.S. Defenbaugh, K.M. Keys, J.B. Smerick, B. Budowle, Validation of short tandem repeats (STRs) for forensic usage: performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples, Journal of Forensic Sciences 46 (647) (2001) 60.

[2] B. Budowle, A.J. Onorato, T.F. Callagham, A.D. Manna, A.M. Gross, R.A. Guerreri, et al., Mixture Interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework, Journal of Forensic Sciences 54 (2009) 810–821.

[3] P. Gill, J. Buckleton, Commentary on: Budowle B, Onorato AJ, Callaghan TF, della Manna A, Gross AM, Guerrieri RA, Luttman JC, McClure DL. Mixture interpretation: Defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. J Forensic Sci 2009;54 (4):810-21, Journal of Forensic Sciences 55 (2010) 265–268.

[4] 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 Genetics 17 (2016) 125.

[5] **Scientific Working Group on DNA Analysis Methods (SWGDAM), SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories, (2010) . [Accessed] 23 November 2014** http://www.fbi.gov/hq/lab/html/codis_swgdam.pdf.

[6] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Science International: Genetics. 7 (2013) 516–528.

[7] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, et al., Validating TrueAllele® DNA Mixture Interpretation, Journal of Forensic Sciences. 56 (2011) 1430–1447.

[8] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Science International: Genetics. 4 (2009) 1–10.

[9] P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, Forensic Science International: Genetics. 7 (2013) 251–263.

[10] **Scientific Working Group on DNA Analysis Methods (SWGDAM), Guidelines for the validation of probabilistic genotyping systems, (2015) . [Accessed] 3 October 2016.** http://media.wix.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515. pdf.

[11] D. Taylor, Using continuous DNA interpretation methods to revisit likelihood ratio behaviour, Forensic Science International: Genetics 11 (2014) 144–153.
[12] **FBI Laboratory, National DNA Index System (NDIS) Operational Procedures Manual, Version Effective January 1, 2015, (2015) . [Accessed] April 4, 2016** https://www.fbi.gov/about-us/lab/biometric-analysis/codis/ndis-procedures-manual/view.
[13] **FBI Quality Assurance Standards for Forensic DNA Testing Laboratories, (2011) . [Accessed] 19 November 2014** http://www.fbi.gov/about-us/lab/biometric-analysis/codis/qas-standards-for-forensic-dna-testing-laboratories-effective-9-1-2011.
[14] **Scientific Working Group on DNA Analysis Methods (SWGDAM), Guidelines for validation of DNA analysis methods, (2016) . [Accessed] 1 February 2017** https://media.wix.com/ugd/4344b0_813b241e8944497e99b9c45b163b76bd.pdf.
[15] J.-A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Science International: Genetics 7 (2013) 296–304.
[16] C. Brookes, J.-A. Bright, S. Harbison, J. Buckleton, Characterising stutter in forensic STR multiplexes, Forensic Science International: Genetics 6 (2012) 58–63.
[17] D. Taylor, J. Buckleton, Bright J-A: Factors affecting peak height variability for short tandem repeat data, Forensic Science International: Genetics 21 (2016) 126–133.
[18] D. Taylor, J.-A. Bright, C. McGoven, C. Hefford, T. Kalafut, J. Buckleton, Validating multiplexes for use in conjunction with modern interpretation strategies, Forensic Science International: Genetics 20 (2016) 6–19.
[19] J.-A. Bright, E. Huizing, L. Melia, J. Buckleton, Determination of the variables affecting mixed MiniFiler DNA profiles, Forensic Science International: Genetics 5 (2011) 381–385.
[20] J.-A. Bright, J. Turkington, J. Buckleton, Examination of the variability in mixed DNA profile parameters for the Identifiler multiplex, Forensic Science International: Genetics 4 (2009) 111–114.
[21] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Science International 64 (1994) 125–140.
[22] National Research Council II, National Research Council Committee on DNA Forensic Science, The Evaluation of Forensic DNA Evidence, National Academy Press, Washington, D.C, 1996.
[23] B. Budowle, T.R. Moretti, A.L. Baumstark, D.A. Defenbaugh, K.M. Keys, Population data on the thirteen CODIS core short tandem repeat loci in African Americans, US, Caucasians, Hispanics, Bahamanians, Jamaicans and Trinidadians, Journal of Forensic Sciences 44 (1999) 1277–1286.
[24] B. Budowle, P. Collins, P. Dimsoski, C. Ganong, L. Hennessy, C. Leibelt, et al., Population data on the STR loci D2S1338 and D19S433, Forensic Sci Communications (2001) 3.
[25] B. Budowle, B. Shea, S.J. Niezgoda, R. Chakraborty, CODIS STR. loci data from 41 sample populations, Journal of Forensic Sciences 46 (453) (2001) 89.
[26] T.R. Moretti, B. Budowle, J.S. Buckleton, Notice of Amendment of the FBI's STR Population Data Published in 1999 and 2001, Journal of Forensic Sciences 60 (2015) 1114–1116.
[27] J.-A. Bright, D. Taylor, J. Curran, J. Buckleton, Searching mixed DNA profiles directly against profile databases, Forensic Science International: Genetics 9 (2014) 102–110.
[28] D. Taylor, J.-A. Bright, J. Buckleton, The 'factor of two' issue in mixed DNA profiles, Journal of Theoretical Biology 363 (2014) 300–306.
[29] J.-A. Bright, I.W. Evett, D. Taylor, J.M. Curran, J. Buckleton, A series of recommended tests when validating probabilistic DNA profile interpretation software, Forensic Science International: Genetics 14 (2015) 125–131.
[30] J.-A. Bright, D. Taylor, C.E. McGovern, S. Cooper, L. Russell, D. Abarno, et al., Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles, Forensic Science International: Genetics 23 (2016) 226–239.
[31] J.M. Curran, J.S. Buckleton, C.M. Triggs, What is the magnitude of the subpopulation effect, Forensic Science International 135 (2003) 1–8.
[32] J. Buckleton, J. Curran, J. Goudet, D. Taylor, A. Thiery, B.S. Weir, Population-specific F_{ST} values for forensic STR markers: A worldwide survey, Forensic Science International: Genetics 23 (2017) 91–100.
[33] C.M. Triggs, J.M. Curran, The sensitivity of the Bayesian HPD method to the choice of prior, Science & Justice 46 (2006) 169–178.
[34] D. Taylor, J.-A. Bright, J. Buckleton, J. Curran, An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations, Forensic Science International: Genetics 11 (2014) 56–63.
[35] J.-A. Bright, J.M. Curran, J.S. Buckleton, The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation, Forensic Science International: Genetics 12 (2014) 208–214.
[36] **European Network of Forensic Science Institutes, ENFSI Guideline for Evaluative Reporting in Forensic Science, (2015) . [Accessed] 3 March 2017** http://enfsi.eu/news/enfsi-guideline-evaluative-reporting-forensic-science/.
[37] Association of Forensic Science Providers. Standards for the formulation of evaluative forensic science expert opinion, Science & Justice 49 (2009) 161–164.

Forensic Science International: Genetics 34 (2018) 11–24

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Internal validation of STRmix™ – A multi laboratory response to PCAST

Jo-Anne Bright[a,*], Rebecca Richards[a], Maarten Kruijver[a], Hannah Kelly[a], Catherine McGovern[a], Alan Magee[b], Andrew McWhorter[c], Anne Ciecko[d], Brian Peck[e], Chase Baumgartner[f], Christina Buettner[g], Scott McWilliams[g], Claire McKenna[h], Colin Gallacher[i], Ben Mallinder[i], Darren Wright[j], Deven Johnson[k], Dorothy Catella[l], Eugene Lien[m], Craig O'Connor[m], George Duncan[n], Jason Bundy[o], Jillian Echard[p], John Lowe[q], Joshua Stewart[r], Kathleen Corrado[s], Sheila Gentile[s], Marla Kaplan[t], Michelle Hassler[u], Naomi McDonald[v], Paul Hulme[w], Rachel H. Oefelein[x], Shawn Montpetit[y], Melissa Strong[y], Sarah Noël[z], Simon Malsom[A], Steven Myers[B], Susan Welti[C], Tamyra Moretti[D], Teresa McMahon[E], Thomas Grill[F], Tim Kalafut[G], MaryMargaret Greer-Ritzheimer[H], Vickie Beamer[I], Duncan A. Taylor[J,K], John S. Buckleton[a,L]

[a] Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland, 1142, New Zealand
[b] Forensic Science Ireland, Ireland
[c] Texas Department of Public Safety, Houston Laboratory, United States
[d] Midwest Regional Forensic Laboratory, Andover, MN, United States
[e] Centre of Forensic Sciences, Toronto, Canada
[f] Texas Department of Public Safety, Austin Laboratory, United States
[g] Wyoming State Crime Laboratory, United States
[h] Austin Police Department, City of Austin, TX, United States
[i] Scottish Police Authority (SPA), United Kingdom
[j] Idaho State Police Forensic Services, United States
[k] Sacramento District Attorney's Office Laboratory of Forensic Services, CA, United States
[l] Oakland County Sheriff's Office, MI, United States
[m] New York City Office of Chief Medical Examiner (OCME), United States
[n] Broward Sheriff's Office Crime Laboratory, FL, United States
[o] Florida Department of Law Enforcement, United States
[p] Connecticut DESPP Division of Scientific Services, United States
[q] Key Forensic Services Ltd., Warrington Laboratory, United Kingdom
[r] Texas Department of Public Safety, Corpus Christi Laboratory, United States
[s] Onondaga County Center for Forensic Sciences, NY, United States
[t] Oregon State Police Laboratory (OSP), United States
[u] San Diego County Sheriff's Regional Crime laboratory, United States
[v] Texas Department of Public Safety, Lubbock Laboratory, United States
[w] Cellmark Forensic Services, United Kingdom
[x] DNA Labs International, United States
[y] San Diego Police Department Crime Laboratory, CA, United States
[z] Laboratoire de sciences judiciaires et de médecine légale (LSJML) Montréal, Canada
[A] Key Forensic Services Ltd., Norwich Laboratory, United Kingdom
[B] California Department of Justice Bureau of Forensic Services, United States
[C] Department of Forensic Sciences Laboratory, Washington DC (DFS), United States
[D] Federal Bureau of Investigation (FBI), United States
[E] Forensic Science Northern Ireland, Northern Ireland
[F] Erie County Central Services Laboratory, Buffalo, NY, United States
[G] US Army Criminal Investigation Laboratory (USACIL), United States
[H] DuPage County Sheriff's Crime Laboratory, IL, United States
[I] Scottsdale Police Department Crime Laboratory, AZ, United States
[J] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[K] School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia
[L] University of Washington, Department of Biostatistics, Seattle, WA 98195, United States

* Corresponding author.
E-mail address: jo.bright@esr.cri.nz (J.-A. Bright).

https://doi.org/10.1016/j.fsigen.2018.01.003
Received 20 August 2017; Received in revised form 29 November 2017; Accepted 6 January 2018
Available online 08 January 2018
1872-4973/

ARTICLE INFO

Keywords:
PCAST
STRmix
Forensic DNA
Probabilistic genotyping
Continuous models
Validation

ABSTRACT

We report a large compilation of the internal validations of the probabilistic genotyping software STRmix™. Thirty one laboratories contributed data resulting in 2825 mixtures comprising three to six donors and a wide range of multiplex, equipment, mixture proportions and templates. Previously reported trends in the LR were confirmed including less discriminatory LRs occurring both for donors and non-donors at low template (for the donor in question) and at high contributor number. We were unable to isolate an effect of allelic sharing. Any apparent effect appears to be largely confounded with increased contributor number.

1. Introduction

In 2016, the President's Council of Advisors on Science and Technology (PCAST) issued a report [1] and subsequently an addendum [2]. This report discussed a number of forensic disciplines. Included amongst these was the interpretation of complex DNA mixtures. PCAST defined a complex mixture as any profile with three or more donors. The report noted perceived limits to the proof of validity of the use of probabilistic genotyping (PG) in some situations as of September 2016. In particular they highlighted gaps regarding high ratio and high contributor number mixtures. PCAST considered validity proven for mixtures containing "*three contributors where the person of interest comprises at least 20% of the sample.*" [2]. They noted that the "*few studies that have explored 4- or 5-person mixtures often involve mixtures that are derived from only a few sets of people (in some cases, only one).*" [2]. They call for the expansion of empirical studies, testing the validity and reliability of PG methods across a broader relevant range of profile types.

PCAST limited themselves for proof of validity to empirical studies published in the peer reviewed literature. There are a number of published reports describing the validation of various probabilistic genotyping software by the developers. These include the New York City Office of Chief Medical Examiner's FST Tool [3], TrueAllele® [4], and STRmix™ [5]. More recently the validation of GenoProof Mixture 3 [6] and *Kongoh* [7] has been reported.

PCAST also perceived there was a gap in "*the need for clarity about the scientific standards for the validity and reliability of forensic methods.*" [1]. The Scientific Working Group on DNA Analysis Methods (SWGDAM) [8] and International Society for Forensic Genetics (ISFG) [9] have both published comprehensive guidelines that inform how to test a probabilistic genotyping system to ensure reliability and validity of results.

At the time of the PCAST report there was a considerable number of empirical studies already undertaken by various laboratories who had implemented, or were in the process of implementing, STRmix™. These followed the SWGDAM guidelines [10,11]. They were not published in the peer reviewed literature largely because it is the policy of many journals not to publish such material. Some of these studies are already in the public domain on websites (see for example [12,13]).

Since the appearance of the PCAST report, the Federal Bureau of Investigation Laboratory, Quantico, has published its STRmix™ internal validation in the peer reviewed literature [14], also in accordance with the SWGDAM guidelines. This publication reports 277 mixtures with two to five donors and a range of mixture ratios and templates.

In this work we report a further study of 2825 mixtures compiled from 31 laboratories (including multi laboratory systems) who are using STRmix™ in casework (28/31) or currently validating STRmix™ for future use in casework (3/31). Mixtures of three, four, five, and six contributors were specifically targeted in order to address the criticisms of PCAST.

We aim to specifically address the deficiencies described by PCAST in their report by addressing the following points:

(1) How well does the method perform as a function of the number of contributors to the mixture? How well does it perform when the number of contributors to the mixture is *unknown*?

(2) How does the method perform as a function of the number of alleles shared among individuals in the mixture? Relatedly, how does it perform when the mixtures include related individuals?

(3) How well does the method perform – and how does accuracy degrade – as a function of the absolute and relative amounts of DNA from the various contributors?

We address point 1 in experiment 1 by analysing all submitted mixtures assuming the *apparent* number of contributors. The apparent number of contributors (N) was determined blind by the submitting laboratory following their own standard operating procedures. Note that this resulted in all six person mixtures being analysed as assuming less than six. Additionally, we have assumed N + 1 for a subset of the data within experiment 2. Point 2 we address by interrogating the data in experiment 1 with respect to the amount of allelic sharing. Point 3 we address by conducting H_p and H_d true tests on mixtures in experiment 1.

In this work the developers of STRmix™ did not generate or choose the data that was analysed by individual (non-developing) laboratories and they have not censored any data from the results. This adheres to the call by PCAST for work to be carried out in conjunction between developers and non-developing organisations.

There is a fourth point to the list in the PCAST report:

(4) Under what circumstances – and why – does the method produce results (random inclusion probabilities) that differ substantially from those produced by other methods?

We do not address point (4) within this paper, however work is ongoing to address it across a number of continuous and semi-continuous platforms.

2. Methods

2.1. Data submission

Participating laboratories submitted ground truth known profiles originating from three to six contributors that had previously been interpreted as part of their STRmix™ internal validation studies. Profiles were submitted as analysed data in the form of text or Excel files. In addition, laboratories provided reference profiles for the known contributors, their validated laboratory specific settings, and the apparent number of contributors to each profile. The apparent number of contributors was determined by the submitting laboratories following their own standard operating procedures. The apparent number of contributors was used as the true number of contributors to a crime profile is never known.

2.1.1. Data description

Apparent three, four and five person mixtures were interpreted by staff at ESR (New Zealand) using STRmix™ V2.5.02. No apparent single source or two person mixtures were interpreted as PCAST, perhaps erroneously, decreed foundational validity to be already established for these [1]. In total there were 2825 mixtures interpreted from 31 different laboratories generated using eight different STR multiplexes and analysed on two different types of capillary electrophoresis (CE) instruments.

The STRmix™ settings used for the interpretation were those determined by the contributing laboratory. These included per allele stutter ratios (back and forward, where determined), allele and stutter peak height variance distributions, analytical thresholds, saturation,

and drop-in parameters. For each interpretation, eight MCMC chains of 100,000 burn-in accepts and 50,000 post burn-in accepts were used.

The number of profiles submitted, multiplex, PCR cycle number, CE instrument used, and number of mixtures interpreted for each participating laboratory are provided in Table 1. Note some laboratories submitted profiles generated using more than one multiplex (kit) and some were multi laboratory systems, submitting profiles from different laboratories within the one system. Many of the laboratories undertook dilution series to prepare mixtures for interpretation. These were typically made by taking DNA from a few donors, often staff members, and mixing them in different combinations and ratios. PCAST noted that "In human molecular genetics, an experimental validation of an important diagnostic would typically involve hundreds of distinct samples." (PCAST pg 81). Each different combination of genotypes is a unique contributor combination.

The number of the unique contributor combinations for each mixture type is given in Table 1. For example, there were twelve combinations of different contributors for the apparent three person mixtures submitted by Lab 01. In total there were 25 apparent three person mixtures from Lab 01, hence 12/25 in Table 1. For all laboratories, there were 205 unique three contributor profiles, 132 unique four contributor profiles, and 14 unique five contributor profiles. Within the STRmix™ deconvolution, template is modelled per contributor [11]. The mode of the post burn-in proposals for template per contributor was used to calculate mixture proportion. The mixture proportions as determined by STRmix™ (sorted by ascending proportion for contributor 1, constrained as the 'major' contributor) are plotted for each apparent N in Fig. 1. At least one contributor in 69.5% of the apparent three person mixtures, 96.5% of the apparent four person mixtures and all of the apparent five person mixtures contained less than 20% of the sample.

PCAST calls for an investigation to be conducted into how a method *"performs as a function of the number of alleles shared among individuals in the mixture"*. In Fig. 2 we provide the distribution of allele sharing for known contributors in the mixtures, broken down by the true number of contributors to a mixture. Allele sharing (AS) is defined as the fraction of alleles in the donors collectively that appear in two or more donor genotypes. The upper tail (> 0.80 proportion AS) for the three and four contributor mixtures are a known family group consisting of a mother, father, and their two biological children that was investigated by one participating laboratory.

2.2. Experiment 1

For each profile, likelihood ratios (*LRs*) were calculated for the true donors and 10,000 false donors. The profiles of the 10,000 non-donors were created by simulation using the FBI Caucasian allele frequencies for each multiplex. All *LRs* were calculated using the Caucasian allele

Table 1
A list of the contributing laboratories, multiplex (kit) used, PCR cycle number, and CE instrument. The total number of mixtures interpreted per laboratory are sorted by apparent number of contributors with the number of unique contributor combinations and minimum minor proportion as determined by STRmix™ indicated.

					Number of each mixture type Unique contributor combinations/total (Minimum minor contribution)		
Lab	Samples submitted (true N)	Kit	Cycle Number	CE	Apparent 3p	Apparent 4p	Apparent 5p
L01	$N_3 = 24, N_4 = 23$	Fusion 5C	28	3130	12/25(7%)	12/22(7%)	–
L02	$N_3 = 19, N_4 = 24$	Identifiler™ Plus	28	3500	4/21(6%)	3/22(6%)	–
L03	$N_3 = 88, N_4 = 128, N_5 = 48$	GlobalFiler™	29	3500	5/87(3%)	6/161(< 1%)	2/16(5%)
L04	$N_3 = 3, N_4 = 3$	NGM SElect™	30	3130	1/3(10%)	1/3(6%)	–
L05	$N_3 = 39, N_4 = 37$	Fusion 6C	29	3130	5/50(3%)	4/26(< 1%)	–
L06	$N_3 = 28, N_4 = 69$	Identifiler™ Plus	28	3130	4/67(28%)	2/30(12%)	–
L07	$N_3 = 29, N_4 = 30$	Identifiler™ Plus	28	3130	4/36(2%)	1/23(2%)	–
L08	$N_3 = 19, N_4 = 20$	Fusion 6C	29	3500	2/24(7%)	1/15(4%)	–
L09	$N_3 = 28, N_4 = 8, N_5 = 6$	Fusion 5C	30	3500	4/28 (1%)	2/8(2%)	1/6(6%)
	$N_3 = 22, N_4 = 22$	Identifiler™ Plus	29	3500	1/22 (1%)	1/22 (2%)	–
L10	$N_3 = 29, N_4 = 52, N_5 = 12$	GlobalFiler™	28	3500	4/64 (3%)	4/29 (1%)	–
L11	$N_3 = 69, N_4 = 42$	GlobalFiler™	28	3500	2/69 (< 1%)	2/42 (1%)	–
L12	$N_3 = 28, N_4 = 32$	NGM SElect™	29	3500	2/38 (5%)	1/22 (5%)	–
L13	$N_3 = 3, N_4 = 3$	NGM SElect™	30	3130	1/3 (9%)	1/3 (3%)	–
	$N_3 = 3, N_4 = 3$	PowerPlex® ESI17 Pro	30	3130	1/3 (13%)	1/3 (6%)	–
L14	$N_3 = 10, N_4 = 13$	PowerPlex® 16 HS	30	3130	2/16 (7%)	1/7 (5%)	–
L15	$N_3 = 26$	PowerPlex® ESI17 Fast	30	3130	11/26 (2%)	–	–
	$N_3 = 28$	PowerPlex® ESI17 Fast	30	3500	11/28 (2%)	–	–
L16	$N_3 = 29, N_4 = 11$	Identifiler™ Plus	28	3130	9/38 (4%)	1/2 (5%)	–
L17	$N_3 = 26, N_4 = 32$	GlobalFiler™	29	3500	2/32 (4%)	1/26 (1%)	–
L18	$N_3 = 97, N_4 = 46$	Fusion 5C	29	3130	7/108 (7%)	3/35 (2%)	–
L19	$N_3 = 28, N_4 = 30$	Identifiler™ Plus	29	3130	9/37 (3%)	15/21 (2%)	–
L20	$N_3 = 22, N_4 = 23, N_5 = 12$	GlobalFiler™	29	3500	9/42 (< 1%)	4/13 (5%)	1/2 (1%)
L21	$N_3 = 43, N_4 = 39$	Fusion 6C	29	3500	14/59 (4%)	9/23 (1%)	–
L22	$N_3 = 62, N_4 = 65, N_5 = 11$	GlobalFiler™	29	3500	27/69 (3%)	25/64 (1%)	2/5 (7%)
L23	$N_3 = 72, N_4 = 64$	Fusion 6C	29	3500	6/83 (1%)	4/53 (< 1%)	–
	$N_3 = 159, N_4 = 60$	Identifiler™ Plus	28	3130	4/161 (1%)	3/58 (< 1%)	–
L24	$N_3 = 35, N_4 = 36$	GlobalFiler™	29	3500	4/37 (3%)	3/34 (2%)	–
L25	$N_3 = 20, N_4 = 24$	GlobalFiler™	29	3500	1/20 (5%)	1/24 (6%)	–
L26	$N_3 = 18, N_4 = 12$	Identifiler™ Plus	28	3130	17/25 (6%)	3/5 (< 1%)	–
L27	$N_3 = 51, N_4 = 42$	Identifiler™ Plus	28	3500	5/71 (3%)	2/22 (< 1%)	–
L28	$N_3 = 12, N_4 = 77, N_5 = 76, N_6 = 65$	Fusion 5C	29	3500	6/24 (3%)	7/151 (< 1%)	6/55 (< 1%)
L29	$N_3 = 52, N_4 = 52$	GlobalFiler™	28	3500	2/53 (3%)	1/51 (1%)	–
L30	$N_3 = 31, N_4 = 42$	GlobalFiler™	29	3500	4/42 (4%)	3/31 (< 1%)	
L31	$N_3 = 63, N_4 = 99, N_5 = 17$	GlobalFiler™	29	3500	3/80 (1%)	4/85 (< 1%)	2/14 (< 1%)
		TOTAL Number of each mixture type unique combinations/total (minimum minor contribution)			205/1591 (< 1%)	132/1136 (< 1%)	14/98 (< 1%)

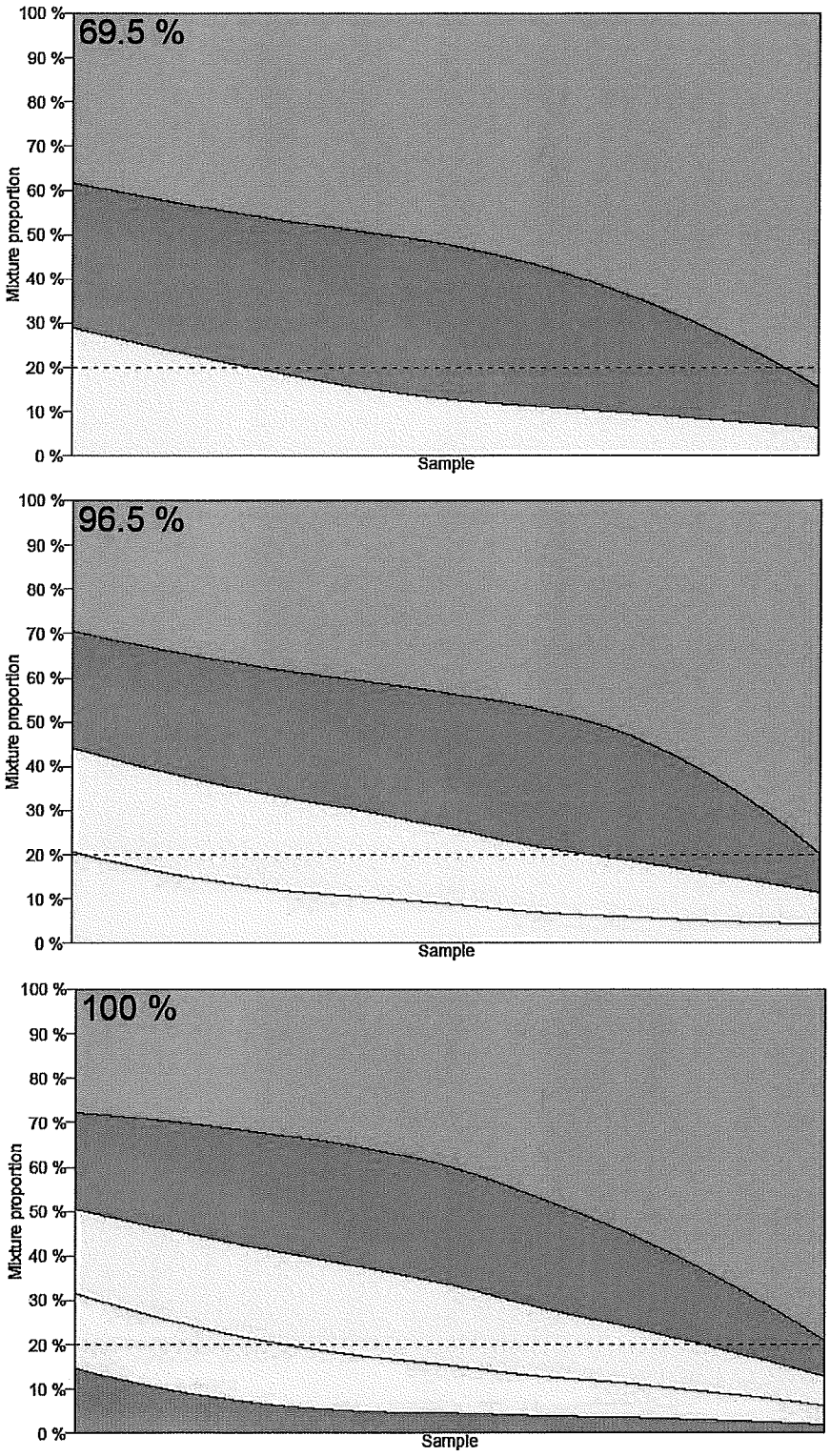


Fig. 1. Mixture proportions as calculated by STRmix™ and sorted by ascending proportion plotted by apparent N where 1a is apparent three, 1b apparent four and 1c apparent five N. Plots are smoothed for improved readability.

frequencies from the FBI expanded CODIS core set [15] and a theta (F_{ST}) of 0.01. The propositions considered were:

H_p: the DNA originated from the person of interest (either true or false donor) and N-1 unknown contributors

H_d: the DNA originated from N unknown contributors

where N was the apparent number of contributors.

Average peak height (*APH*) was calculated for each contributor by averaging the peak heights of the unmasked alleles (not shared between contributors and not in back stutter positions of any other contributor alleles). Alleles that had dropped out were assigned a height of half the laboratory's analytical threshold (AT).

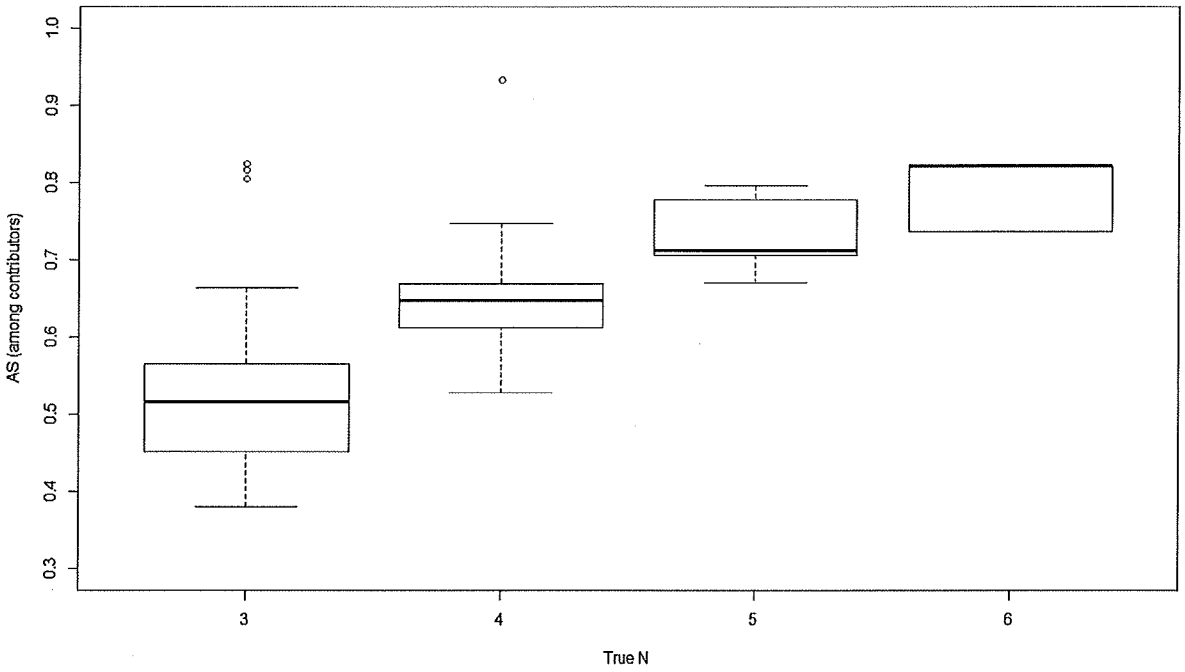


Fig. 2. Distribution of allele sharing (AS) for known contributors to mixtures, plotted by true N.

2.3. Experiment 2

For one laboratory the three and four contributor profiles were analysed at both the apparent number of contributors (N) and one greater (N + 1). For these mixtures, apparent N was the same as known N. In practise, when analysed as N + 1 a non-existent contributor with true mixture proportion 0 has been added to reflect this ambiguous contributor being present at trace amounts. The mixture proportion for this additional contributor was constrained to be low, but not necessarily zero, using the informed mixture proportion prior function in STRmix™ [16]. The *LR*s for the true donors and 10,000 non-donors were assigned as per Experiment 1.

3. Results

3.1. Data review

The summary statistics for each interpretation were reviewed prior to review of the *LR*. These statistics included the Gelman-Rubin convergence statistic, average $\log_{10}$(likelihood) of the post burn-in MCMC, the average of the post burn-in allele variance parameter, and the average of the post burn-in stutter variance parameter. These values can be used as diagnostics of the interpretation, to check for adequate MCMC convergence. They are designed to help assess a STRmix™ deconvolution result. No profiles required reinterpretation based on the review of the diagnostics.

The *LR*s were also reviewed as part of data quality checks. Large inclusionary *LR*s ($LR > > 1$) for false contributors and exclusionary *LR*s ($LR < 1$) for true contributors where the *APH* was relatively high were investigated. For any given mixture, there is a chance that a given false contributor will have sufficient matching alleles, by chance, to give an $LR > 1$. Likelihood ratios for false contributors above 10,000 are provided in Table 2. Following Taylor et al. [17];

1) The average *LR* for false contributors should be about 1.
2) The probability of observing a likelihood ratio of x or larger from an unrelated non-donor is no more than 1 in x.

These two statements fo rm the basis for assessing false contributor tests. In an experiment on 10,000 false contributors we would expect approximately one $LR \geq 10{,}000$, plausibly 10 above 1000 and 100 above 100. This work reports the comparison of approximately 20 million false contributors. The average *LR* for all false contributors is approximately 0.12. The reason that this average is below one is because the genotypes that would lead to the highest *LR*s (and so contribute significantly to the average) were not happened across in the number of H_d true tests performed.

The fraction of allele sharing for the twenty highest false contributors ranged from 0.61 to up to 0.98 of the alleles with the mixture (Table 2).

False exclusions were observed for known contributors where the apparent number of contributors was fewer than the ground truth

Table 2
Summary of large inclusionary *LR*s for false contributors and percentage of overlapping alleles.

Number	Kit	Apparent N	Known N	*LR*	Fraction of allele sharing
1	GlobalFiler™	3	3	505,924	0.81
2	Identifiler Plus™	3	3	379,716	0.90
3	GlobalFiler™	4	4	197,907	0.98
4	GlobalFiler™	3	4	134,486	0.83
5	GlobalFiler™	4	4	88,022	0.98
6	GlobalFiler™	4	5	53,019	0.93
7	Fusion 6C	3	3	47,062	0.85
8	Fusion 5C	3	3	43,065	0.78
9	Fusion 5C	3	3	26,874	0.80
10	GlobalFiler™	3	3	19,340	0.67
11	Fusion 5C	3	3	17,582	0.61
12	Identifiler Plus™	3	4	16,995	0.80
13	Fusion 5C	4	4	15,765	0.80
14	Identifiler Plus™	3	3	14,446	0.87
15	NGM SElect™	3	4	13,717	0.78
16	GlobalFiler™	4	5	12,135	0.93
17	Fusion 5C	4	6	11,188	0.93
18	Fusion 5C	3	3	10,896	0.80
19	Fusion 5C	3	3	10,309	0.82
20	Identifiler Plus™	3	3	10,298	0.80

number of contributors. This was an expected result [18,19]. By way of explanation we present an example of a true five contributor mixture interpreted assuming four contributors. Fig. 3 is a stylised electropherogram for one locus (SE33) with peaks and their corresponding height. STRmix™ has modelled the minor peaks as stutters of the eight alleles all above 800 rfu. Assuming four contributors and eight alleles, each contributor must be heterozygous at this locus. One known contributor who is homozygous at this locus (genotype 18,18) is therefore excluded ($LR_{SE33} = 0$) as a contributor under the assumption of four contributors. A second individual (genotype 12,23.2) is a poor fit to the profile assuming four contributors given the large peak imbalance for these alleles resulting in a low weight and subsequent *LR* at this locus ($LR_{SE33} = 0.01$).

False exclusions were also observed due to human error if, for example, an incorrect reference profile was supplied. Human errors were all corrected and the *LR*s reassigned. Another common reason for a false exclusion was due to the lack of separation of alleles during capillary electrophoresis. This occurred when peaks that differed by one base pair (for example a 9.3/10 at TH01) were not separated sufficiently during electrophoresis and one was subsequently not designated at analysis [14]. In all identified occasions an allele corresponding with a minor contributor was 'hidden' within the shoulder of an allele from a major contributor. Affected loci were identified by reviewing the electropherogram, and the locus was subsequently ignored during the interpretation.

3.2. Results for experiment 1

Violin plots [20] showing the densities of $\log_{10}(LR)$ per *APH* range are provided in Fig. 4 through 6 for apparent three, four and five contributor mixtures, respectively. The percentage of non-contributors giving $LR = 0$ is given at the bottom of each plot. The plots show the general trends for both H_p and H_d results.

Plots of $\log_{10}(LR)$ versus *APH* for all mixtures are given in the Supplementary material Figs. S1 through S9, plotted by apparent number of contributors. These plots are also separated into H_p true (*LR*s for true donors) and H_d true results (*LR*s for 10,000 false donors) and H_p and H_d true combined in order to help visualise the trends. In order to facilitate comparison between plots the axis scales have been retained for the same N. For the H_p true results where apparent N differed from the true N these results are indicated with a different plotting symbol. *LR* results of 0 (exclusions) have been plotted at −40 on the $\log_{10}$ scale. Normalisation of the CE platform (3130 versus 3500) had no effect on the trends present in the data and is not shown.

The vertical line of points in Fig. S8 at 50 rfu where $\log_{10}(LR) > 1$ are two siblings from a family study that included their biological father and mother. Due the complete allele sharing with both parents the *APH* for both siblings were calculated at half the AT, which is artificially low.

Fig. 4, Fig. 5 and Fig. 6 show the same trends as seen in previous work [14,21], with the addition of information regarding the consequence of over or underestimating the number of contributors. With increased information present within the profile (either by greater amounts of DNA, or by fewer contributors) the power to discriminate contributors from non-contributors increases, and there is a divergence of the *LR* from neutrality. Also consistent with previous findings [18], the underestimation of the number of contributors tends to either have little effect on the *LR* or will tend to exclude known contributors. This occurs because genotype sets possessing unreal allele pairings are forced to be given weight within the analysis. Interestingly this exclusionary effect was reduced as mixture complexity increased to the point that there were no exclusions produced from underestimating the number of contributors in five person mixtures (Fig. S1). We surmise that this is an effect of the increased allele sharing generally seen in higher order mixtures (Fig. 2) meaning that there are increased opportunities for genotype sets to possess the genotypes of the known contributors, even when their number is underestimated.

A plot of $\log_{10}(LR)$s for profiles generated using Identifiler™ Plus 28 cycles analysed on a 3130 or 3500 are plotted in Figs. S10 and S11 for the apparent three and four person mixtures, respectively (Supplementary material). As a visual aid we have added smoothed trend lines (LOWESS lines) for instrument type. These trend lines give a rough idea of the relationship between $\log_{10}(LR)$ and *APH* for different cases. Any trend line is a compromise between smoothness and error. We did not get materially different results when trying other trend lines

Peak	Height
12	892
14	116
15	1104
17	155
18	1899
22.2	186
23.2	2334
24.2	147
25.2	1386
26.2	1508
27.2	1410
30.2	89
31.2	953

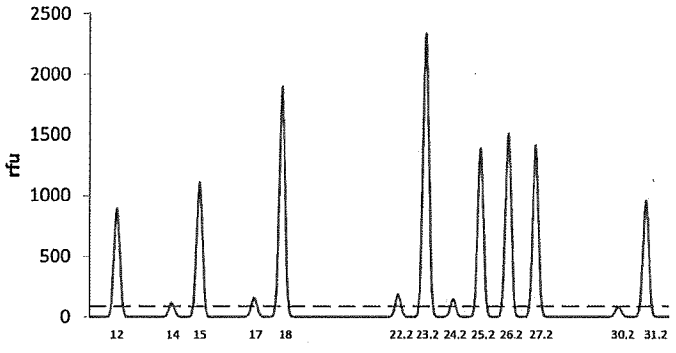


Fig. 3. Stylised locus electropherogram with tabulated peak designations and their corresponding heights for a true five person mixture interpreted assuming four contributors.

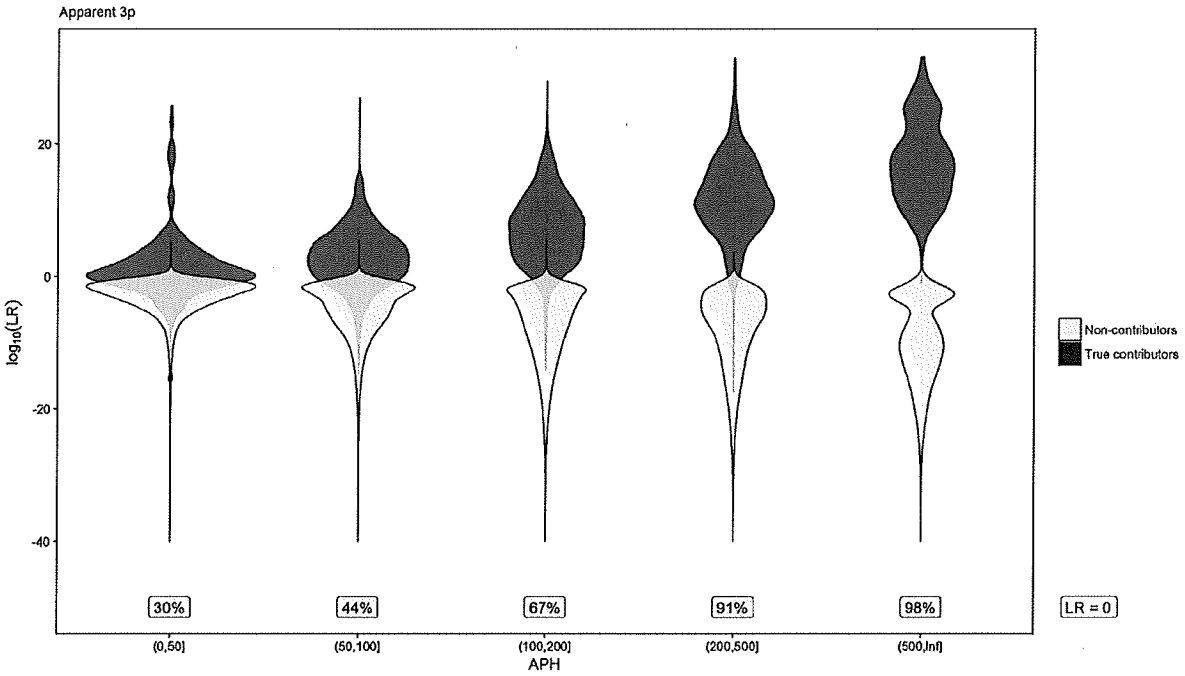


Fig. 4. Violin plot of $\log_{10}(LR)$ versus *APH* for apparent three contributor mixtures.

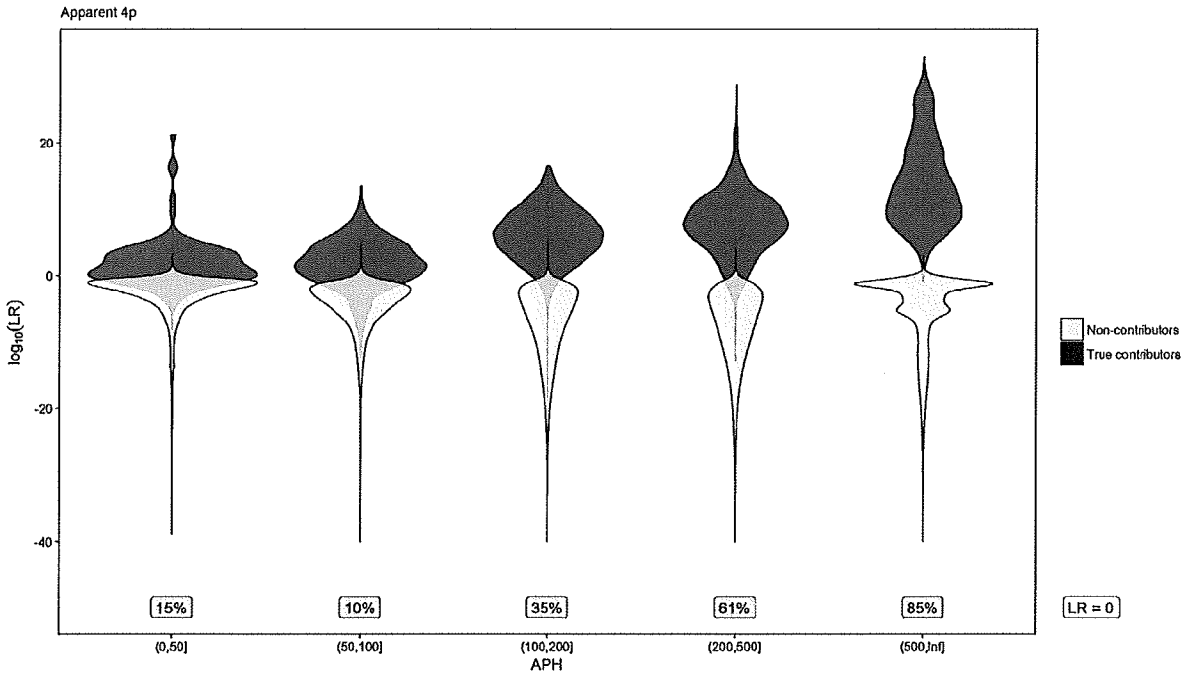


Fig. 5. Violin plot of $\log_{10}(LR)$ versus *APH* for apparent four contributor mixtures.

available in the ggplot2 package [22].

Applied Biosystems report a three- to fourfold increase in rfu scale with the 3500 models over the older Applied Biosystems 3100 and 3130 instruments [23]. This is evidenced by a general shift in the trend lines for the 3500 to the right in Figs. S10 and S11. The lines converge at high *APH* where the individual contributor profiles are likely fully represented and trend to $\log_{10}(LR) = 0$ as *APH* decreases.

Plots of $\log_{10}(LR)$s for true contributors identified by kit type are given in Figs. S12 and S13 for the apparent three and four person mixtures, respectively (Supplementary material). The LOWESS trend lines for kit type are modelled. These plots indicate the performance of the difference kits over *APH* for submitted profiles. As the profiles analysed are not the same between the different kits they are not suitable for comparing performance of the different kits. However, they do give an indication of general trends. As an example, comparing the trend lines for Identifiler™ versus GlobalFiler™ mixtures, at higher per contributor *APH* the $\log_{10}(LR)$s are higher for GlobalFiler™ profiles, most likely due to the additional loci within the GlobalFiler™ kit compared with the Identifiler™ Plus kit. $\text{Log}_{10}(LR)$ values for Identifiler™ profiles are generally higher at low contributor *APH* compared to GlobalFiler™ profiles, however. This could be due to the increased variability of the GlobalFiler™ profiles, all of which were analysed on

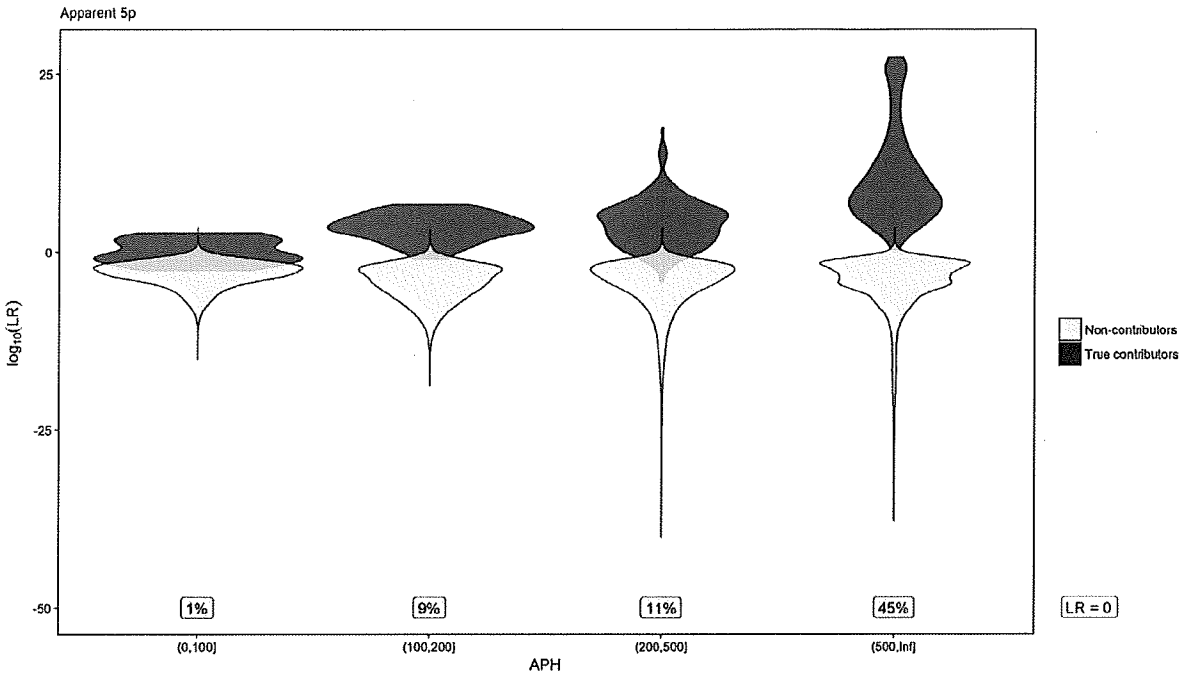


Fig. 6. Violin plot of $\log_{10}(LR)$ versus *APH* for apparent five contributor mixtures.

3500 instruments, in some cases with cycle numbers greater than 28 [24]. A comparison of the Fusion 5C and Fusion 6C trend lines illustrates the increase in discrimination achieved by adding the highly polymorphic STR locus SE33 resulting in generally higher $\log_{10}(LR)$s.

3.3. Results for experiment 2

The *LR*s for H_p true under the assumption of N and N + 1 contributors are presented in Fig. 7. Within Fig. 7 the size of the plotting symbols is relative to the contributor's proportion of the mixture. The *LR*s for H_d true are summarised in Figs. 8 and 9.

The results shown in Fig. 7 demonstrate some findings that are important for DNA mixture interpretation:

1. The general result was a decrease in the *LR* for true contributors after the assumption of an additional contributor to the mixture. The additional proposed contributor is interacting with the true contributors, diffusing the genotype weights, hence lowering the *LR*.
2. When a proposed person of interest aligns with the dominant component in a mixed DNA profile, the support for their inclusion to a mixture will not be markedly altered by an increase in the number of contributors under which the DNA profile is analysed. This is consistent with earlier findings [18].
3. Even when only donating a minor component of the total DNA, the change in *LR* produced by increasing the number of contributors is still not extreme. In no instances has an increase in the number of contributors seen an *LR* that strongly favours inclusion shift to one that favours exclusion.

We also consider the effect of contributor overestimation on H_d true tests. Fig. 8 shows the distribution of H_d true $\log_{10}(LR)$ values for three person mixtures when considered as originating from three (N) or four

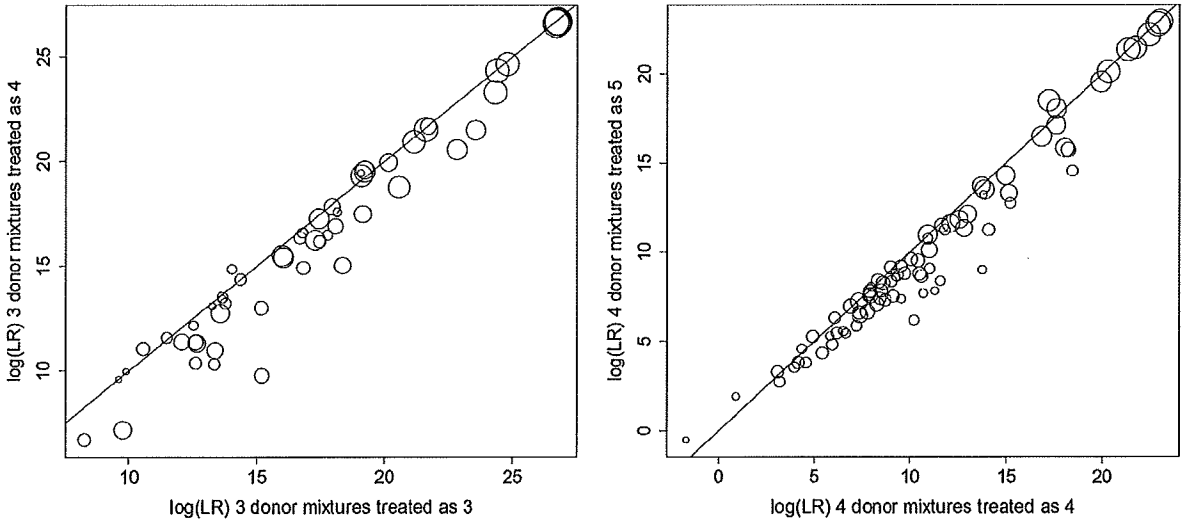


Fig. 7. The *LR*s for H_p true for three and four person mixtures from one laboratory under the assumption of N and N + 1 contributors. The $x = y$ line is shown. The size of the plotting symbol represents the mixture proportion of the donor.

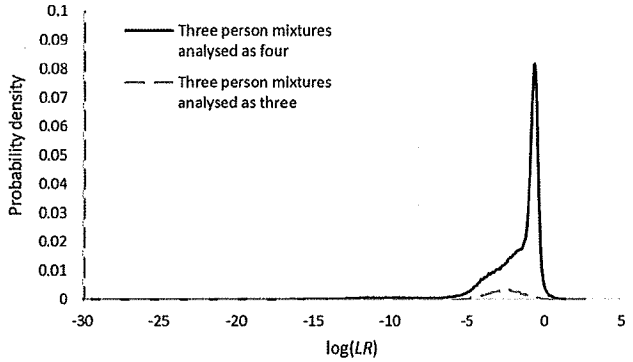
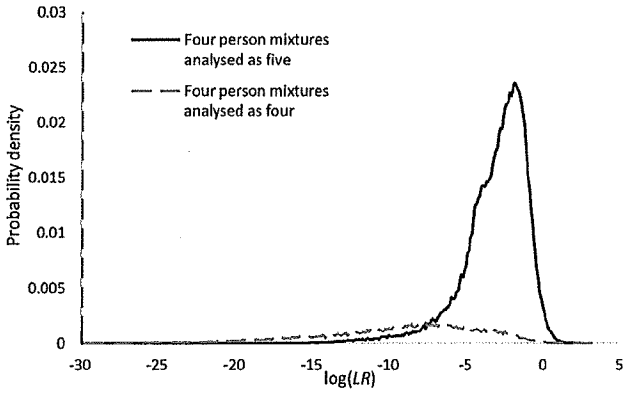


Fig. 8. The *LRs* for H_d true for three person mixtures from one laboratory under the assumption of N and N + 1. The bulk of the distribution for the three person mixtures analysed as three is at $LR = 0$ (90% of all *LRs*) represented by $\log_{10}(LR) = -30$.



Fig, 9. The *LRs* for H_d true for four person mixtures from one laboratory under the assumption of N and N + 1. 81% of four person mixtures analysed as four resulted in $LR = 0$, represented by $\log_{10}(LR) = -30$.

(N + 1) contributors. Fig. 9 shows the results of the same analysis but when considering four person mixtures as originating from either four (N) or five (N + 1) individuals. The bulk of the distribution for the three person mixtures analysed as three is at $LR = 0$ (90% of all *LRs*) represented by $\log_{10}(LR) = -30$ in Fig. 8. In Fig. 9, 81% of four person mixtures analysed as four resulted in $LR = 0$, again represented by $\log_{10}(LR) = -30$.

Figs. 8 and 9 show that, when analysed using the true number of contributors, the instances of H_d true comparisons that lead to outright exclusions is greatly increased. Put another way, inflating the number of contributors leads to an increase in non-zero *LRs*. In fact, the most common occurrence from inflating the number of contributors is that during deconvolution the additional proposed contributor is assigned a very low template (near 0) and can possess any genotype (including complete dropout) with relatively even weight. This is visually seen in Figs. 8 and 9 by the peak of $\log_{10}(LR)$s just below 0.

3.4. Allele sharing

A demonstration of the effect that allele sharing has on the *LR* is confounded by other factors that affect the magnitude of the *LR*, such as:

- The amount of DNA that the individual has donated to the sample,
- The mixture proportions of the contributors (mixtures at an even mixture proportion will tend to have lower *LRs*, due to the reduction in information that peak heights provide to determine genotype sets),
- Masking of minor contributors in stutter positions of major contributors.

An individual that shares 100% of alleles with the other contributors to a mixture can still have their genotype resolved completely, based on peak heights, given the right circumstances (as seen in Fig. S8 for the family set). The ability to use peak heights in this way is one of

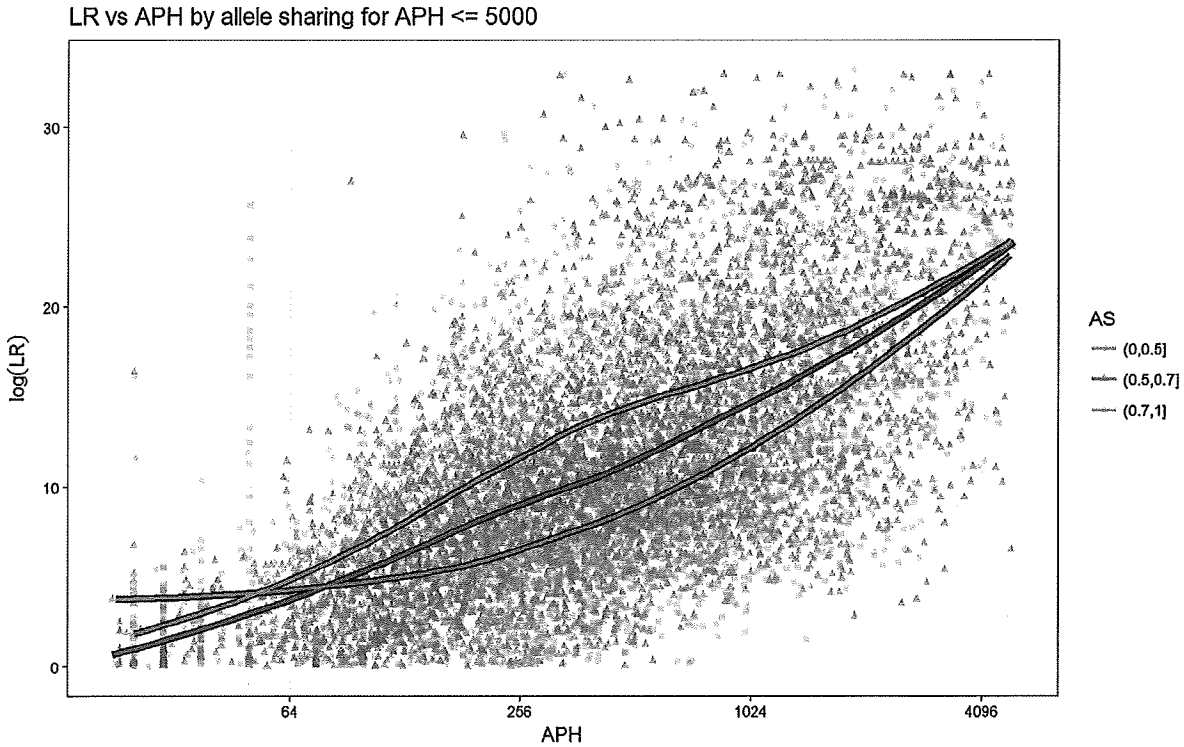


Fig. 10. The size of the $\log_{10}(LR)$ by considering differing amounts of input DNA (*APH*) and amount of allelic sharing (AS). The set of data points with high AS (0.7,1] are a family set (father, mother, children) where all alleles from the children are masked by the parents and therefore *APH* was set to half of the AT.

the main drivers for the differences in *LRs* produced between fully and semi-continuous systems. In Fig. 10 we show the *LR* (on $\log_{10}$ scale) for all data in the study, broken up into three categories of allele sharing, 0 to 0.5, 0.5–0.7 and 0.70–1.0. The lines in Fig. 10 are LOWESS lines to demonstrate the general trends of the data.

From Fig. 10, it appears that the greater the allele sharing, the less the power there is to discriminate a true contributor from a non-contributor. This trend is intuitive as it would be expected that the more an individual's alleles are already accounted for by others in the mixture, the less 'need' there is for someone possessing those alleles to reasonably explain the observed peaks in the mixture. However, further experimentation shows that this apparent trend is totally confounded by the number of contributors to the mixture. Fig. 11 shows the same style of result as Fig. 10, but plotted by number of contributors. In Fig. 11 the recovered weight of evidence is plotted, that is, $\log_{10}(LR)/\log_{10}(1/RMP)$. RMP is the conditional match probability following the Balding and Nichols model [25] and a theta (F_{ST}) of 0.01. Carrying out this transformation accounts for the different profiling systems that are being combined in this *meta*-analysis. In these plots the *y*-axis is bounded by one demonstrating that the *LR* cannot exceed one divided by the random match probability.

The trend seen in Fig. 2 is that higher order mixtures tend to have true contributors that share more alleles (because there are more of them to potentially share), and Figs. S1–S9 demonstrate that higher order mixtures tend to have less discrimination power. Therefore, there is a correlation between allele sharing and *LR* evident in Fig. 10, particularly at low *APH*. In Fig. 11 this trend disappears, showing that it is an effect of number of contributors, and not allele sharing, that is the main driver to *LR* change.

In Fig. 12 we plot a density plot of $\log_{10}(LR)/\log_{10}(1/RMP)$ by the amount of allele sharing of the non-contributors with the true contributors. The $\log_{10}(LR)/\log_{10}(1/RMP)$ cannot exceed one, which would indicate a fully resolved component. Inspection of Fig. 12 shows that as the fraction of shared alleles increases the $\log_{10}(LR)/\log_{10}(1/RMP)$ for the non-contributor increases. As allele sharing of the non-contributors with the true contributors decreases, the $\log_{10}(LR)/\log_{10}(1/RMP)$ decreases with more observations around zero, indicated by the broadening of shape. Fig. 12 shows that non-contributors are unlikely to yield large *LRs* even if they share many alleles with the true contributors. In other words, non-contributors that share most of their alleles with the mixture's donors can typically still be excluded because the peak heights make their inclusion unlikely.

On the other hand, Fig. 6 shows that true contributors can yield *LRs* close to the inverse of the single source match probability even in five person mixtures. This means that at least this mixture donor's component is almost fully resolved on the basis of peak heights. This may be expected, for instance, in a 10:1:1:1:1 mixture where the major may be clearly resolved by simply 'eyeballing' the electropherogram.

4. Discussion

4.1. Performance of the system with regards to contributor number

In principle, we observe less discriminatory *LRs* for true and non-contributors when the number of assigned contributors increases. This has been demonstrated previously using STRmix™ [14,21]. This does not mean that mixed DNA profiles containing more contributors are less reliable, just that they are less informative with respect to potential contributors.

The *true* number of contributors to a crime profile is never known. Within this work we have used the apparent number of contributors when interpreting the mixtures. Apparent N was determined by each submitting laboratory using their own validated methods. The assigned N can be fewer than the true N when individuals within a profile have "dropped out" (their alleles falling below the detection limit of the CE) and within mixtures of contributors with high amounts of allele sharing (an extreme example being mixtures of related individuals). Apparent N may be assigned a number higher than true N in the presence of artefacts, such as stutter, that are larger than expected. This assignment can be confounded in saturated profiles.

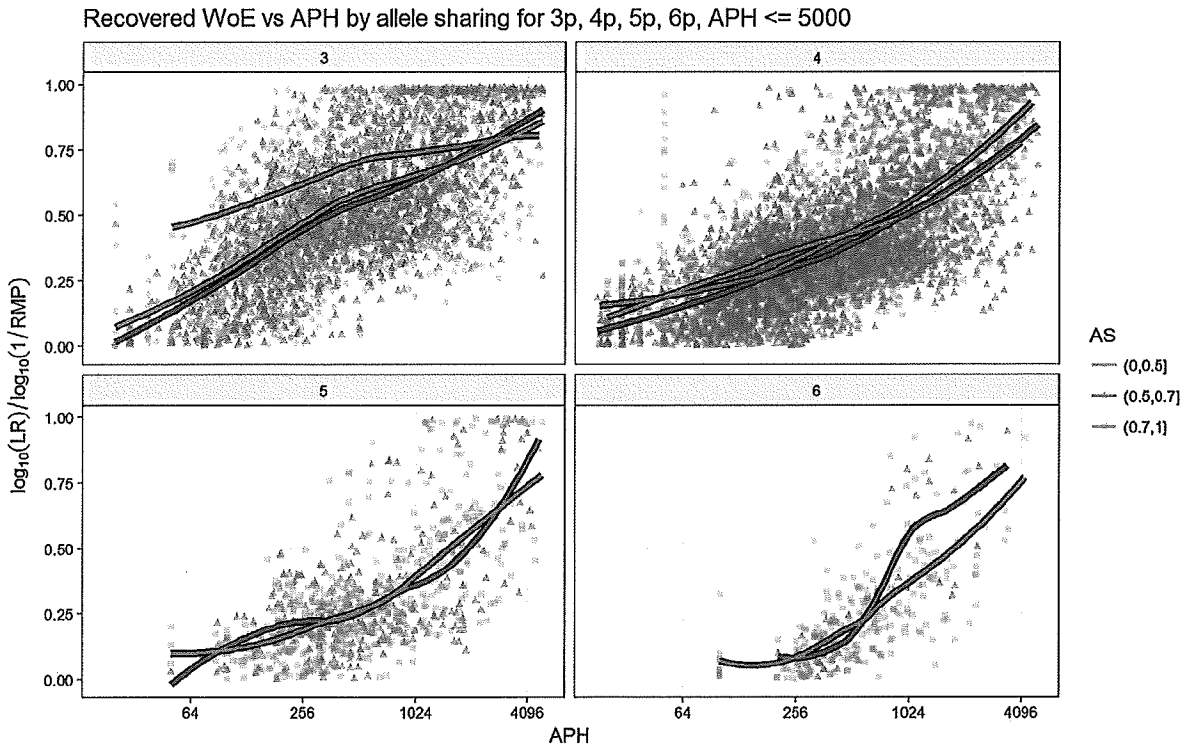


Fig. 11. The size of the recovered weight of evidence $\log_{10}(LR)/\log_{10}(1/RMP)$ by considering differing amounts of input DNA (*APH*) and amount of allelic sharing (AS) plotted by true number of contributors.

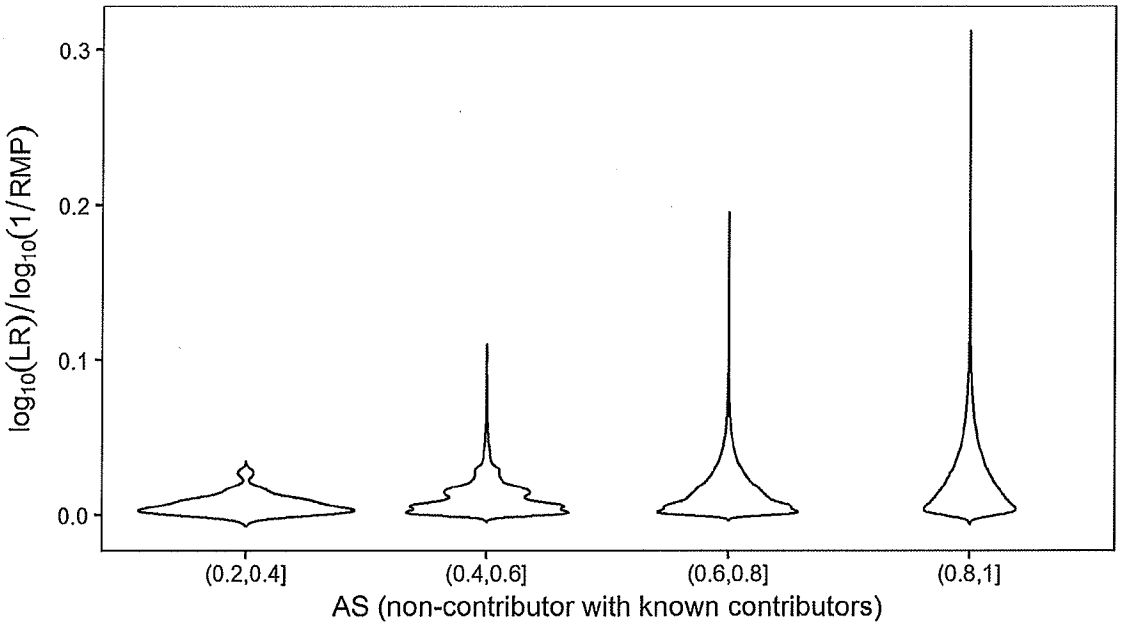


Fig. 12. Density plot of $\log_{10}(LR)/\log_{10}(1/RMP)$ by the amount of allele sharing of the non-contributors with the true contributors.

As the number of contributors to a DNA profile increases, the DNA mixture becomes more complex. Figs. S1 through S9 show *LRs* generated for H_p and H_d true for apparent three, four and five person mixtures plotted against *APH*. As the number of contributors to the mixture increases the *LRs* trend towards one. This holds true for both H_p and H_d true although the effect for H_d true data is less clear given the number of data. As the number of contributors to a mixture increases, so too do the potential genotype combinations that can explain the observed data. This results in an overall reduction in the weights assigned to each genotype set, as these weights are spread across more potential genotype sets. This behaviour was previously described by Taylor [21].

When overestimating the number of contributors to a mixture (N + 1) the *LR* generally decreased for true contributors. This can be explained by STRmix™ spreading the weights for the true donors across more genotype sets. For four person mixtures the magnitude of the effect on the *LR* for known contributors was somewhat dependent on the proportion that the donor contributed to the mixture. The effect was greater for minor contributors to the mixture and less for major contributors (represented by more data points on the $x = y$ line within Fig. 7). Overestimating the number of contributors had little or no effect on the *LR* of the major contributor to the mixture, demonstrated by the largest circles sitting on the $x = y$ trend line. In these cases the additional proposed contributor was modelled as a trace contributor, sharing alleles with the true minor contributors to those mixtures and having little effect on the major. For the three person mixtures the effect was more visible across a range of mixture proportions. This was likely due to similarities in mixture proportions of the different contributors, with no obvious major contributors.

The effect of overestimation of the number of contributors was also determined for non-contributors using H_d true tests. When assuming N + 1 the number of occurrences of non-contributors resulting in non-exclusionary *LRs* increased. During deconvolution the additional

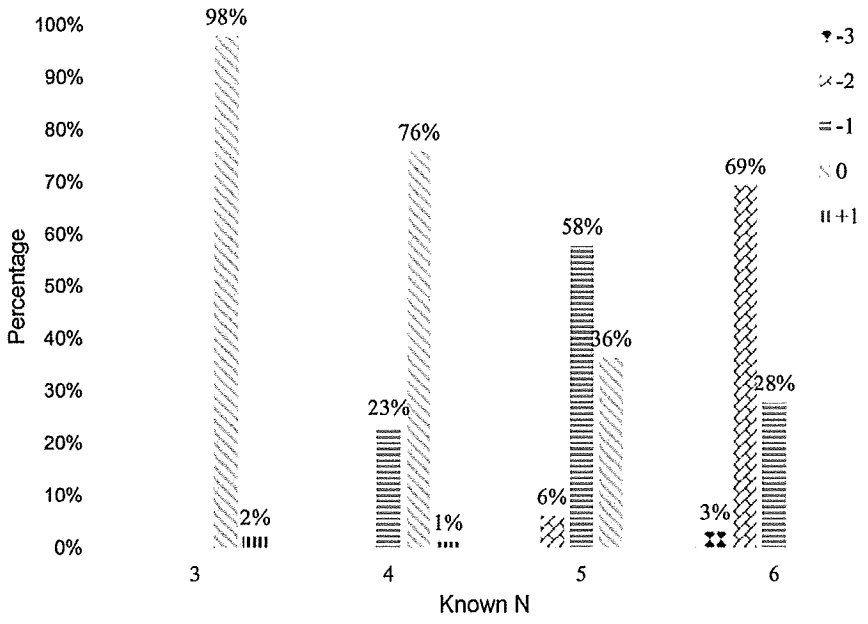


Fig. 13. Plot of percentage of mixtures showing various differences between apparent N and known N against known N. As an example, −1 indicates apparent N was one fewer than known N.

proposed contributor is assigned very low template and can possess any genotype leading to these results.

In summary, overestimation of the number of contributors generally leads to lower *LR*s for true contributors (Fig. 7) and an increase in *LR*s for non-contributors (Fig. 8).

Underestimating the number of contributors can result in false exclusions of true donors. In this study, this is seen when apparent N is fewer than true N. This is demonstrated in the H_p true plots within the Supplementary material where apparent N that differs from known N are indicated with a different plotting symbol.

When assigning N, for false donors the only risk is overestimation, as there is a small increase in the number of very low grade false inclusions. With respect to the *LR* for true donors, you are either correct or conservative when N is either under or overestimated.

In Fig. 13 we provide a plot showing the level of over and underestimation of the apparent N compared to the known N in this study.

Fig. 13 shows that an underestimation of N was more common than an overestimation of N. There are three broad reasons why N might be underestimated:

1) One contributor has donated so little DNA that their presence is unseen in the DNA profile, we call this the tiny minor scenario;
2) Contributors are present so that one or more is completely masked by others in the profile, and in a way so that peak height does not reveal their presence. This is the hidden contributor scenario;
3) There is a combination of multiple low-level contributors that, due to some masking and some dropout, produce a profile where the apparent number of contributors is fewer than the known number of contributors. This is the low level donors' scenario.

Each of these is discussed in turn below.

4.1.1. The tiny minor

Any profile is a result of fragments of DNA that have been aliquoted from a DNA extract and then amplified during PCR. There exists a possibility that no DNA fragments from a minor DNA donor have been sampled for PCR. We first ask what we consider to be the correct number of contributors; the number of different individual's DNA in the DNA extract, or the number of different individual's DNA in the PCR? If it is the former, then we would ask; if the individual has contributed so little DNA that the observed fluorescence in the DNA profile is not affected by their presence, then what purpose is served by considering them as a contributor? We note that many of the underestimates of number of contributors in this study arise from such situations.

4.1.2. The hidden contributor

Consider a DNA profile where multiple individuals, are contributing to a DNA profile, however they possess sufficient allelic overlap so that the DNA profile appears as a lower order mixture. The apparent number of contributors being lower than the known number of contributors relies on the DNA profile being formed in such a way that peak imbalances will not indicate the true number of contributors. For example, a combination of two individuals who are homozygous at each locus, combined in equal proportions to a DNA sample will always appear single source. However, this risk of multiple contributors being combined to meet theses specifications is very remote, and artificial. It only tends to occur in mixtures of family members, such as a child and their parents donating equal amounts of DNA to a sample. The Coble et al. [26] experiment is valuable but does not take into account peak heights, and so the study does not reflect the information that peak heights provide analysts in their assignment of N. This is evident in the difference between the results obtained by Coble et al. and our work. For example, Coble et al. reported the probability of a known five-person mixture presenting as an apparent five person mixture was less than 0.01, whereas in our study, based on human assignment, this probability is 0.36 (and noting that many of the remaining mixtures fall into the tiny minor and low donor scenarios).

4.1.3. The low level donors' scenario

This scenario is where there are multiple low level contributors, who are present in low amounts such that they exhibit significant dropout and so in combination the apparent number of contributor is fewer than the known number of contributors. This is a scenario that could plausibly occur with reasonable probability when multiple low level contributors are present (see [16] for an exploration of this). Experimentation has shown that very low level contributors will yield *LR*s of approximately one. It is likely that when analysed under the known number of contributors, all true (and a majority of false) contributors give this neutral *LR* value. In other words, the profile does not have the information in order to distinguish true from false donors. If analysed as the apparent number of contributors then the likely outcome is an exclusion of the known contributors (and more exclusions of non-contributors). The primary difference in *LR* between known and apparent number of contributors is between neutral and possibly exclusionary, which we could argue presents less risk of misleading a court.

4.1.4. Overestimating the number of contributors

Our studies show that the chance of overestimating N in relation to the known value is less common that underestimation and cannot be predicted so easily by simulation as in Coble et al. [26]. It requires two events to occur:

1) There is a stochastic event, such as a peak imbalance, high stutter or drop-in, which occurs at an improbable level,
2) The analyst interpreting the profile feels that the out-of-place fluorescence has resulted in a profile that is more likely to exist if it has originated from more contributors that the known number of contributors.

Fig. 7 shows that the effect of overestimation of N is relatively mild on known contributors to a DNA profile. STRmix™ assigns near-zero mass to the non-existent contributor, leaving the other contributors relatively unchanged. The largest effect is to decrease the *LR* for minor known contributors. For non-contributors, Fig. 8 shows the effect that has previously been described, i.e. that an overestimation of N tends to increase low-level *LR*s for non-contributors. In effect the experiment is showing the practical functioning of the catch-all statement suggested earlier.

Our findings show that as mixture complexity increases, the ability of an analyst to designate the known number of contributor is reduced. As explained, it is actually often the apparent number of contributors that is the more appropriate value to choose for analysis. In assigning apparent number of contributors the overwhelming result is alignment with the desired trends in *LR*s with regards to profile complexity and DNA amount (i.e. those described in [21], where known number of contributors was used for all analyses) are obtained. In the rare circumstances where the known contributors were not supported as donors of DNA to the profile, this was due to one of the three underestimate conditions described above in 4.1.1 through 4.1.3 above.

4.2. Performance as a function of amount of allele sharing

Within Fig. 10 the trend is that the greater the allele sharing, the less the power to discriminate a true contributor from a non-contributor. However, this relationship is dominated by the number of contributors within the mixture (as seen in Fig. 11). Higher order mixtures result in less informative *LR*s. This effect is related more to the number of contributors within a mixture than the amount of allele sharing between contributors within the mixture. There is a relationship between the number of contributors and proportion of allele sharing within a mixture. It has previously been shown that the probability of a higher order mixture appearing as having originated from one fewer individual

based on allele count alone is high [26,27]. For example, Coble et al. calculated the probability of a six contributor profile appearing as a five contributor profile based on allele count as 0.8599 for the GlobalFiler™ 24 locus multiplex [26]. The study by Coble et al. did not take into account peak height, thereby making the values in their study a worst case scenario.

4.3. Performance of the system with regards to amount of DNA

In principle, we observe less discriminatory *LRs* for true and non-contributors when the *APH* (template) decreases per contributor. Again, this does not mean that mixed DNA profiles with contributors containing less DNA are unreliable, just they are less informative with respect to the true and non-contributors.

PCAST describe limits on PG reliability based on mixture proportion and number of contributors. Per contributor template is more informative of *LR* than mixture proportion. With respect to mixture proportion, the limit is not the software but the hardware. For example, assuming a minor contributor's alleles within a mixture are present just above the analytical threshold of a 3130 (typically 50 rfu) and a major contributor's alleles are at the saturation limit (typically 7000 rfu), this would be maximum mixture proportion of 140:1. 2293 out of the 2825 submitted profiles had at least one component who contributed less than 20% of the sample.

5. Conclusion

In their review of published literature validating probabilistic genotyping, PCAST surmised that the limits of foundational validity extended to three person mixtures where the person of interest made up at least 20% of the profile. What was not taken into account during the PCAST review was a wealth of unpublished validation material residing in laboratories that had validated (or were in the process of validating) probabilistic genotyping software. Due to our involvement with STRmix™ we are aware of the breadth of such validation material for STRmix™ specifically, and assume that similar material must be present for other probabilistic genotyping systems. A disconnect exists between the PCAST desire for laboratories to publish their validation material in peer reviewed journals and the general resistance to such publications by the journals themselves. This is for the completely understandable reason that they are generally not novel, or, individually, of general interest to the forensic community.

> *PCAST has said "When further studies are published, it will likely be possible to extend the range in which scientific validity has been established to include more challenging samples. As noted above, such studies should be performed by or should include independent research groups not connected with the developers of the methods and with no stake in the outcome."*

There has already been an example of published material that extend the PCAST limits, from the Forensic Biology laboratory at the Federal Bureau of Investigation [14]. We add to that published work, by compiling the STRmix™ validation material from 31 laboratories, which allows a novel look at data spanning laboratory technology and process. PCAST highlighted four key areas that they felt additional validation would be merited:

(1) How well does the method perform as a function of the number of contributors to the mixture? How well does it perform when the number of contributors to the mixture is *unknown*?
(2) How does the method perform as a function of the number of alleles shared among individuals in the mixture? Relatedly, how does it perform when the mixtures include related individuals?
(3) How well does the method perform—and how does accuracy degrade—as a function of the absolute and relative amounts of DNA from the various contributors?
(4) Under what circumstances—and why—does the method produce results (random inclusion probabilities) that differ substantially from those produced by other methods?

We address points 1 to 3 in this study. It is unknown whether further addendums will be released by the PCAST group, or whether there are any plans for a follow-up study in the future. The material we provide here demonstrates a foundational validity of, at least, the STRmix™ software method for complex, mixed DNA profiles to levels well beyond the complexity and contribution levels suggested by PCAST. The study was done in accordance with the specific manner outlined in the PCAST report.

Acknowledgements

This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of their organisations. The authors would like to thank Professor James Curran for his help in creating the plots in Fig. 1.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2018.01.003.

References

[1] President's Council of Advisors on Science and Technology, PCAST Releases Report on Forensic Science in Criminal Courts, (2016).
[2] President's council of advisors on science and technology, An Addendum to the PCAST Report on Forensic Science in Criminal Courts, (2016).
[3] A.A. Mitchell, J. Tamariz, K. O'Connell, N. Ducasse, Z. Budimlija, M. Prinz, et al., Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in, Forensic Sci. Int. Genet. 6 (2012) 749–761.
[4] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, Validating TrueAllele® DNA mixture interpretation, J. Forensic Sci. (2011) 2011.
[5] J.-A. Bright, D. Taylor, C.E. McGovern, S. Cooper, L. Russell, D. Abarno, et al., Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles, Forensic Sci. Int. Genet. 23 (2016) 226–239.
[6] F.M. Götz, H. Schönborn, V. Borsdorf, A.-M. Pflugbeil, D. Labudde, GenoProof Mixture 3—New software and process to resolve complex DNA mixtures, Forensic Sci. Int. Genet. Suppl. Ser. (2017).
[7] S. Manabe, C. Morimoto, Y. Hamano, S. Fujimoto, K. Tamaki, Development and validation of open-source software for DNA mixture interpretation based on a quantitative continuous model, PLoS One 12 (2017) e0188183.
[8] Scientific Working Group on DNA Analysis Methods (SWGDAM), Guidelines for the Validation of Probabilistic Genotyping Systems, (2015).
[9] M.D. Coble, J. Buckleton, J.M. Butler, T. Egeland, R. Fimmers, P. Gill, et al., DNA Commission of the International Society for Forensic Genetics: recommendations on the validation of software programs performing biostatistical calculations for forensic genetics applications, Forensic Sci. Int. Genet. 25 (2016) 191–197.
[10] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int. Genet. 7 (2013) 516–528.
[11] J.-A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Sci. Int. Genet. 7 (2013) 296–304.
[12] The New York City Office of Chief Medical Examiner, Internal Validation of STRmix™ V2.4 for Fusion NYC OCME, (2016).
[13] District of Columbia Department of Forensic Science Forensic Science Laboratory Forensic Biology Unit. Internal Validation of STRmix™ V2.3, (2015).
[14] T.R. Moretti, R.S. Just, S.C. Kehl, L.E. Willis, J.S. Buckleton, J.-A. Bright, et al., Internal validation of STRmix for the interpretation of single source and mixed DNA profiles, Forensic Sci. Int. Genet. 29 (2017) 126–144.
[15] T.R. Moretti, L.I. Moreno, J.B. Smerick, M.L. Pignone, R. Hizon, J.S. Buckleton, et al., Population data on the expanded CODIS core STR loci for eleven populations of significance for forensic DNA analyses in the United States, Forensic Sci. Int. Genet. 25 (2017) 175–181.
[16] D. Taylor, J. Buckleton, J.-A. Bright, Does the use of probabilistic genotyping change the way we should view sub-threshold data, Aust. J. Forensic Sci. 49 (2017) 78–92.
[17] D. Taylor, J. Buckleton, I. Evett, Testing likelihood ratios produced from complex DNA profiles, Forensic Sci. Int. Genet. 16 (2015) 165–171.
[18] J.-A. Bright, D. Taylor, J. Curran, J. Buckleton, Searching mixed DNA profiles directly against profile databases, Forensic Sci. Int. Genet. 9 (2014) 102–110.
[19] J.-A. Bright, J.M. Curran, J.S. Buckleton, The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation, Forensic Sci. Int.

Genet. 12 (2014) 208–214.
[20] J.L. Hintze, R.D. Nelson, Violin plots a box plot-density trace synergism, Am. Stat. 52 (1998) 181–184.
[21] D. Taylor, Using continuous DNA interpretation methods to revisit likelihood ratio behaviour, Forensic Sci. Int. Genet. 11 (2014) 144–153.
[22] H. Wickham, Ggplot2–Elegant Graphics for Data Analysis, 2nd edition, Springer-Verlag, New York, 2016.
[23] Applied Biosystems User Bulletin Applied Biosystems® 3500/3500xL Genetic Analyzer, Life Technologies internal report, Foster City, CA), 2011.
[24] D. Taylor, J. Buckleton, J.-A. Bright, Factors affecting peak height variability for short tandem repeat data, Forensic Sci. Int. Genet. 21 (2016) 126–133.
[25] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (1994) 125–140.
[26] M.D. Coble, J.-A. Bright, J.S. Buckleton, J.M. Curran, Uncertainty in the number of contributors in the proposed new CODIS set, Forensic Sci. Int. Genet. 19 (2015) 207–211.
[27] J.S. Buckleton, J.M. Curran, P. Gill, Towards understanding the effect of uncertainty in the number of contributors to DNA stains, Forensic Sci. Int. Genet. 1 (2007) 20–28.