Article:

Buckleton, J. S., Bright, J. -A., Gittelson, S., Moretti, T. R., Onorato, A. J., Bieber, F. R., Budowle, B., Taylor, D. A. (2018). **The Probabilistic Genotyping Software STRmix: Utility and Evidence for its Validity**. *Journal of Forensic Sciences*, 64(2), 393–405.

The Probabilistic Genotyping Software STRmix: Utility and Evidence for its Validity

John Buckleton[1,2*], Jo-Anne Bright[1], Simone Gittelson[3], Tamyra Moretti[4], Anthony Onorato[4], Frederich Bieber[5], Bruce Budowle[6] and Duncan Taylor[7,8]

[1] *Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland 1142, New Zealand*

[2] *Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand*

[3] *Centre of Forensic Science, University of Technology, Sydney, P.O. Box 123, Broadway, NSW 2007, Australia*

[4] *DNA Support Unit, Federal Bureau of Investigation Laboratory, 2501 Investigation Parkway, Quantico, VA 22135*

[5] *Center for Advanced Molecular Diagnostics, Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, 75 Francis Street, Boston, MA 02115*

[6] *Center for Human Identification, Department of Microbiology, Immunol-ogy, and Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107*

[7] *Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia*

[8] *School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia*

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

Abstract

Forensic DNA interpretation is transitioning from manual interpretation based usually on binary decision making towards computer based systems that model the probability of the profile given different explanations for it, termed probabilistic genotyping (PG). Decision making by laboratories to implement probability-based interpretation should be based on scientific principles for validity and information that supports its utility, such as criteria to support admissibility. The principles behind STRmix™ are outlined in this paper and include standard mathematics and modeling of peak heights and variability in those heights. All PG methods generate a likelihood ratio (LR) and require the formulation of propositions. Principles underpinning formulations of propositions include the identification of reasonably assumed contributors. Substantial data have been produced that support precision, error rate, and reliability of PG, and in particular STRmix™. A current issue is access to the code and quality processes used whilst coding. There are substantial data that describe the performance, strengths and limitations of STRmix™, one of the available PG software.

Keywords

Forensic science, DNA; probabilistic genotyping; validation; STRmix™

A common binary method, dating to the 1990s, for the interpretation of short tandem repeat (STR) typing results from forensic casework analyses, while valid (1-3), had two primary limitations: first, for some specimens, a substantial amount of profile data could not be used for calculating statistical weight, resulting in more inconclusive results; and second, a number of laboratories faced challenges in interpretation of complex forensic DNA mixtures.

In recent years, more sophisticated approaches to applying the fundamental principles of DNA mixture interpretation have been incorporated into customized software that expands the capabilities of the forensic analyst. These tools bring together refined methods of biological modeling, probability, and computational power that provide more meaningful empirical assignments of evidentiary weight.

Substantial data have been generated and accumulated that demonstrate the utility of probabilistic genotyping (PG). As exemplified herein with one software solution for PG, STRmix™, this document focuses on sound practices for forensic DNA mixture interpretation, the attendant statistical analyses using STRmix™, and other application issues related to admissibility. The topics covered in this effort include:

1. An introduction to PG, the likelihood ratio (*LR*), and setting propositions,
2. Validity of STRmix™,
3. A discussion about admissibility, peer review, code disclosure, and independent testing of STRmix™,
4. A discussion on assigning the number of contributors, and
5. The effect of relatives in mixtures.

This resource should guide the reader in becoming familiar with the salient features of STRmix™, as well as its strengths and limitations.

Introduction to probabilistic genotyping

The interpretation of forensic DNA mixture evidence is moving towards PG. The Scientific Working Group on DNA Analysis Methods (SWGDAM) defines PG as "*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…*" (4). Biological modeling is based on numerical criteria that can be encoded into the software to aid in interpretation of DNA profile characteristics such as peak height, base pair size, stutter, DNA degradation, allele dropout, and drop-in. The conceptual basis for PG was in place by 2000 (5-9). Advances to this initial concept were made and encoded in the software LoComatioN (10). Workable PG solutions were not implemented into routine forensic casework until about 2009 following advancements and the development of other programs, such as TrueAllele® (11).

Broadly there are two categories of PG software: Semi-continuous and fully continuous. The key difference between them is that semi-continuous models do not consider allele peak heights, while fully continuous methods make direct use of such information. Both the semi- and fully continuous methods assess the probability of observing the mixed DNA profile given proposed genotypes for the contributors. The semi-continuous methods assign a probability to the profile given a genotype combination, and the number is in the continuous interval [0,1]. The mathematical details are not described herein, but all the PG solutions utilize some form of 'nuisance parameter' for a factor that must be accounted for in the process. In most semi-continuous applications, this parameter is the assignment of the probability of allele dropout. The semi-quantitative model by Slooten (12) in the software product MixKin removes the nuisance parameter by the preferred method of integration. The programs LRmix (13), LikeLTD (14), or Lab Retriever (15) use plug-in values or the value derived by the method of maximum likelihood estimation (MLE) rather than from the integral. In the case of LRmix, an allele dropout value is assigned, often following a sensitivity analysis; LikeLTD assigns the nuisance parameter by MLE; and in the case of Lab Retriever, this value is assigned using a form of logistic regression that does not account for degradation (see (16) for a review of some logistic regression methods). Lab Retriever contains an additional approximation to a population genetic model introduced for computational convenience (17). This approximation is unlikely to have any large effect.

STRmix™, TrueAllele® (11), and GenoProof Mixture 3 (18) are fully continuous methods that are based on a Markov chain Monte Carlo (MCMC) resampling method (19). The use of MCMC is not novel and has been used to solve many complex problems within chemistry, physics, biology, statistics, and computer science. The continuous model software *Kongoh* (20) utilizes MLE. Other continuous solutions of which we are aware include LikeLTD-ht (14), DNAmixtures, (21, 22) and EuroForMix (23).

Introduction to likelihood ratios (*LRs*)

The outputs of all PG software are *LR*s. The *LR* is a ratio of the probability, the probability density, or quantities proportional to either probability or density of some specific observations or findings when considering two alternative (i.e., mutually exclusive) propositions. As applied to forensic DNA typing, the ratio, in its simplest form (i.e., a single source specimen), expresses the probability of the DNA evidence if a person of interest (POI) rather than an unknown individual is the source of the DNA.

Bayes' theorem follows immediately from the laws of probability and in the current context may be expressed in the following form: Posterior odds = *LR* × prior odds

Whatever the odds are on the person of interest (POI) being a contributor without considering the DNA evidence (i.e., the prior odds), this theorem describes that these odds should be increased (or decreased) by *LR* times upon considering the DNA evidence. In practice, it is the *LR* rather than posterior odds that is typically presented in court.

Naming the propositions

Ian Evett in "What is the probability this blood came from that person?" (24) recognized the work of Dennis Lindley on probability and Bayesian theory. Evett used the terms C (contact) and $\bar{C}$ (non-contact) to describe alternate propositions used for the *LR*. Subsequently, H_p and H_d were introduced as the prosecution and defense hypotheses, respectively (25). The prosecution proposition is usually straightforward (i.e., the defendant is the source of the DNA on the evidence), while the defense proposition can vary substantially. Argument can arise about assertions, such as:

1. The expert should not assume what the defense proposition may be,
2. The defense is entitled to all propositions consistent with exoneration and should not be constrained to one proposition, and
3. The defense is not obligated to provide a proposition.

Using the terms prosecution and defense may contribute to some contention when considering propositions. Therefore, as with the earlier C and $\bar{C}$ espoused by Evett, alternate propositions without such descriptors might be sensible. H_p and H_d could readily be replaced with, for example, H_1 and H_2 (15, 26), H_1 and H_a (where '*a*' stands for alternate), or H_C and $H_{\bar{C}}$ (referring to contributor and non-contributor), to avoid the implications of the '*p*' and '*d*' labels.

The two propositions used for the *LR* must be exclusive, meaning that they cannot both be true at the same time. They should also be exhaustive (cannot both be false at the same time) within the context of the specific case. For example, consider that the prosecution asserts that the defendant is the source of the DNA, and the alternate proposition is that the source of DNA is a random person unrelated to the defendant. In this case, H_p and H_d may both be false, for example the DNA could be from the defendant's brother, in which case these propositions are not exhaustive. However, propositions should always be considered in light of the accepted background information (*I*), and if a proposition is very unlikely or impossible given *I*, then it need not be considered by an analyst. In the context of the case discussed above in which *I* is that the defendant has no brother, this possibility need not be considered. We direct the reader to Biedermann et al. (29) for a more in depth discussion on setting propositions.

Transfer and persistence of DNA

The concept of hierarchy of propositions is well established (27, 28). Gittelson et al. (29) discussed this concept more recently. Propositions are classified into four levels: offence, activity, source, and sub-source.

- Offence level propositions describe the issue for the fact finder which is one of guilt or innocence.
- Activity level propositions describe the activity that deposited the DNA.
- Source level refers to the origin of the body fluid or cell type examined.
- Sub-source level refers to the origin of the DNA (i.e., donor).

Gittelson et al. (29) suggested a requirement for interpretation: "*Due attention must be paid to the position in the hierarchy of propositions that can be considered. This information must be effectively conveyed to the court to avoid the risk that an evaluation at one level is translated uncritically and without modification to evaluation at a higher level.*"

They further stated: "*We cannot emphasise the importance of this enough. A DNA match may inform decisions about the source of the DNA, but decisions about an activity, say sexual intercourse versus social contacts, involve additional considerations beyond the DNA profile*."

Transfer and persistence of DNA are relevant for an activity level evaluation of the DNA results, whereas PG software and other interpretation and statistics methods evaluate the DNA results at the sub-source level. Discussion of transfer and persistence therefore has nothing to do with PG. However, a sub-source level evaluation of the DNA results may be necessary for evaluating the findings with regard to a pair of activity level propositions.

Effect of different propositions when using STRmix™

The assignment of propositions should be made from the relevant background information (29, 30). The following issues should be considered:

- Which, if any, of the known individuals may be reasonably assumed to be contributors (29-31),
- The number of contributors to the profile (discussed later in this paper) (32-36),
- How to deal with multiple POIs (29, 30),
- How to deal with evidential items associated with neither the POI nor the victim (29, 30), and

The effects of these considerations are summarized here.

Assumption of the presence of an individual's DNA in a mixture

If a genotyped person, say the complainant in a sexual assault, can reasonably be expected to have donated DNA to the sample and the profile suggests their presence, then that person should be included under both the prosecution and defense propositions.

There are three principles that could be applied when making this decision.

First, any person should be assumed to be a contributor if the presence of his/her DNA is reasonably expected and the mixture is explained well by their inclusion. One reasonable expectation of the person's DNA being present (e.g., Option 1 in Table 1) is if the item of evidence is derived from an intimate sample of this person such as a vaginal swab. This concept can reasonably be extended to other items associated with the person, for example their clothing. Accordingly, it is important that any such assumption of the presence of the person's DNA be stated/documented.

Second, the contributor proposition should align with the scientific explanation of the evidence informed by any legitimate background information.

Third, reasonable alternate propositions consistent with non-contribution should be considered. For example, in a mixed DNA profile, it may be in the interests of the defense to include any person's DNA under both H_1 and H_2 as long as this inclusion is consistent with their own non-contribution

Table 1. Various options for the propositions H_1 and H_2. V is for victim, P is the person of interest and U is an unknown person.

H_1	H_2	
V+ P	V + U	Option 1
	U + U	Option 2
	P + U	Option 3
U + P	U + U	Option 4
U + V		Option 5

How to deal with multiple POIs

Consider a situation where there are two POIs termed P_1 and P_2. A crime stain is found, and the DNA mixture profile can be explained fully if P_1 and P_2 are the contributors (Table 2). For demonstration purposes a two-person mixture is assumed (but the three approaches described here can extend to higher order mixtures). U stands for an unknown individual, usually considered to be unrelated to either P_1 or P_2, although in STRmix™ this assumption may be relaxed to a relative in most, but not all, of the situations illustrated below. Generally, either

Approach 1 or 2 (Table 2) is acceptable. Approach 2 has slightly more power to distinguish contributors from non-contributors. It should be used when it aligns with the prosecution allegation.

Approach 3 runs the risk of a major contributor with a high *LR* (if analyzed separately) ‘carrying’ a non-contributor or a weak/trace contributor with a low *LR* (if analyzed separately) into the final high *LR* for P_1 + P_2 that could be misleading if reported. This approach should be predicated on separate tests for P_1 and P_2 which both return *LR*s >1. Approach 3 only should be used in the unlikely event that background information determines that the DNA must originate from both P_1 and P_2 or neither of them.

Table 2. Three approaches to assigning propositions when there are multiple POIs. Propositions are H_1 and H_2. P_1 and P_2 are POIs and U is an unknown person.

Approach 1			Approach 2			Approach 3		
H_1	H_2		H_1	H_2		H_1	H_2	
P_1 + U	U + U	The results are given in the report	P_1 + U	U + U	The results are in the notes and not the report. They are used in an exploratory manner to inform the inclusion of P_1 and P_2 separately before testing them both together.	P_1 + U	U + U	The results are in the notes and not the report. They are used in an exploratory manner to inform the inclusion of P_1 and P_2 separately before testing them both together.
P_2 + U			P_2 + U			P_2 + U		
P_1 + P_2		The result is in the notes and not the report. It is used to check that both P_1 and P_2 may be included together.	P_1 + P_2	P_1 + U	The results are given in the report	P_1 + P_2		The results are given in the report
				P_2 + U				

How to deal with evidential items not demonstrably associated (before DNA testing) with either the POI or the victim

As an example, consider a situation in which a two-person DNA mixture was recovered from somewhere not particularly closely associated with the victim, such as a stain on a bed sheet in a room at a house where a party occurred. The alleged victim, V, states that she was raped in this room by the accused. The stain can be explained as a mixture of the victim and the person of interest, P.

Initially five options for sets of propositions may be considered (Table 1). The contributor hypothesis may pose H_1 as V + P. The non-contributor hypothesis has the option of Option 1, 2, or 3 for H_2 (or any pair or all three of these). Note that Option 1 is almost always more favorable to the defendant (i.e., a lower *LR*) than Option 2.

Option 3 suggests the presence of P but not V. It may be difficult for the defense to motivate this option in the context of the case. This option requires that the DNA is from the person of interest and another individual. This proposition asks for a rejection of the victim's statement and an explanation of P's DNA in the very room where the rape is alleged to have occurred.

This leaves H_2 as either V + U or U + U (whichever is considered to be the most reasonable given the case's circumstances). Note that the option U + U will likely lead to higher *LR*s than the other options. Using the option 2 set, one could misleadingly produce a very high *LR* from a major aligned with V and a small scatter of trace alleles consistent with P. Since V + U is consistent with non-contribution and in the defendant's interest we suggest that it should be used.

This approach would lead us to suggest Option 1, in which we have assumed the presence of the complainant, V. This would seem to run contrary to the SWGDAM (31) suggestion that people only be used as conditioning genotypes if their DNA is reasonably expected. One could easily state that no reasonable expectation exists since the item is not strongly associated with either P or V. However the appearance of V under H_1 comes about because that is indeed the contributor allegation, and under H_2 because it is consistent with non-contribution and is in the defense's interests.

Options 4 or 5 may be used as exploratory investigations to check the inclusion of V and P separately before proceeding to the set actually used for the interpretation, which would be both V and P for H_1.

In very complex situations, a 'search strategy' (akin to proposition sets 4 and 5) may be the only recourse available to the scientist. We note that this approach is suitable in an investigative framework, and some check of the potential for co-contribution should be made. An approach has been proposed for these types of cases in Section 4 of Buckleton et al. (30).

Applicability of probabilistic genotyping to forensic DNA typing results and usage matters

Since the use of any tool or technique must be supported both scientifically and for admissibility, we provide some information about PG, and in particular STRmix™, to support its application for interpretation of DNA profiles derived from forensic evidence.

The general acceptance test

General acceptance of the method is one of the Daubert admissibility criteria (37) and is the primary criterion of the Frye standard (38). Most PG software applications are based on established mathematical principles. For example, the MCMC algorithm is not novel (Markov published the first of his papers on this topic in 1906). Other components of MCMC were developed in the middle of the twentieth century; "*Monte Carlo methods were born in Los Alamos, New Mexico during World War II, eventually resulting in the Metropolis algorithm in* *the early 1950s… MCMC was brought closer to statistical practicality by the work of Hastings in the 1970s.*" (39)

MCMC is a widely-used technique and is considered a mainstream statistical tool. It is used in real estate market prediction (40), earthquake and rock fracturing (41), electricity capacity modelling (42), weather prediction (43), betting (44), climate (45), computational biology (46), computational linguistics (47), genetics (48), engineering (49), physics (50), aeronautics (51), stock market prediction (52), and social science (53). The key papers describing the algorithms used within the MCMC are Metropolis et al. (54), with 37,506 cites in Google Scholar (as at May 27th, 2018), and Hastings (55), with 12,229 cites providing some measure of their widespread acceptance and use. Searching scientific literature for 'Markov chain Monte Carlo' returns more than 512,000 records.

There is substantial interest in PG as evidenced by the number of modern PG software programs that have been developed, or are being developed, by researchers with very strong mathematical or statistical backgrounds (11, 13, 19, 21, 56-61). These efforts and recommendations indicate strong support for the general acceptance of PG. While different PG software programs tend to differ in some details, there is a very substantial commonality of principle between these software tools (i.e., they all produce *LR*s and all model the probability or probability density of the profile given all plausible genotypes). The differences among programs do not indicate a lack of consensus on general acceptance of PG as an analytical/statistical method or of individual software programs. These efforts are strong support for the general acceptance of PG. Both SWGDAM (4) and ISFG (62, 63) give recommendations for validation for laboratories that adopt PG for mixture interpretation.

At preparation of this manuscript STRmix™ is in use in at least 46 laboratories worldwide as their predominant method for the interpretation of DNA profiles in forensic casework. The laboratories using STRmix™ reside in the USA (n=31), Australia (n=7), England (n=2), Scotland (n=1), Republic of Ireland (n=1), Canada (n=2), Finland (n=1), and New Zealand (n=1).

Peer review, independent testing, and further general acceptance

Peer review is another criterion of the Daubert standard that may be considered by the gate keeper. Oxford (https://en.oxforddictionaries.com/definition/peer_review) defines peer review as: "*Evaluation of scientific, academic, or professional work by others working in the same field*." The scientific concepts that underpin PG software are verified by publication in peer reviewed journals and wider comments published elsewhere (11, 13-19, 21, 22, 29, 32, 33, 35, 58, 60, 64-93).

Additional peer-review has been achieved for many PG solutions by those laboratories that performed internal validation studies (11, 80, 94). Such internal validation studies typically are not published, as journals tend not to find such studies novel. These data, however, are available for review, if desired by the courts, as part of the formal discovery process, and some are available online (95-98).

PG and the PCAST Report

Recently, the President's Council of Advisors on Science and Technology (PCAST) proffered criticisms about the foundational validity of some forensic disciplines (99). While the 2016 PCAST Report favored the use of PG for forensic DNA mixture interpretation, PCAST neither evaluated the current state-of-knowledge regarding PG at the time of its review nor up to disbandment of the Council in 2017. PCAST considered validity proven for the use of PG for up to three-person DNA mixtures where the minor contributor is greater than 20% of the mixture (amended to the POI being 20% in the Report addendum) and for two-person mixtures where the minor profile is greater than 10%. If taken literally, according to PCAST one cannot reliably interpret mixtures – to include minor as well as major contributors – where the minor contribution is below 10% (footnote 216 of the original report). This statement relative to the major contributor is obviously unfounded. It is likely that PCAST was referring to assessment of the minor and assumed the major could be analysed. PCAST also incorrectly perceived gaps in proof of validity with high contributor numbers and mixture contributions less than 20%.

The PCAST Report assessed proof of validity by empirical studies published in the peer reviewed literature as of 2016 and did not review the totality of available data even at that time. This partial assessment is unfortunate, as publication of all validation studies is difficult since many journals preclude or discourage publication of most internal validation studies and many laboratories do not see it as their role to publish. By taking this limited stand, the PCAST Committee members did not avail themselves of the totality of data that was accessible at the time of their review and subsequent report being issued.

Fortunately, validation data exist, peer-reviewed literature is available, and limitations of the applications are described. One example of such work with STRmix™ (80) is summarized in Table 3. This work covers 1 to 5 person mixtures at much greater ratios and lower templates than referred to in the PCAST Report.

Recently, the internal validation data from 31 laboratories using or validating STRmix™ was compiled and interpreted (hereafter "internal compilation study" (94)) specifically to address the points raised within the PCAST Report. This study concluded that this combined dataset "*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*." These efforts, representing a substantial resource commitment, are a collation of the validation studies from 31 laboratories and demonstrate that there is support for interpreting a minor contributor much less than 20%, and in fact down to 0% (present but not observed), of the total DNA present in the mixture. As the template tends towards 0 the *LR* tends to approximately 1.

Table 3. A summary of the tests undertaken in the internal validation studies by the FBI.

Number of contributors	Input DNA range (per contributor) ng	Contributor ratio range	Number interpreted	Number of true contributors tested	Number of false contributors tested
2	0.006 to 0.9	10:1 to 1:1	105	202	22,504
3	0.021 to 1.0	16:1:1 to 1:1:1	64	192	13,620
4	0.050 to 3.2	16:1:1:1 to 1:1:1:1	84	336	17,808
5	0.016 to 1.25	10:1:1:2:2 to 1:1:1:1:1	24	120	5,256
A selection of 172 one-, two- and three-person profiles were interpreted as originating from two, three and four individuals, respectively. The true contributors and 200 non-contributors were tested. These experiments entertain ratios below that espoused as a threshold by the PCAST Report.					

Disclosure of the algorithms

An issue that has arisen during court proceedings (or during discovery requests) is that there is a need to have access to the source code of PG software to ensure proper peer review of the validity/reliability of the software. Use of open source software has been advocated (for example by ISFG https://www.isfg.org/) because:

1. It allows all parties (including the defense) to have access to software, and
2. There is a possibility that review by third parties could improve the code.

Regarding the first point, freeware is accessible without restriction or cost, which may be desirable to some potential users. In contrast, commercially-available software comes at a cost but includes continued support and quality control measures. Users need to evaluate these aspects of open source and commercially-available software when deciding to implement PG software. Discovery regarding commercial software may involve the code or an executable program but often with cost recovery. STRmix™ comes with a user manual and data on validation associated with the current version. Moreover, STRmix™ code has been and remains available for court purposes under a non-disclosure agreement and supervision, but not to competing software developers. To date the STRmix™ code has been provided via this process thrice (100).

While code can be made available for legal proceedings, informed empirical testing, which is the basis of validation studies (both developmental and internal), is the best way to assess performance and critically evaluate the results produced by a software tool. Indeed, troubleshooting a PG tool by the developers is typically performed without reference to the source code. STRmix™ has a facility called 'extended output' that is available to users as a useful assessment and diagnostic tool. This function outputs the proposed genotypes and other variables and the probability density for each step in the MCMC process. One can then attempt to reproduce these values by independent calculation(s). The code is only accessed to rectify an error, for example identified during this extended output review, or to modify the program.

In response to the second point, one would expect that freeware, being more accessible, would enable substantial third party improvement. There are a few instances to date, however, where programming input to open source PG software has occurred from external parties (e.g., Lab Retriever programmer Kirk Lohmueller made suggestions to the LikeLTD programmers (14) (see pg. 27), whilst cloning parts of LikeLTD, and David Balding (personal communication) had input from students). All software needs funding for support and development, and non-commercial software may require public funding or donations (see https://scieg.org/support-our-work/).

STRmix™ has received input on miscode detection and suggestions for improvement from many sources, such as collaborators, based on empirical testing, or applied usage of the software and/or extended output functions.

Both commercial and open source software should come with additional support beyond an accompanying manual. STRmix™ requires substantial mandatory training provided by experts, as well as making available a user help desk. As one moves to more complex ways of interpreting DNA profiles, proper training is vital to reduce the problems of misuse that may occur. The risks associated with little, improper, or no training and quality control are serious and warrant substantial consideration. This concept was not lost on the Court in a Daubert challenge ruling, stating (101):

"*As the source code cannot be altered by anyone except the programmers, there is an additional layer of internal controls that govern the STRmix's operation.*"

Support throughout validation and implementation and participation in a software user group are also critically beneficial to the development and execution of reliable standard operating procedures.

Error rate

Error and error rate are general scientific issues but also arise in the forensic setting (error rate of the method is one of the Daubert admissibility criteria) (37). Determining the error rate is not always straightforward, largely because error has various meanings; in DNA interpretation, it is determined very much by the sample. With sound quality assurance practices and standard operating procedures derived from judicious validation studies, error may be addressed and reduced.

In the context of forensic cases, concerns surround false associations and false exclusions. A fair justice system would generally favor a false exclusion over a false association. Accordingly, overstating the strength of the evidence when a POI (or victim) cannot be excluded as a potential contributor should be avoided. In forensic science a tendency to understate the evidential weight is termed conservativeness. The STRmix™ software incorporates key features to drive the *LR* towards a conservative (lower) result.

Most discussions on error rate surround the concepts of a true state and a declared state. For example, one could input true donors and see if the software outputs a declaration of "true donor." The output of all PG software, however, is not a declaration of true or non-donors but a *LR*, which is a number on a continuous scale. This approach differs from a categorical declaration of inclusion or exclusion in a similar manner to how a probability would differ from a statement of certainty. The result of a true donor indicated as a non-donor could be termed something like evidence supporting H_2 and a false donor indicated as a true donor could be termed something like support for H_1.

Support for H_1 for a non-contributor is directly related to the strength (or quality) of the DNA profile (75, 102). Hence there will not be one rate of support for H_1 for all DNA profiles examined by PG but a different rate for each sample.

An important distinction to understand is that support for H_1 for a non-donor can be due to the non-donor sharing many alleles with the profile, which is different from an event of "software error." Studies on STRmix™ suggest that the software itself does not contribute to the support for H_1 (75) under this scenario beyond that expected by overlap of the alleles of the non-donor and the profile. Some of the best evidence for this comes from Turing's rule. Turing's rule translated to DNA states that the average *LR* for a large sample of non-donors should be 1. Trials with STRmix™ show the average *LR* to be approximately 1 or, when various levels of conservancy are included, less than 1 (75, 94, 102). Considerable research has been undertaken that allows informed statements to be made about the potential uncertainty associated with *LR*s (19, 35, 66, 68). It is very difficult for operator error of the software or false information about a known contributor to cause a false inclusion. There are methods, based on importance sampling (102), that are sufficiently fast and allow massive (up to 10^{30} or even greater) non-donor tests to be run during validation and if needed on a particular case basis. Importance sampling creates a biased sample by drawing from a distribution of importance, in the DNA example these are profiles likely to produce high *LR*s. Since the sample is biased it is necessary to readjust for the bias after simulation. In a large set of mixtures compiled from 31 laboratories (94), all large (over 10,000) *LR*s for non-donors were investigated; in all instances the non-donors had high allelic overlap with the profile. This is a correct result. This empirical assessment of non-donor inclusion rate provides additional support on the reliability of STRmix™. To date there have been no detected instances of a high *LR* using STRmix™ software where there were not also many alleles in common between the donor and the profile.

There are several situations where a false exclusion (*LR* <1 for a true donor) may occur:

1. The contribution of a true donor is low resulting in only a few alleles above the analytical threshold, or for other reasons the PCR does not generate an optimum profile (i.e., the contributor’s alleles are not detected or are poorly represented, or stutter heights appear disproportionately due to stochastic amplification),
2. Incorrect typing information for a true contributor(s) is used in the analysis (to include a sample mix-up) (this can be for the tested POI or, for a conditioned analysis, another contributor whose DNA is assumed to be present in a mixture),
3. An operator error, notably not removing an artifact before STRmix™ analysis, or
4. The number of contributors assigned is too few.

Diagnostics output by STRmix™ align with human judgment and thus allow for a human check of the results (a desirable and recommended feature). In some instances of false exclusions, the output data may indicate an *LR* <1 for a single locus, with high *LR*s at all other loci, suggesting a possible error in the input data. In such situations, or others in which the output from the software and the human operator disagree, the operator should review the input data to determine if they are correct (e.g., artifacts removed and alleles properly recorded). A retained artifact should be removed or a mislabeled allele corrected, and PG should be carried out again.

Although one might suggest that an error rate should be considered for the software – operator pair, there is in fact no reasonable way to calculate the operator error rate. As the process described here allows for correction, the reported result of such a software – operator pair is not an error for which an appropriate rate could be calculated. Any such corrections should be documented and, together with the output of the software and the original profile(s), are subject to technical review and, if desired, other independent review. Lastly, error rates, at best are some indirect indication of performance, but are not indicative or predictive of error in a specific case. The important question is whether an error occurred in the case, which can best be addressed by review or by re-calculation, in this situation, of the PG results (103).

Coding standards and miscodes

There are no coding standards designed specifically for forensic software. To date, the software has generally been evaluated in systems (i.e., testing the overall process for generating reliable results) and by empirical approaches. Neither the SWGDAM nor ISFG recommendations on validating PG software require accreditation by any software standardization organization, instead seeming to prefer the systems approach for validation. Coding standards in the wider industry, such as the Institute of Electrical and Electronics Engineers (IEEE), might provide guidance when testing the performance of software (104). Third party assessment, developmental validation studies, internal validation studies, and adherence to gathering community feedback tend to provide more scrutiny than only a single entity assessment.

As all software programs may have coding faults (miscodes), even with the most diligent scrutiny, developers and maintainers of a software must gather information by continuous testing and from users to identify miscodes. It is important to be transparent and disclose any miscodes discovered that affect the numerical result, so stakeholders are informed and can have confidence that key software is subjected to critical quality review and is continuously improving. Many probabilistic genotyping software such as LRmix Studio (105), Lab Retriever (106), STRmix™ (107), EuroForMix(23), and LikeLTD (14) have disclosed miscodes. The consequences of miscodes should also be investigated and disclosed.

Validation testing selects a broad range of samples but cannot test the myriad possible ways DNA profiles may present in real forensic casework (11, 108, 109). However, one can have confidence that a breadth of normal usage is well tested, as exemplified by Bright et al. with more than 2,825 mixtures of DNA from three to six contributors (94).

Precision of the output

Inherent in the validation of forensic methods is the assessment of variability. There has been some misuse of terms associated with precision. We reprise some definitions in Table 4, following those of (31, 110).

It is well-known that measurement error is associated with diagnostics, such as DNA typing. For example, the highly reliable generation of STR results using optimum input amounts of DNA is considered repeatable and reproducible. Yet if a sample were re-amplified or re-injected, no one would expect that precisely the same peak heights would be obtained and that the various peaks of a profile would be in the exact same relative proportions. Precision and accuracy are assessed within some range of measurement that is determined through validation studies. These studies are important for determining the limitations of a methodology, including usage of the STRmix™ software (4).

Table 4. Terms relating to variability

Repeatability	The degree, within measurement error, to which the same result(s) is obtained for a sample when the assay is repeated by the same operator and/or detection instrument.
Reproducibility	The degree, within measurement error, to which the same result(s) is obtained for a sample when the assay is repeated between/among different operators and/or detection instruments.
Precision	The degree of mutual agreement among a series of individual measurements, values and/or results. Precision depends only on the distribution of random errors and does not relate to the true value or specified value.
Accuracy	Degree of conformity of a measured quantity to its actual (true) value.
Objective	Little or no judgment required by the analyst
Subjective	Some judgment required by the analyst
Biased	A measurement is systematically above or below the true value.

STRmix™ uses a MCMC resampling method. The MCMC resampling strategy will create run-to-run variability. For STRmix™, the random number generator is run starting from a seed and, in the default setting, that seed is set from the clock. Thus, one should expect a degree of variation in the results. Validation should address the degree of variation as a basis for determining appropriate operating procedures and reporting criteria.

Recognition of variability is a positive aspect of PG, or for that matter any methodology measuring a target of interest. PG results are based on modeling which is the best attempt available at producing a rational and supportable answer that is based on solid mathematics and extensive empirical work. When judgment is exercised in relation to any modeling decision, we favor decisions that tend to reduce the value of the *LR*, an approach we call conservative. Indeed, introducing a continuous DNA interpretation system helped to better recognize the uncertainty inherent in assigning a *LR*.

Reliability of PG at low template

STRmix™ has now been extensively tested on profiles generated from optimum template levels down to extinction (see in particular (68, 74, 75, 80, 94)), as well as across a range of constructed mixture types as encountered in forensic casework with respect to total template amount (i.e., optimal to trace), contributor proportions (i.e., similar to disparate relative contributions within a given mixture) and other features such as allele sharing. The trend is that the *LR* tends towards 1 for both true donors and non-donors as peak heights of the contributor in question become lower. This finding is true whether the other contributors are also low or high in average peak height. Trials have been undertaken where the minor contributor is not observable (0%). In such cases STRmix™ reports a *LR* close to 1, usually between a log(*LR*) of -3 to 3. These results demonstrate that STRmix™ reliably reports that the profile is close to uninformative with respect to whether the POI, at zero template and hence not there, is a contributor or not.

Moretti et al. (80) reported a total of 277 two-, three-, four- and five person mixtures, prepared using DNA from thirteen contributors with varying individual template amounts (ranging 0.006-3.2 ng) and total template amounts (ranging 0.019-4 ng) tested using STRmix™ (Table 3). Ratios ranged from equal contributions (i.e. 1:1 to 1:1:1:1:1) to up to a maximum major contribution of approximately 95% (e.g., 20:1), as well as various intermediate proportions. This study also assessed the effect on the *LR* of assigning the number of contributors for STRmix™ analysis as one more than the target number of contributors. The word “target” is used herein to mean the number of contributors input into the mock sample. This usage is differentiated from the “true” number since one or more of the target contributors may be in such a low amount that it is not realistically present at all.

Bright et al. (94) reported a large compilation of the results from 31 different laboratories of internal validation studies of STRmix™. There were 2,825 mixtures generated using eight different STR multiplexes and analysed on two different types of CE instruments. These mixtures comprised three to six donors, with contributor templates down to extinction and a wide range of mixture proportions. The apparent number of contributors as interpreted by the respective laboratories was assigned as 3, 4 or 5 for PG analysis. Though some trace contributors were not observed in the electropherograms, the assigned *LR*s were appropriate based on the data present for each contributor (i.e., tending to 1 with less information present). Bright et al. also observed lower *LR*s for true contributors and more *LR*s near 1 for non-donors when the number of assigned contributors increased. This finding using STRmix™ also was reported by Taylor (68) and Moretti et al. (80) and demonstrated that mixed DNA profiles containing more contributors are reliably reported as less informative.

These studies also support that the average peak height (APH) of the contributor is a good indicator of information content. It is the low peak heights rather than the extreme ratios that lead to uninformative profiles. For example, testing two low-level contributors with similar APHs (a 1:1 mixture) presents more of a challenge to the software than does a 1:20 mixture, since the genotype of the higher contributor has less uncertainty and helps to inform the genotype of the lower contributor. The PCAST position placed significance solely on contributor ratio, ignoring the important component of template amount. Empirical testing (80, 94) demonstrates that the positions stated in the PCAST Report are unsupported and the use of complete data should be considered when evaluating performance of PG software. Typical results are shown in Figure 1. The *LR* tends to approximately, but not exactly, 1 for both true and non-donors as the template is reduced. This is the correct result.

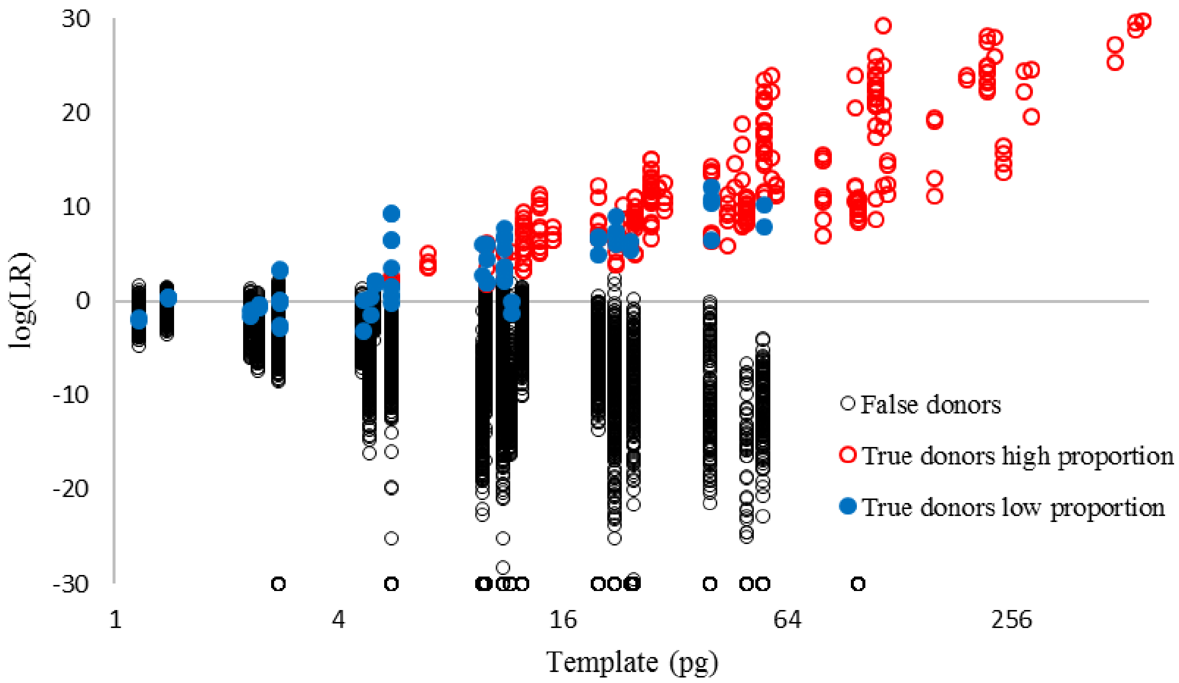


Figure 1. A plot of $\log_{10}(LR)$ vs template (pg) for each donor in a four-person mixture, prepared across a range of template amounts and contributor ratios, tested using STRmix™. For the non-donor tests, the template is assigned as the lowest template of the four true donors. For those samples with template above 1 pg, 194 non-donors were tested against the profile. We have also added a fictional contributor, effectively at template 0pg and tested against 100,000 non-donors. Due to plotting limitations, these samples are represented in this plot at template 0.5 pg. For the true donor tests the data have been divided into proportion above 10% (high N = 275) and those with proportion below 10% and down to 0% (low N = 72). As the template diminishes, the *LR*s for both the true and non-contributors tend towards 1 (a log(*LR*) of zero is marked with a central horizontal line in the graph).

Number of contributors (NoC)

The number of contributors to a DNA mixture profile from a casework sample is typically unknown and is correctly described as a nuisance variable. A nuisance variable is something needed to do the computation but not available directly from the data. However, depending on the typing results, assigning a NoC to a DNA mixture can range from fairly straightforward to particularly challenging. When the assignment of NoC is more challenging there are approaches (described herein) available that can provide results that tend to understate the strength of the evidence.

There are two general *LR* approaches in use, often driven by the capabilities of the software.

These are:

1. Assigning a NoC that is the same under both H_1 and H_2 (termed constrained NoC).
2. Allowing the NoC under each proposition to differ (termed unconstrained NoC).

Constrained NoC

Under this process a NoC is assigned to the profile based on the number of allelic peaks and their heights, often after consideration of artifacts. While most constrained NoC determinations are performed manually, there are software tools available to assist if needed, such as NOCIt (111) based on Monte Carlo methods, PACE (112) based on machine learning, and methods using Bayesian networks (113) and maximum likelihood (114).

Herein only assignment of NoC by a human operator is discussed (initially drop-in is not considered to simplify the discussion). The DNA profile should be examined and any obvious artifacts such as spikes and pull-up discounted. Peaks in back or forward stutter positions below some preset value derived from internal validation may be considered stutter, or stutter and allelic. If considered solely stutter, such peaks can be discounted in determining NoC. The peaks that remain should be considered potentially allelic. The minimum NoC may be determined based on peak count alone. It is important, then, to consider whether the observed peak heights of the alleles can be supported by this preliminary assignment. If peak imbalances are unrealistic with the preliminary NoC, then at least one additional contributor should be added. After this initial review of the evidentiary DNA profile it is desirable to determine which, if any, contributors should be expected under both H_1 and H_2. This assessment may include, for example, the victim and a consensual partner. The profile of the POI should not be examined at this stage.

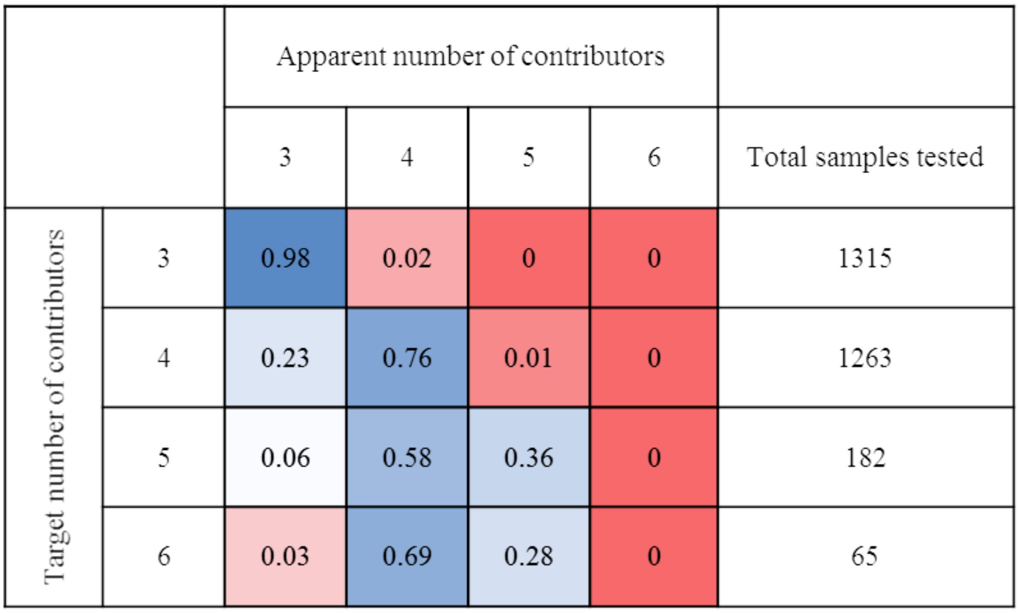
Depending on the laboratory's internal validation studies assessing detection sensitivity and allele drop-in, few trace peaks may be discounted for the NoC assignment as potentially being attributed to drop-in. This is usually fewer than a preset number (often only up to two are permitted) and lower than a height established from empirical studies. Although discounted when assigning NoC, these peaks must not be discounted from the analysis of evidential value.

With this general process a reasonable NoC can be assigned, but there is no guarantee that this estimate is the actual NoC of the sample. Confidence in the assignment varies depending on the complexity of the mixture. Fortunately, any reasonable discrepancy in NoC assignment seems to have a minor effect on the deconvolution or *LR* (35, 66, 80, 94).

Even for controlled studies, such as those performed during validation, the actual NoC used to generate a mixture (target *N*) may not be the same as the number *observed* in the mixture (*N*). Simply stated: low level contributors may be too low to observe, because a minimum amount of template is needed for any DNA sample to be detectable. Hence, even in mock samples, the number of contributors may not be accurately represented (note that these issues relate to the minor and trace contributors, as major contributors tend to be well represented in mock samples).

The internal validation compilation study (94) described this effect of estimating various NoC to a mixture. Figure 2 (derived from Figure 13 of Bright et al. (94)) shows the level of over and under-estimation of the apparent NoC (*N*) determined following the individual laboratories' protocols compared to the target *N* in their respective studies. Overestimation of *N* generally led to similar or lower *LR*s for true contributors. Underestimation of *N* resulted in exclusions of true contributors, usually affecting the lower/lowest quantity contributor(s).

These data support the view that when assigning *N*, for false contributors, the risk is overestimation of *N*, as there is an increase in the number of very low grade adventitious hits. With respect to the *LR* for true contributors, when *N* is either under or overestimated, the result is conservative. Hence, if the *LR* is large, for example larger than 1000 and there is uncertainty in *N*, there is confidence in the *LR* if *N* is correct and if not correct, the *LR* is more conservative than the already built in buffers (such as the use of a conservative population genetic model (115) and reporting a lower bound on the *LR* (116)).



		Apparent number of contributors				
		3	4	5	6	Total samples tested
Target number of contributors	3	0.98	0.02	0	0	1315
	4	0.23	0.76	0.01	0	1263
	5	0.06	0.58	0.36	0	182
	6	0.03	0.69	0.28	0	65

Figure 2. Heat map of the fraction of prepared DNA mixtures as interpreted with various differences between the apparent number of contributors (NoC) and the target number (target NoC). The higher numbers are blue and the lower numbers red.

Moretti et al. (80) reported the effect on the *LR* of assuming an incorrect *N* by both increasing (*N*+1) and decreasing (*N*-1) from the most plausible 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 for both true contributors and 200 non-contributors, which then were converted to lower bounds on the *LR*. For true contributors (H_1-true), the majority of lower bounds on the *LR*s under the assumptions of *N* and *N*+1 contributors were similar (within one order of magnitude); for 13% of the analyses, the lower bound on the *LR* decreased by more than one order of magnitude. With regard to non-contributors under the incorrect assumption of an additional contributor, fewer were excluded outright, though overall 94.3% returned lower bounds on the *LR* <1.

As a means of examining the impact of assuming too few contributors without returning an exclusion outright, Moretti et al. (84) artificially created three mixtures from a two-person mixture (1:5 contributor ratio) by adding a “third” contributor in the range 50-200 rfu, constructed as if it was a child of the two true 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). Note that a *LR*=0 is a practical rounding off, as in theory a *LR* should not be assigned a value of 0 (80).

Management of the uncertainty of NoC in real casework (usually for complex profiles or low level contributors) is easily achieved by testing plausible values for NoC. All outcomes of plausible analyses should be retained, and one or a few may be reported.

It may be tempting to revisit the NoC after examination of the POI’s profile. For example, it may be possible to sustain the inclusion of the POI by adding a contributor. This approach cannot be entertained with software that uses the constrained NoC approach.

Unconstrained NoC

There is no requirement for the contributor (H_1) and non-contributor (H_2) hypotheses to specify the same number of contributors when calculating *LR*s. Some PG, for example LRmix, LikeLTD, and Lab Retriever, can perform calculations with different NoC for each proposition.

Under this approach when considering NoC under H_1 whether by human judgment or software, the genotype of the POI may be considered. This may lead to the situation where NOC under H_1 is one larger than under H_2 in order to accommodate the POI. We are unaware of any publications addressing the likely effect of this approach.

Software implementing the unconstrained NoC approach can also test an increase in NoC under H_2 while leaving NoC under H_1 at the assigned value. Again we are unaware of a publication outlining the effect of this.

Recently Slooten and Caliebe (117) published a result that is likely to advance this discussion. If we consider a constrained NoC then there will be different *LR*s for each value of the NoC (termed LR_n where n is the NoC). We want the overall *LR* which we define as the *LR* where the number of contributors is not known and is treated, correctly, as a nuisance variable. Slooten and Caliebe show that the overall *LR* is the weighted average of the LR_n values under one reasonable assumption that we discuss later. This is a useful finding.

In their simplest solution (they offer several) the weights for the weighted average are $\Pr(N = n|G_C, G_P, H_2)$ where G_{C} is the profile of the crime stain and G_P is the profile of the POI. Since we consider H_2, G_P can be removed from the conditioning yielding $\Pr(N = n|G_C, H_2)$. It is likely that only a few values of *n* need to be considered, maybe often only one or two.

The assumption that leads to this result is that *n* is equally likely under H_1 and H_2, specifically $\Pr(N = n|H_1) = \Pr(N = n|H_2)$. Note that the conditioning does not contain G_{c} or G_{s}, and hence is informed only by whether or not the POI is a donor. This assumption is likely to be true or approximately true in the vast majority of cases.

Subjectivity

Some have suggested that subjectivity equates with bias with attendant negative outcomes. Instead, subjectivity does not automatically imply a bias that will result in an error, and objectivity does not automatically imply an absence of bias that will render an interpretation free from error. The forensic community has begun to appreciate the risks of contextual and conformational bias and is attempting to address these risks and their potential negative effects in a number of ways (see (118) for a discussion). As an example, the use of suspect-driven bias is clearly an unacceptable practice for deciding which loci in a mixture profile may exhibit allele dropout (2, 119).

However, the stark view that subjectivity equates to committing error belies the current thinking about cognitive bias. As Jeanguenat et al. (118) recently reminded readers:

“*Cognitive contamination or bias is inherent in all human beings due to the architecture and operation of the brain. However, it is important to understand that although bias exists it does not always result in an incorrect interpretation, just as enacting bias reduction steps will not guarantee that laboratory results will be error free. Nevertheless, forensic scientists should continue to improve and seek mechanisms to minimize error due to bias*.”

The idea that subjectivity will inevitably lead to unfair outcomes is incorrect. Indeed, there are methods that have been invoked with intentional bias. For example, some practitioners have generally set stochastic thresholds (STs) rather high to reduce the chances of declaring a false match to a reference sample, at the expense of producing more inconclusive results. SWGDAM defines the ST as “the value above which it is reasonable to assume that allelic dropout has not occurred within a single-source sample” (120). While such a practice does not make use of some potentially useful data, it is "biased" in order to avoid a more egregious error (i.e., a false inclusion).

The best approach to control the potential negative effects of bias is by providing proper training and education to DNA analysts using these software tools. All humans are susceptible to various biases, and forensic scientists are no exception (118). The more cognizant individuals are to the risks of bias, the better they will be to develop strategies and procedures to minimize their effects. Thinking that one can overcome bias by force of will is a dangerous misconception.

Many PG programs are designed intentionally to yield a lower bound on the *LR* that factors in several elements of uncertainty and thereby “biases” the analysis toward conservatism (i.e., reduced statistical weight). For example, in STRmix™, the highest posterior density (HPD)

LR, which if enabled, accounts for the expected amount of run-to-run variation from the Monte Carlo effect and variation in allele probabilities (121). In addition, most PG programs use the population genetic model of Balding and Nichols (122) or close variations. This model has been shown to result in *LR*s that are conservative (115, 123-125)) and is used in TrueAllele®, LRmix, STRmix™, EuroForMix, and LikeLTD. A close approximation is used in Lab Retriever. Another key parameter in the population genetic model is the coancestry coefficient *θ*, which is typically set in a way that tends toward lowering the *LR* result. It is either set towards the upper end of the plausible range, or a distribution is used based on a diverse set of populations (33). LRmix and LikeLTD utilize the size bias correction described by Balding (126) which produces conservative assignments on average. Balding (127) described a way to combine the contribution of various relatives and unrelated people, which is implemented in STRmix™ (71) and can be set to give an allowance regarding relatedness. Together or separately, all these features are biased towards cautious statements of evidential weight in relation to the contributor proposition. Software has been proposed as a way to eliminate bias by promoting that it is completely objective (109). This assertion is problematic in that it does not appreciate the effects of bias and dissuades one from embracing the need to consider bias. No one is impervious to bias including those that develop software. While software allows for better repeatability and reproducibility, one should be cognizant that those who develop the software have inserted their own ideas about how the program works best, such as what aspects to weight more so than others. Users of PG software should be informed about the limitations of the software and potential pitfalls and not rely exclusively on software output. Users must apply their training and expertise in DNA profile interpretation and evaluate the software output by visually comparing the PG results with the original DNA profile. For most profiles, the PG results can be assessed to be intuitively correct, or not, by a properly trained DNA analyst. If one were to rely solely and blindly on the PG output, error could occur. For example, a false exclusion of a POI due to incomplete resolution of a TH01 minor 10 allele from the major contributor's 9.3 allele was shown by Moretti et al. (80) This "false exclusion" was a limitation of CE instrument resolution and not due to the PG software. However, Moretti et al. (80) overcame the limitation with an accompanying manual assessment.

To summarize the issues of bias:

1. Bias is an inherent characteristic of human beings.
2. Training about cognitive bias is an important aspect of good science, and particularly so for forensic science.
3. Software can have inherent bias, some of which is desirable.
4. Some PG developers and many users are trained on aspects where bias can impact the decision process.
5. Additional review beyond relying solely on software output is recommended as an additional layer to reduce the effects of bias.

PG software removes some decisions or gives substantial support to these decisions. Remaining aspects of subjectivity in STRmix™ include some artifact management of spikes and pull-up at the initial analysis of the electropherogram and an assignment of an exact or approximate number of contributors.

Recently an inter-laboratory study was published (128) that reports the results of 15 different laboratories using LRmix on the same profile. This profile had been constructed from donors with features deliberately chosen to make interpretation difficult (personal communication, MC Márquez). The *LR*s reported vary between 2.6 × 10^3 and 3.2 × 10^{14}. This range drew attention in court and suggested some fault in PG. However the paper clearly describes that the variation arises not from the software but from subjective decisions regarding allele and stutter determinations (see Figure 1 from (128)). This finding supports the use of models for stutter as in STRmix™ rather than relying solely on subjective human decisions and may also indicate a need for training that should accompany use of software.

Conclusion

PG had been in gestation for some time from before 2000 (5). The first forensic DNA case of which we are aware that utilized PG methods occurred with TrueAllele® in 2009. Large scale deployment of PG software, and STRmix™ in particular, to forensic laboratories began in 2012. The efforts to bring PG to fruition, including the initial theoretical development for human identification applications based on STR typing (5-8), span almost two decades, and thus its use today should not be misconstrued as some sudden novel technology. To the contrary, with the maturation of STR typing technologies, great strides in the application of probabilistic solutions to biological phenomena, and the development of several software options, the "coming of age" of PG has been recognized by forensic laboratories. Facilitated by guidance documents from SWGDAM for validating PG systems in 2015 (4), followed by the European Network of Forensic Science Institutes (ENFSI) in 2017 (129), empirical studies performed by many have demonstrated the utility and reliability of PG for mixture analysis and enabled the implementation of reliable procedures for the application of this technology to forensic casework.

Our goal in this paper is to provide and address information and issues that relate to PG in general and STRmix™ in particular through our own experience. Our intent is to be informative. By understanding the strengths and limitations of any PG software, users and stakeholders will better understand the system and hopefully use it in a thoughtful manner for the public good.

Funding

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 the U.S. Department of Justice. Names of commercial manufacturers are provided for identification purposes only, and inclusion does not imply endorsement of the manufacturer, or its products or services, by the FBI. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the U.S. Government.

References

1. Budowle B, Onorato AJ, Callagham TF, Manna AD, Gross AM, Guerreri RA, et al. 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.

2. Bieber FR, Buckleton JS, Budowle B, Butler JM, Coble MD. Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion. BMC Genet. 2016;17(1):125.

3. Bille T, Bright J-A, Buckleton J. Application of Random Match Probability Calculations to Mixed STR Profiles. J Forensic Sci. 2013;58(2):474-85.

4. Scientific Working Group on DNA Analysis Methods (SWGDAM). Guidelines for the Validation of Probabilistic Genotyping Systems. 2015 [updated 2015; cited 3 October 2016]; Available from: http://media.wix.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf.

5. Gill P, Whitaker JP, Flaxman C, Brown N, Buckleton JS. An investigation of the rigor of interpretation rules for STR's derived from less than 100 pg of DNA. Forensic Sci Int. 2000;112(1):17-40.

6. Evett IW, Buffery C, Willott G, Stoney D. A guide to interpreting single locus profiles of DNA mixtures in forensic cases. J Forensic Sci Soc. 1991;31(1):41-7.

7. Evett IW, Gill PD, Lambert JA. Taking account of peak areas when interpreting mixed DNA profiles. J Forensic Sci. 1998;43(1):62-9.

8. Pálsson B, Pálsson F, Perlin M, Gudbjartsson H, Stefánsson K, Gulcher J. Using quality measures to facilitate allele calling in high-throughput genotyping. Genome Res. 1999;9(10):1002-12.

9. Perlin MW, Burks MB, Hoop RC, Hoffman EP. Toward fully automated genotyping: allele assignment, pedigree construction, phase determination, and recombination detection in Duchenne muscular dystrophy. Am J Hum Genet. 1994;55(4):777-87.

10. Gill P, Kirkham A, Curran J. LoComatioN: A software tool for the analysis of low copy number DNA profiles. Forensic Sci Int. 2007;166(2-3):128-38.

11. Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, et al. Validating TrueAllele® DNA mixture interpretation. J Forensic Sci. 2011;56:1430-47.

12. Slooten K. Accurate assessment of the weight of evidence for DNA mixtures by integrating the likelihood ratio. Forensic Sci Int Genet. 2017;27:1-16.

13. Prieto L, Haned H, Mosquera A, Crespillo M, Alemañ M, Aler M, et al. Euroforgen-NoE collaborative exercise on LRmix to demonstrate standardization of the interpretation of complex DNA profiles. Forensic Sci Int Genet. 2014;9:47-54.

14. Balding DJ, Steele CD. likeLTD v6.0: an illustrative analysis, explanation of the model, results of validation tests and version history. 2015 [updated 2015; cited 16 January 2018]; Available from: https://sites.google.com/site/baldingstatisticalgenetics/software/likeltd-r-forensic-dna-r-code.

15. Inman K, Rudin N, Cheng K, Robinson C, Kirschner A, Inman-Semerau L, et al. Lab Retriever: a software tool for calculating likelihood ratios incorporating a probability of drop-out for forensic DNA profiles. BMC Bioinformatics. 2015;16(1):298.

16. Buckleton J, Kelly H, Jo-Anne Bright, Taylor D, Tvedebrink T, Curran JM. Utilising allelic dropout probabilities estimated by logistic regression in casework. Forensic Sci Int Genet. 2014;9:9-11.

17. Balding DJ, Buckleton J. Interpreting low template DNA profiles. Forensic Sci Int Genet. 2009;4(1):1-10.

18. Götz FM, Schönborn H, Borsdorf V, Pflugbeil A-M, Labudde D. GenoProof Mixture 3—New software and process to resolve complex DNA mixtures. Forensic Sci Int Genet Suppl Ser. 2017;6:e549-e51.

19. Taylor D, Bright J-A, Buckleton J. The interpretation of single source and mixed DNA profiles. Forensic Sci Int Genet. 2013;7(5):516-28.

20. Manabe S, Morimoto C, Hamano Y, Fujimoto S, Tamaki K. Development and validation of open-source software for DNA mixture interpretation based on a quantitative continuous model. PLoS ONE. 2017;12(11):e0188183.

21. Cowell RG, Lauritzen SL, Mortera J. Probabilistic expert systems for handling artifacts in complex DNA mixtures. Forensic Sci Int Genet. 2011;5(3):202-9.

22. Cowell R, Graversen T, Lauritzen S, Mortera J. Analysis of forensic DNA mixtures with artefacts. J R Stat Soc Ser C Appl Stat. 2015;64(1):1-48.

23. Bleka Ø, Storvik G, Gill P. EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts. Forensic Sci Int Genet. 2016;21:35-44.

24. Evett IW. What is the Probability that This Blood Came from that Person? A Meaningful Question. J Forensic Sci Soc. 1983;23:35-9.

25. Evett IW, Weir BS. Interpreting DNA Evidence – Statistical Genetics for Forensic Scientists. Sunderland: Sinauer Associates, Inc., 1998.

26. Marsden CD, Rudin N, Inman K, Lohmueller KE. An assessment of the information content of likelihood ratios derived from complex mixtures. Forensic Sci Int Genet. 2016;22:64-72.

27. Cook R, Evett IW, Jackson G, Jones PJ, Lambert JA. A hierarchy of propositions: Deciding which level to address in casework. Sci Justice. 1998;38(4):231-40.

28. Evett IW, Jackson G, Lambert JA. More on the hierarchy of propositions: exploring the distinction between explanations and propositions. Sci Justice. 2000;40(1):3 - 10.

29. Gittelson S, Kalafut T, Myers S, Taylor D, Hicks T, Taroni F, et al. A Practical Guide for the Formulation of Propositions in the Bayesian Approach to DNA Evidence Interpretation in an Adversarial Environment. J Forensic Sci. 2016;61(1):186-95.

30. Buckleton J, Bright J-A, Taylor D, Evett I, Hicks T, Jackson G, et al. Helping formulate propositions in forensic DNA analysis. Sci Justice. 2014;54(4):258-61.

31. Scientific Working Group on DNA Analysis Methods (SWGDAM). Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories. 2017 [updated 2017; cited 17 March 2017]; Available from: https://media.wix.com/ugd/4344b0_2a08f65be531488caa8037ed55baf23d.pdf.

32. Buckleton J, Bright JA, Taylor D. Forensic DNA evidence interpretation. 2nd ed. Florida, USA: CRC Press, 2016.

33. Bright J-A, Taylor D, McGovern CE, Cooper S, Russell L, Abarno D, et al. Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles. Forensic Sci Int Genet. 2016;23:226-39.

34. Coble MD, Bright J-A, Buckleton JS, Curran JM. Uncertainty in the number of contributors in the proposed new CODIS set. Forensic Sci Int Genet. 2015;19:207-11.

35. Bright J-A, Curran JM, Buckleton JS. The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation. Forensic Sci Int Genet. 2014;12:208-14.

36. Curran JM, Buckleton J. Uncertainty in the number of contributors for the European Standard Set of loci. Forensic Sci Int Genet. 2014;11(1):205-6.

37. Daubert et al. v Merrell Dow Pharmaceuticals Inc., 509 US 579 (1993). 1993.

38. Frye v The United States of America, 54 AppDC 46, 293Fed 1013 (1923). 1923.

39. Robert C, Casella G. A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. Statistical Science. 2011;26(1):102-15.

40. Gargallo P, Miguel JA, Salvador MJ. MCMC Bayesian spatial filtering for hedonic models in real estate markets. Spat Stat. 2017;22:47-67.

41. Pan X, Zhang G, Chen H, Yin X. McMC-based AVAZ direct inversion for fracture weaknesses. J Appl Geophy. 2017;138:50-61.

42. Almutairi A, Hassan Ahmed M, Salama MMA. Use of MCMC to incorporate a wind power model for the evaluation of generating capacity adequacy. Electr Pow Syst Res. 2016;133:63-70.

43. Berrocal VJ, Raftery AE, Gneitting T, Steed RC. Probabilistic weather forecasting for winter road maintenance. J Am Stat Assoc. 2010;105(490):522-37.

44. Crowder M, Dixon M, Ledford A, Robinson M. Dynamic modelling and prediction of English Football League matches for betting. Statistician. 2002;51(2):157-68.

45. Wang G, Chen S. Evaluation of a soil greenhouse gas emission model based on Bayesian inference and MCMC: Model uncertainty. Ecol Modell. 2013;253:97-106.

46. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003;19(12):1572–4.

47. Johnson M, Griffiths T, Goldwater S. Bayesian Inference for PCFGs via Markov Chain Monte Carlo. Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference: Association for Computational Linguistics; 2007;139-46.

48. Bieber FR, Brenner CH, Lazer D. Finding criminals through DNA of their relatives. Science. 2006;312:1315-6.

49. Beck JL, Au S-K. Bayesian Updating of Structural Models and Reliability using Markov Chain Monte Carlo Simulation. J Eng Mech. 2002;128(4).

50. Vinay K, Jeremy JD. Markov-Chain Monte Carlo Reconstruction of Emission Measure Distributions: Application to Solar Extreme-Ultraviolet Spectra. Astrophys J. 1998;503(1):450.

51. Yang P, Bifeng S, Qing H, Baiyu O. A Direct Simulation Method for Calculating Multiple-hit Vulnerability of Aircraft with Overlapping Components. Chin J Aeronaut. 2009;22(6):612-9.

52. Rey C, Rey S, Viala JR. Detection of high and low states in stock market returns with MCMC method in a Markov switching model. Econ Model. 2014;41:145-55.

53. Jackman S. Estimation and Inference Are Missing Data Problems: Unifying Social Science Statistics via Bayesian Simulation. Polit Anal. 2000;8(4):307-32.

54. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of state calculations by fast computing machines. J Chem Phys. 1953;21:1087-91.

55. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970;57:97--109.

56. Puch-Solis R, Clayton T. Evidential evaluation of DNA profiles using a discrete statistical model implemented in the DNA LiRa software. Forensic Sci Int Genet. 2014;11:220-8.

57. Puch-Solis R, Rodgers L, Mazumder A, Pope S, Evett I, Curran J, et al. Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters. Forensic Sci Int Genet. 2013;7(5):555-63.

58. Cowell RG, Lauritzen SL, Mortera J. Probabilistic modelling for DNA mixture analysis. Forensic Sci Int Genet Suppl Ser. 2008;1(1):640-2.

59. Gill P, Haned H. A new methodological framework to interpret complex DNA profiles using likelihood ratios. Forensic Sci Int Genet. 2013;7(2):251-63.

60. Lohmueller K, Rudin N. Calculating the weight of evidence in low-template forensic DNA casework. J Forensic Sci. 2013;58(1):234-59.

61. Mitchell AA, Tamariz J, O'Connell K, Ducasse N, Budimlija Z, Prinz M, et al. Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in. Forensic Sci Int Genet. 2012;6(6):749-61.

62. Coble MD, Buckleton J, Butler JM, Egeland T, Fimmers R, Gill P, 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. 2016;25:191-7.

63. Gill P, Gusmão L, Haned H, Mayr WR, Morling N, Parson W, et al. 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 Sci Int Genet. 2012;6(6):679-88.

64. Bright J-A, Evett IW, Taylor D, Curran JM, Buckleton J. A series of recommended tests when validating probabilistic DNA profile interpretation software. Forensic Sci Int Genet. 2015;14:125-31.

65. Bright J-A, Stevenson KE, Curran JM, Buckleton JS. The variability in likelihood ratios due to different mechanisms. Forensic Sci Int Genet. 2015;14:187-90.

66. Bright J-A, Taylor D, Curran J, Buckleton J. Searching mixed DNA profiles directly against profile databases. Forensic Sci Int Genet. 2014;9:102-10.

67. Kelly H, Bright J-A, Buckleton JS, Curran JM. A comparison of statistical models for the analysis of complex forensic DNA profiles. Sci Justice. 2014;54(1):66-70.

68. Taylor D. Using continuous DNA interpretation methods to revisit likelihood ratio behaviour. Forensic Sci Int Genet. 2014;11:144-53.

69. Taylor D, Bright J-A, Buckleton J, Curran J. An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations. Forensic Sci Int Genet. 2014;11:56-63.

70. Taylor D, Bright J-A, Buckleton J. The ‘factor of two’ issue in mixed DNA profiles. J Theor Biol. 2014;363:300-6.

71. Taylor D, Bright J-A, Buckleton J. Considering relatives when assessing the evidential strength of mixed DNA profiles. Forensic Sci Int Genet. 2014;13:259-63.

72. Taylor D, Bright J-A, Buckleton J. Interpreting forensic DNA profiling evidence without specifying the number of contributors. Forensic Sci Int Genet. 2014;13:269-80.

73. Taylor D, Bright J-A, McGovern C, Hefford C, Kalafut T, Buckleton J. Validating multiplexes for use in conjunction with modern interpretation strategies. Forensic Sci Int Genet. 2016;20:6-19.

74. Taylor D, Buckleton J. Do low template DNA profiles have useful quantitative data? Forensic Sci Int Genet. 2015;16:13-6.

75. Taylor D, Buckleton J, Evett I. Testing likelihood ratios produced from complex DNA profiles. Forensic Sci Int Genet. 2015;16:165-71.

76. Taylor DA, Bright JA, Buckleton J. Commentary: A "source" of error: Computer code, criminal defendants, and the constitution. Front Genet. 2017;8:33.

77. Cooper S, McGovern C, Bright JA, Taylor D, Buckleton J. Investigating a common approach to DNA profile interpretation using probabilistic software. Forensic Sci Int Genet. 2015;16:121-31.

78. Bright J-A, Taylor D, Gittelson S, Buckleton J. The paradigm shift in DNA profile interpretation. Forensic Sci Int Genet. 2017;31:e24-e32.

79. Taylor D, Bright J-A, Kelly H, Lin M-H, Buckleton J. A fully continuous system of DNA profile evidence evaluation that can utilise STR profile data produced under different conditions within a single analysis. Forensic Sci Int Genet.31:149-54.

80. Moretti TR, Just RS, Kehl SC, Willis LE, Buckleton JS, Bright J-A, et al. Internal validation of STRmix for the interpretation of single source and mixed DNA profiles. Forensic Sci Int Genet. 2017;29:126-44.

81. Taylor D, Buckleton J, Bright J-A. Factors affecting peak height variability for short tandem repeat data. Forensic Sci Int Genet. 2016;21:126-33.

82. Bright J-A, Stevenson KE, Coble MD, Hill CR, Curran JM, Buckleton JS. Characterising the STR locus D6S1043 and examination of its effect on stutter rates. Forensic Sci Int Genet. 2014;8(1):20-3.

83. Bright J-A, Neville S, Curran JM, Buckleton JS. Variability of mixed DNA profiles separated on a 3130 and 3500 capillary electrophoresis instrument. Aust J Forensic Sci. 2014;46(3):304-12.

84. Bille TW, Weitz SM, Coble MD, Buckleton J, Bright J-A. Comparison of the performance of different models for the interpretation of low level mixed DNA profiles. ELECTROPHORESIS. 2014;35(21-22):3125-33.

85. Bright J-A, Curran J, Buckleton J. Modelling PowerPlex® Y stutter and artefacts. Forensic Sci Int Genet. 2014;11:126-36.

86. Bright J-A, Taylor D, Curran JM, Buckleton JS. Developing allelic and stutter peak height models for a continuous method of DNA interpretation. Forensic Sci Int Genet. 2013;7(2):296-304.

87. Bright J-A, Taylor D, J.M. C, Buckleton JS. Degradation of forensic DNA profiles. Aust J Forensic Sci. 2013;45(4):445-9.

88. Bright J-A, McManus K, Harbison S, Gill P, Buckleton J. A comparison of stochastic variation in mixed and unmixed casework and synthetic samples. Forensic Sci Int Genet. 2012;6(2):180-4.

89. Brookes C, Bright J-A, Harbison S, Buckleton J. Characterising stutter in forensic STR multiplexes. Forensic Sci Int Genet. 2012;6(1):58-63.

90. Perlin MW, Sinelnikov A. An information gap in DNA evidence interpretation. PLoS One. 2009;4(12):e8327.

91. Mortera J, Dawid AP, Lauritzen SL. Probabilistic expert system for DNA mixture profiling. Theor Popul Biol. 2003;63:191-205.

92. Mortera J. Analysis of DNA mixtures using Bayesian networks. In: Green P, Hjort NL, Richardson S, editors. Highly Structured Stochastic Systems. Oxford: Oxford University Press; 2002.

93. Kelly H, Bright J-A, Kruijver M, Cooper S, Taylor D, Duke K, et al. A sensitivity analysis to determine the robustness of STRmix™ with respect to laboratory calibration. Forensic Sci Int Genet. 2018;35:113-22.

94. Bright J-A, Richards R, Kruijver M, Kelly H, McGovern C, Magee A, et al. Internal validation of STRmix™ – A multi laboratory response to PCAST. Forensic Sci Int Genet. 2018;34:11-24.

95. The New York City Office of Chief Medical Examiner. Internal Validation of STRmix™ V2.4 for Fusion NYC OCME. 2016 [updated 2016; cited 22 April 2017]; Available from: http://www1.nyc.gov/assets/ocme/downloads/pdf/STRmix-V2-4-Fusion-5C-Validation%20Summary.pdf.

96. District of Columbia Department of Forensic Science Laboratory. Part II: Internal Validation of STRmix™ V2.4 Using the GlobalFiler™ PCR Amplification Kit and 3500/3500XL Genetic Analyzer. February 24, 2017. 2017 [updated 2017; cited 19 January

2018]; Available from: https://dfs.dc.gov/sites/default/files/dc/sites/dfs/page_content/attachments/STRmix%20v2.4%20Validation%20Report.pdf.

97. District of Columbia Department of Forensic Science Laboratory. Internal Validation of STRmix™ V2.4. December 30, 2015. 2015 [updated 2015; cited 19 January 2018]; Available from: https://dfs.dc.gov/sites/default/files/dc/sites/dfs/page_content/attachments/STRmix%20Validation.pdf.

98. California Department of Justice. STRmix V2.0.6 BFS Casework Internal Validation Summaries. 2016 [updated 2016; cited 18 January 2018]; Available from: https://epic.org/state-policy/foia/dna-software/EPIC-16-02-02-CalDOJ-FOIA-20160219-STRmix-V2.0.6-Validation-Summaries.pdf.

99. President's Council of Advisors on Science and Technology. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. 2016 [updated 2016; cited 22 April 2017]; Available from: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.

100. DPP v Tuite (Ruling number 3). 2017 VSC 442.

101. State of Florida v. Dwayne Cummings, Case No. 2016-CF-239. Case No 2016-CF-239; 2016.

102. Taylor D, Curran JM, Buckleton J. Importance sampling allows Hd true tests of highly discriminating DNA profiles. Forensic Sci Int Genet. 2017;27:74-81.

103. Harmon R, Budowle B. Questions About Forensic Science. Science. 2006;311(5761):607-10.

104. Adams N, Koppl R, Krane D, Thompson W, Zabell S. Letter to the Editor—Appropriate Standards for Verification and Validation of Probabilistic Genotyping Systems. J Forensic Sci. 2018;63(1):339-40.

105. LRmix Studio. LRmix Studio Release Notes. [cited 31 January 2018]; Available from: http://lrmixstudio.org/download/lrmixstudio-2.1.3-CommunityEdition-distribution.zip.

106. Lab Retriever. Version 2.2.1 released Nov 21, 2014. [cited 31 January 2018]; Available from: https://scieg.org/wp-content/uploads/2017/07/Lab_retriever_2.2.1.rtf.

107. Buckleton JS. A summary of the seven identified miscodes in STRmix™. [cited 31 January 2018]; Available from: https://johnbuckleton.files.wordpress.com/2017/12/a-summary-of-the-seven-identified-miscodes-in-strmix.pdf.

108. Butt N, Bauer D, Perlin MW. Validating TrueAllele® interpretation of DNA mixtures containing up to ten unknown contributors

American Academy of Forensic Sciences 70th Annual Meeting, Seattle, WA, 23-Feb-2018, 2018. 2018.

109. Perlin M. Objective DNA Mixture Information in the Courtroom: Relevance, Reliability and Acceptance. International Symposium on Forensic Science Error Management: Detection, Measurement and Mitigation, 2015; National Institute of Standards and Technology. Arlington, Virginia. 2015.

110. Budowle B, Connell ND, Bielecka-Oder A, Colwell RR, Corbett CR, Fletcher J, et al. Validation of high throughput sequencing and microbial forensics applications. Investigative Genetics. 2014;5(1):9.

111. Swaminathan H, Grgicak CM, Medard M, Lun DS. NOCIt: A computational method to infer the number of contributors to DNA samples analyzed by STR genotyping. Forensic Sci Int Genet. 2015;16:172-80.

112. Marciano MA, Adelman JD. PACE: Probabilistic Assessment for Contributor Estimation— A machine learning-based assessment of the number of contributors in DNA mixtures. Forensic Sci Int Genet. 2017;27:82-91.

113. Biedermann A, Bozza S, Konis K, Taroni F. Inference about the number of contributors to a DNA mixture: Comparative analyses of a Bayesian network approach and the maximum allele count method. Forensic Sci Int Genet. 2012;6(6):689-96.

114. Haned H, Pene L, Lobry JR, Dufour AB, Pontier D. Estimating the Number of Contributors to Forensic DNA Mixtures: Does Maximum Likelihood Perform Better Than Maximum Allele Count? J Forensic Sci. 2011;56(1):23-8.

115. Curran JM, Buckleton JS, Triggs CM. What is the magnitude of the subpopulation effect? Forensic Sci Int. 2003;135(1):1-8.

116. Curran JM, Buckleton JS. An investigation into the performance of methods for adjusting for sampling uncertainty in DNA likelihood ratio calculations. Forensic Sci Int Genet. 2011;5(5):512-6.

117. Slooten K, Caliebe A. Contributors are a nuisance (parameter) for DNA mixture evidence evaluation. Forensic Sci Int Genet.doi: 10.1016/j.fsigen.2018.05.004.

118. Jeanguenat AM, Budowle B, Dror IE. Strengthening forensic DNA decision making through a better understanding of the influence of cognitive bias. Sci Justice. 2017;57(6):415-20.

119. Budowle B, Bieber FR. Report on Review of Mixture Interpretation in Selected Casework of the DNA Section of the Forensic Science Laboratory Division, Department of Forensic Sciences, District of Columbia. [22 April 2015; cited 1 June 2018]; Available from: https://dfs.dc.gov/page/usao-report-april-2015.

120. Scientific Working Group on DNA Analysis Methods (SWGDAM). SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories.

2010 [updated 2010 14 January 2010; cited 1 June 2018]; Available from: http://www.forensicdna.com/assets/swgdam_2010.pdf.

121. Curran JM, Buckleton JS, Triggs CM, Weir BS. Assessing uncertainty in DNA evidence caused by sampling effects. Sci Justice. 2002;42(1):29-37.

122. Balding DJ, Nichols RA. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int. 1994;64:125-40.

123. Curran J, Walsh SJ, Buckleton JS. Empirical support for the reliability of DNA evidence interpretation in Australia and New Zealand. Aust J Forensic Sci. 2008;40(2):99-108.

124. Curran JM, Walsh SJ, Buckleton JS. Empirical testing of estimated DNA frequencies. Forensic Sci Int Genet. 2007;1(3-4):267-72.

125. Lauc G, Dzijan S, Marjanovic D, Walsh S, Curran J, Buckleton J. Empirical support for the reliability of DNA interpretation in Croatia. Forensic Sci Int Genet. 2008;3(1):50-3.

126. Balding DJ. Estimating products in forensic identification using DNA profiles. J Am Stat Assoc. 1995;90(431):839-44.

127. Balding DJ. Weight-of-evidence for forensic DNA profiles. Chichester: John Wiley and Sons, 2005.

128. Barrio PA, Crespillo M, Luque JA, Aler M, Baeza-Richer C, Baldassarri L, et al. GHEP-ISFG collaborative exercise on mixture profiles (GHEP-MIX06). Reporting conclusions: Results and evaluation. Forensic Sci Int Genet. 2018;35:156-63.

129. ENFSI DNA Working Group. Best Practice Manual for the internal validation of probabilistic software to undertake DNA mixture interpretation. 2017 [updated 2017; cited 31 January 2018]; Available from: http://enfsi.eu/wp-content/uploads/2017/09/Best-Practice-Manual-for-the-internal-validation-of-probabilistic-software-to-undertake-DNA-mixture-interpretation-v1.docx.pdf.cc