Document Title:	**DNA Mixture Study: Novel Metrics to Quantify the Intra- and Inter-Laboratory Variability in Forensic DNA Mixture Interpretation**
Author(s):	**Emily Rogers, Roman Aranda IV, Philippa M. Spencer, Denise R. Myers**
Document Number:	**304317**
Date Received:	**March 2022**
Award Number:	**2013-DNR-5042**

DNA Mixture Study: Novel metrics to quantify the intra- and inter-laboratory variability in forensic DNA mixture interpretation

Emily Rogers[a], Roman Aranda IV[b], Philippa M. Spencer[c], and Denise R. Myers[b],

[a]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA
[b]Defense Forensic Science Center, Forensic Exploitation Directorate, Forest Park, GA
[c]Defence Science and Technology Laboratory, Salisbury, U.K.

August 1, 2018

Abstract

Despite the prevalence and weight of forensic DNA evidence in the criminal justice system, little is known concerning the amount of variability in the interpretation of forensic DNA data. Variability in interpretation is affected when the DNA sample is complex, consists of multiple contributors, or the starting DNA template is minimal. Previous DNA mixture interpretation studies have qualitatively indicated that variability exists. We present a wide-scale quantitative assessment, using novel metrics, to measure the precision and accuracy of forensic DNA mixture interpretation. These metrics measure the accuracy and precision of a DNA mixture interpretation for each contributor in a mixture. Results of applying these metrics to the data demonstrate: 1) a significant amount of interpretation variability exists within and between laboratories; 2) accurate and precise interpretations are possible, with accuracy and precision being highly correlated. The quantitative results also indicate the ongoing need for training and benchmarking within laboratories and the need for dissemination of best practices between laboratories.

1 Introduction

Considered a reliable standard in forensics, DNA profiles generated from evidence are routinely entered into court proceedings. Because DNA evidence often plays a significant role in either convicting or exonerating persons of interest, the accuracy and precision of forensic DNA analysis is essential.

Although the science behind DNA profile generation is reliable and repeatable, the interpretation of this data is not free of subjectivity. Previous DNA mixture studies by the National Institute of Standards and Technology (NIST) have indicated variability in interpretation results when the same DNA mixtures were submitted to multiple laboratories [1]. This variability may be compounded as the complexity of a DNA sample increases, but the degree of variability present in DNA mixture interpretation by the forensic community is currently unknown. Thus, the size and the acceptable limits of variability within the forensic DNA community is also unknown. It is important to note that variability does not necessarily imply that an incorrect locus interpretation was generated, but that the analysts may be unable to determine a single, correct genotype, but instead provide a range of possible genotypes in which the correct genotype is included.

The purpose of this study was to assess the state of DNA mixture interpretation in the forensic DNA community. Specifically, this study investigates the variability in the precision and accuracy of DNA examiners' mixture interpretations given *.fsa* files. While other DNA mixture studies have been conducted, results have been reported on a broad, mainly qualitative level. The results of this study are presented as follows: 1) we developed novel metrics to quantify a DNA examiner's accuracy and precision in interpreting a variety of DNA mixtures and 2) we use these novel metrics to determine the current variability range within the forensic DNA community with 2- and 3-person DNA mixtures.

The amount of variation that exists, and whether that variation is consistent within and between laboratories, is of interest to the forensic DNA community. Because DNA training and interpretation protocols are deterined by each individual laboratory, we investigated whether intra-laboratory variability, where examiners utilize identical protocols and training, would be significantly different than inter-laboratory variability, where protocol and training differences are expected. The metrics developed by the study also provide insight into strengthening and improving the current state of forensic DNA training and quality control. The quantitative data and novel metrics can be used to benchmark an

examiner's interpretation performance, determine mixture interpretation limitations within a laboratory, and infer whether a new method implemented in a laboratory yields improved precision and accuracy over previous methodologies.

2 Background/Related Works

Current forensic analysis of DNA relies on sections of noncoding DNA, composed of 3-5 repeating base-pair fragments, ranging between 100-450 total nucleotides in length. Known as short tandem repeats (STRs), these repeats occur at multiple locations in the genome and forensic laboratories utilize a select few to generate a genetic profile. The exact number of STRs at a single locus varies widely enough between individuals and, when multiple locations (loci) of STRs are considered in combination, they can be used to discriminate one person from another. During analysis of a DNA sample containing a single contributor, the genetic profile can be determined relatively easily. When additional contributors are added to a sample, the complexity of the sample is increased and it may be difficult to separate the data for each particular contributor.

Extensive research and evaluation has gone into refining the data interpretation of the generated STR data. With its ability to individualize and its popularity in pop culture (television forensics and courtroom dramas), DNA evidence can influence the generation of a verdict [2]. Due to the varying complexity of DNA evidence, results are influenced by the ability of its practitioners to accurately interpret the data and in a manner that can be duplicated by another DNA examiner.

Laboratory accreditation is intended to address these issues by implementing quality controls and establishing quality assurance systems to minimize error and improve consistency. The FBI has generated DNA processing and interpretation guidelines with widespread adoption [3, 4], but the specific interpretation guidelines and limits are largely set by each laboratory. The quality of interpretation and execution is also influenced by the quality of the DNA data generated from a given sample and by the examiner skillset. In addition, interpretation results are likely to vary between laboratories, as individual laboratories determine their own DNA protocols, DNA amplification kits chosen, analytical and stochastic thresholds determined for the data [5], and whether a known reference profile is used to aid analysis [6]. Likewise, variability between examiners within a laboratory may exist, due to interpretation experience versus perceived risk of interpreting a difficult mixture [5, 7, 8, 9, 10, 11].

Sample quality presents its own challenges to interpretation [12]. Factors such as a low level of DNA template [13, 14, 15, 16], can negatively impact interpretation and may contribute to the interpretation variation. In addition, determining the assumed number of contributors (NOC) in a mixture is a non-trivial problem [17, 12, 18]; ambiguity increases as NOC increases and complex, sophisticated computational programs have been developed to address this issue and assist the examiner [19]. Stochastic effects [17, 20] such as the increase in allelic dropout [21, 22] and stutter [23, 24, 25, 26], the overlapping alleles between sources [27, 28] and imbalanced contributor ratios also increase the complexity of the data and interpretation.

The difficulty of mixture interpretation is compounded by the lack of consensus regarding standard methods and protocols for analysis and interpretation of DNA mixtures [5, 29, 30], the use of qualitative versus quantitative methods [29], the type of statistic applied, and the role of software and computational programs. Thus, the state of DNA mixture interpretation with respect to its accuracy and precision remains an important open question. Past inter-laboratory studies include those conducted by NIST [31, 32, 33, 34]. Similar collaborative studies have also been carried out by the European DNA Profiling Group (EDNAP), with qualitative differences being reported. Blind trial testing of multiple laboratories has also been performed, notably by the German DNA Profiling Group (GEDNAP [35, 36]). Results from these previous studies have focused on general trends and qualitative assessments, with reports of "results obtained by the vast majority of participating laboratories who consistently and reproducibly produce correct results" [36]. GEDNAP studies have also identified sources of transcriptional and transpositional errors [36].

By employing novel metrics, this study attempts to quantitatively identify the current range of variation in DNA mixture interpretation within the forensic community.

Results

The novel metrics employed in this study aid in determining the following:

1. The amount of variation and general state of DNA mixture interpretation among participating examiners

   and

2. The amount of variation and general state of DNA mixture interpretation among laboratories

Using the study's accuracy and precision metrics defined in *Materials and Methods*, statistics were calculated for the overall quality of an examiner's DNA interpretation by comparing the examiner generated genotypes to the true, known genotypes of each contributor in a mixture. The variation, or range of genotypes, contained therein is then determined. The variation is quantified by taking the median score of the set and calculating the interquartile range (IQR, the difference between the first and third quartile) of scores.

This enables the data to be calculated into two statistics per grouping: the median score (the middle value) and the IQR as a measure of the variability within each designated group. The median scores help compare results between laboratories. Changes in the metric values also reflect increasing mixture complexity (increasing the number of contributors and/or equalizing contributor ratios), as seen in the provided mixtures. If one laboratory's median score is significantly lower than another laboratory's median for a given mixture, or even the median overall score, this indicates a need for improved interpretation compared to the rest of the participating forensic community. The interquartile range is a direct measure of variability in the interpretation of a given mixture by a laboratory. Thus, a larger IQR implies that examiners within a laboratory are displaying larger variability despite the similar training and protocols.

The scores are calculated at a range of granularity levels to expose any outliers or unusual degrees of variability: the Genotype Interpretation Metric (GIM) and Allelic Truth (AT) metric were calculated at each locus of a mixture, for an individual contributor in a mixture, by overall mixture (including all contributor genotypes of a mixture), by laboratory, and, finally, by grouping laboratories defined by a particular jurisdiction (local, state, federal, and international/other). Due to the amount of data within each granularity, the quantitative analysis focuses on all participating examiners within the study and laboratories with at least five participating examiners. Data is visualized through boxplots and scatterplots. Boxplots directly visualize both the median score (designated as the central red line) and the IQR (designated as the top and bottom limits of the box). Boxplots allow visual inspection of the overall variability distribution for a laboratory's accuracy and precision. Plotting each laboratory's overall metrics (scoring for all examiners in a laboratory), enables outliers to become apparent.

Table 1: Details of the DNA mixture samples sent to participants: Mixtures 1-6 are listed and the number of contributors, the ratio of the major and minor contributors, and whether a reference profile was provided to the examiners is displayed. A reference profile refers to a known contributor that is assumed to be present in the mixture.

Mixture	Contributors	Ratio	Reference?
1	2	3:1	N
2	2	2:1	Y
3	2	3.5:1	N
4	2	4:1	N
5	3	4:1:1	Y
6	3	1:1:1	N

While boxplots visualize median and IQR scores for either accuracy or precision among the labs, scatterplots can directly compare both. We plot either median or IQR scores of accuracy against precision in Figures 4 and 5, with each laboratory represented as a dot with size proportional to the number of participants in a laboratory (note: not all examiners from a laboratory participated in the study). This enables the study to assess whether there is a trade-off between precision and accuracy when an examiner interprets a DNA mixture and whether examiners generally achieve one at the expense of the other.

2.1 Mixture 1, A Baseline Mixture

As a preliminary exploration into interpretation variability, we calculated the described GIM and AT metrics by laboratory region at the local level (city- and county-level), state, federal and the rest of the labs combined together due to their small samples size (international/other), as seen in Figure 1a. While all mixtures in this study exhibit variability, Mixture 1 (see Table 1) is highlighted, which shows the jurisdiction-based metrics for Mixture 1 (Figure 1a). This mixture was designed to be the most favorable mixture provided for interpretation with only two contributors, the largest targeted ratio between major and minor contributor, and clear peak heights in the electropherogram data which averaged well over the provided standard stochastic threshold. While the more difficult mixtures can be expected to have a spectrum of responses, Mixture 1 was designed as a baseline with the least amount of variation expected.

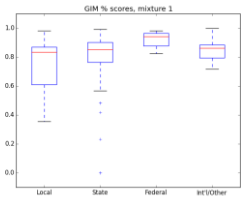
In Mixture 1, GIM and AT variability is found at all laboratory levels, both between laboratories within the same grouping and between groups

(Figure 1a). GIM and AT scores grouped by jurisdiction show small differences in median score but larger differences in IQR (Figure 1a), resulting from larger variation among examiners in the jurisdiction. Federal labs display the highest GIM and AT median scores, as well as the smallest IQR. Local labs show the most variation, with an IQR of slightly above 0.6 to slightly under 0.9 for both GIM and AT scores.

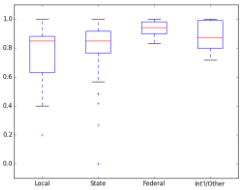
We also investigate whether variation, via GIM and AT scores, is due to distinct differences between laboratories or is distributed evenly between all laboratories. In order to investigate the variability within a lab, we examine the local laboratories using the ID+ amplification kit with at least five participating examiners. Examiners using PP16 kits also displayed similar variation; to clarify the manuscript and to limit redundancy, only ID+ results are discussed in detail. Six such participating laboratories exist and are shown in Figure 1c. Both the GIM and AT scores indicate that the spread of scores is due to variation between the local laboratories and between examiners in the local laboratories. In particular, the differences in median scores between laboratories is more striking and pronounced than in other regional scores, with median GIM and AT scores ranging from around 0.4 to well over 0.8. In terms of AT (Figure 1d), this indicates that certain labs (Laboratories A, E and F) AT scores were consistently 0.8 or greater, while another lab (Laboratory D) AT scores were less than 0.5. Additionally, while labs B and C have median scores between these two extremes, they also exhibit significant variation within their scores: Their IQR's span from approximately 0.6-0.9 of correctly called alleles.

We can further investigate Laboratory C's (n=8 participating examiners) variation by separating metrics by the major and minor contributors in Mixture 1. Figure 1e indicates that while there is a tight consensus in both GIM and AT scores for the major, the variability in overall mixture score comes primarily from the minor profile, whose IQR spans half (0.5) of the entire spectrum.

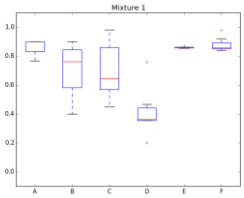
Finally, further investigation of variability of the minor contributor in Mixture 1 by locus (Figure 1g) indicates that there is no consensus in either precision or accuracy across loci. Since the GIM IQR includes a zero score on a majority of loci, this also indicates several examiners deemed the minor contributor as uninterpretable or 'inconclusive' at various loci. Oppositely, several examiners within Laboratory C chose to interpret the minor profile and all but one locus has an AT IQR reaching 1, indicating that more than one examiner correctly fully deconvoluted the minor genotype at each locus. Thus, even with the study-designed baseline Mixture 1 data, Laboratory C displays a wide range of variability.



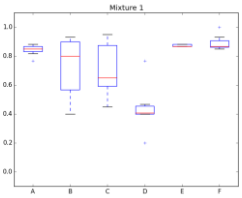
(a) Mixture 1: GIM Metric by Jurisdiction



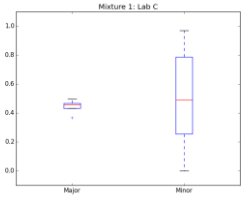
(b) Mixture 1: AT Metric by Jurisdiction



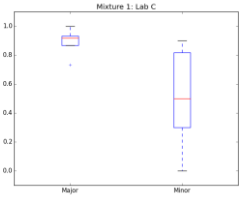
(c) Mixture 1: GIM Metric by Large Laboratories (n≥5 examiners)



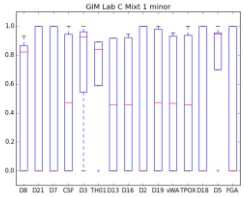
(d) Mixture 1: AT Metric by Large Laboratories (n≥5 examiners)



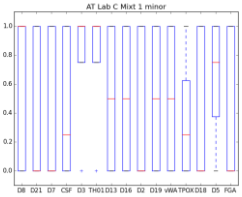
(e) Mixture 1: Contributor GIM Metric for Laboratory C (n=8)



(f) Mixture 1: Contributor AT Metric for Laboratory C (n=8)



(g) Mixture 1: Minor Contributor GIM Metric by Locus for Laboratory C



(h) Mixture 1: Minor Contributor AT Metric by Locus for Laboratory C

Figure 1: Preliminary data exploration of interpretation variability across large local laboratories with participating examiners using the ID+ amplification kit, and analysis at the profile and locus levels. Each level, starting with jurisdiction at the top, occupies a row. The left column shows boxplots of GIM scores, while the right shows boxplots of AT scores. At every level, note the differences between median scores (red line) and IQRs (box height), as well as the correlation between GIM and AT scores.

2.2 Variability in Mixtures 1-6

Since examiner training and protocols are primarily the domain of individual labs, differences in median scores between laboratories may be related to differences in training, amplification kits, and amplification and interpretation protocols. Similarly, differences in lab IQR scores may be due to differences in conformity to protocols, i.e. are examiners implementing the written protocols consistently. To ensure appropriate sample size, we report only those labs with at least five examiners in Figures 2 and 3. For reference, we included and compared statistics against all examiners in the thirteen largest participating laboratories and also against all participating examiners from across all laboratories. Figure 2 indicates that variation exists even in Mixture 1, and increases with more complex mixtures in Mixtures 2-6.

The range of complexity found in the six mixtures is reflected in the wide range of AM and GIM scores generated. To better understand the different scores given to examiners presented with the same *.fsa* mixture data file, we analyzed participants that utilized mixture data generated from the ID+ amplification kits. Because Mixtures 5 and 6 are 3-person mixtures rather than 2-person mixtures (as is the case for Mixtures 1-4), we normalize the $AT(D_M)$ score per mixture by converting it to a percentage of the total score possible. Namely, dividing $GIM(D_M)$ by 30 for a 2-person mixture (30 total pairs of alleles for a 2-person mixture) and 45 for a 3-person mixture (45 total loci for a 3-person mixture) yields the percentage scores found in Table 2. Normalized $AT(D_M)$ scores are similarly obtained by dividing by 60 and 90, respectively, as seen in Table 3.

We compare the variability in precision using the GIM scores and the variability in accuracy using the AT scores in Figures 2 and 3. In Mixture 1, the baseline mixture of the study, significant inter- and intra-laboratory variability exists in the median lab scores. Some laboratories have a median GIM score close to 1.0, while others have a score below 0.7, and a single lab is below 0.5. With intra-laboratory variability, certain laboratories have a tight range of scores between examiners (low IQR), while others have an IQR of over 0.2. Thus, the variation is not consistent between laboratories.

The general spread of AT scores resembles that of the GIM scores, where laboratories that score high on precision also score high on accuracy. It should be noted that an examiner can generate a high GIM score, but not provide the correct genotype and therefore have a lower AT score. Additionally, the GIM IQR of each laboratory is also similar to its AT IQR, indicating that the amount of variability within a lab is consistent with respect

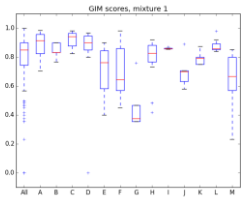
to accuracy as well as precision and varies between participating laboratories. Comparing GIM and AT median and IQR scores directly in Figures 4 and 5 confirms this, with laboratory scores fairly close to the identity diagonal. More rigorous analysis using Spearman's coefficient [37], a nonparametric measure of rank correlation that assesses the relationship between two variables, shows a correlation factor of over 0.9 for almost all mixtures. This indicates a high correlation for median and IQR scores.

This patterns also exists for Mixture 2 (Figure 4). Even so, the range of scores found within laboratories still exhibits significant variation, reaching approximately 0.5 for one laboratory, but laboratories with high accuracy and precision scoring are correlated.

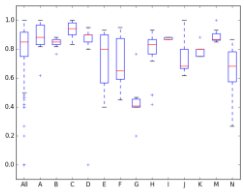
Although the correlation between GIM and AT scores persists in Mixture 3 (Figure 4), the GIM and AT values decrease significantly. This indicates the general difficulty of interpreting a mixture in the absence of an assumed known contributor, with a more difficult Major:minor contributor ratio, and with unpredictable stochastic effects. Nonetheless, certain laboratories and examiners still exhibit a uniformly high level of accuracy and precision. Figure 4 shows the wide scatter of laboratory scores, with median and IQR scores ranging across the spectrum from near 0 (deeming the data inconclusive or uninterpretable) to near 1.0 (full deconvolution of each contributor in the mixture).

The median scores for Mixture 4 in Figure 4 improve to levels similar to that of Mixture 1 and 2, which are both similar in contributor ratio. The median and IQR scores of Mixture 4 span approximately half the spectrum of Mixture 3, from 0.5 to 1.0 for median scores to 0 to 0.5 for IQR scores, as seen in Figure 4. The points lie near the identity diagonal in Figure 4, again indicating correlation in the GIM and AT scores.

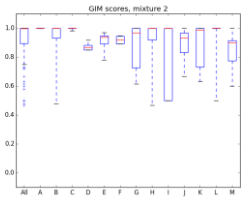
The lowered GIM and AT scores of Mixtures 5 and 6 reflect the increased complexity of a 3-person mixture. Results indicate a majority of the participating laboratories did not attempt to deconvolute such mixtures, either being prohibited by their protocols in attempting to interpret a 3-person mixture or deeming the data uninterpretable; the former is a laboratory wide issue, while the latter is a matter of personal choice. For scoring purposes, both were deemed inconclusive and scored identically. In Mixture 5, a 4:1:1 Major:minor:minor mixture with a known profile provided to the examiner, a majority of the participating laboratories had a median GIM at or below 0.33, a score indicative of reporting the major profile only (the provided reference profile) without deconvoluting any of the minor contributors. Interestingly, a few laboratories deconvolute



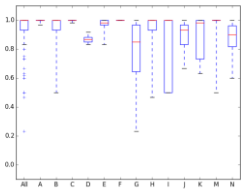
(a) Mixture 1 by GIM



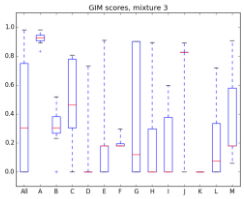
(b) Mixture 1 by AT



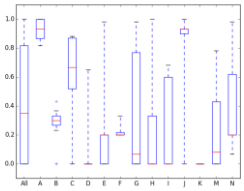
(c) Mixture 2 by GIM



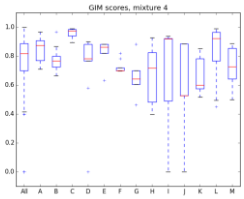
(d) Mixture 2 by AT



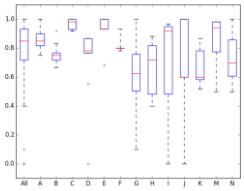
(e) Mixture 3 by GIM



(f) Mixture 3 by AT



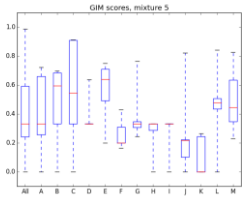
(g) Mixture 4 by GIM



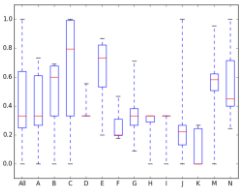
(h) Mixture 4 by AT

Figure 2: Boxplots for 2-person Mixtures $1-4$ of the thirteen labs with five or more participants, giving the distributions of each laboratory's respective scores. Red lines indicate median scores, boxes delimit the IQR, with outliers beyond it. The left column displays the GIM scores for the entire mixture from each laboratory, while the right column displays the AT scores for the entire mixture from each laboratory. Lab designation is not related to those identified in Figure 1.

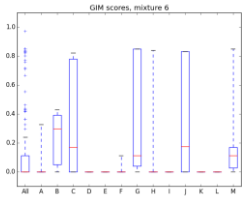
This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.



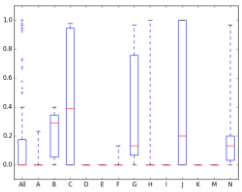
(a) Mixture 5 by GIM



(b) Mixture 5 by AT

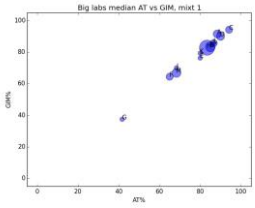


(c) Mixture 6 by GIM

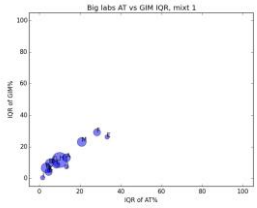


(d) Mixture 6 by AT

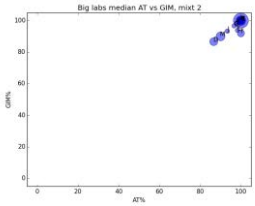
Figure 3: Boxplots for 3-person Mixtures $5-6$ of the thirteen labs with five or more participants, giving the distributions of each lab's respective examiners' scores. Red lines indicate median scores, boxes delimit the IQR, with outliers beyond it. The left column displays the GIM scores for the entire mixture from each laboratory, while the right column displays the AT scores for the entire mixture from each laboratory. Lab designation is not related to those identified in Figure 1.



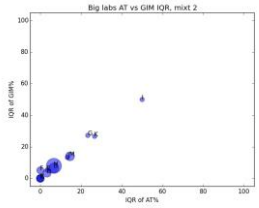
(a) Mixture 1, Median GIM vs AT



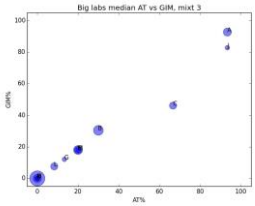
(b) Mixture 1, IQR GIM vs AT



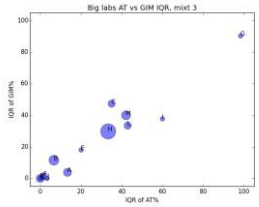
(c) Mixture 2, Median GIM vs AT



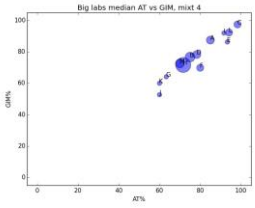
(d) Mixture 2, IQR GIM vs AT



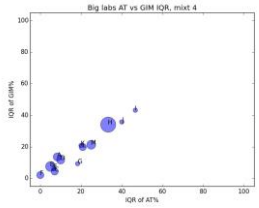
(e) Mixture 3, Median GIM vs AT



(f) Mixture 3, IQR GIM vs AT

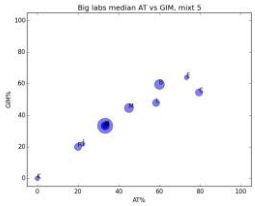


(g) Mixture 4, Median GIM vs AT

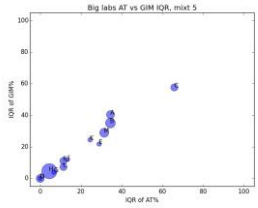


(h) Mixture 4, IQR GIM vs AT

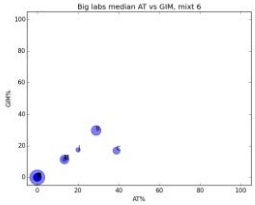
Figure 4: Scatterplots of GIM vs AT scores for 2-person Mixtures 1 −4 of the thirteen labs with five or more participants, defined as "Big" labs. The size of each dot is proportional to the number of participating examiners in the lab. The left column displays the median scores from each lab, while the right column displays the IQR scores. Those with higher GIM and AT scores will coincide with dots near the upper right corner.



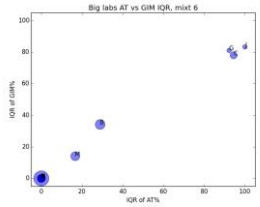
(a) Mixture 5, Median GIM vs AT



(b) Mixture 5, IQR GIM vs AT



(c) Mixture 6, Median GIM vs AT



(d) Mixture 6, IQR GIM vs AT

Figure 5: Scatterplots of GIM vs AT scores for 3-person Mixtures 5−6 of the thirteen labs with five or more participants, defined as "Big" labs. The size of each dot is proportional to the number of participating examiners in the lab. The left column displays the median scores from each lab, while the right column displays the IQR scores. Those with higher GIM and AT scores will coincide with dots near the upper right corner.

the minor profiles, with the GIM score in general lagging slightly behind the AT score. This indicates that laboratories that deconvolute mixture 5 are more cautious, reporting more possible genotype combinations than in other mixtures. However, their accuracy remains high.

Mixture 6 (Table 1) has, unsurprisingly, the lowest metric scores, with a majority of the labs reporting "Inconclusive" for the entire mixture. A few laboratories, however, deconvolute the mixture with varying degrees of success, as seen by a median GIM score above 0 in Figure. 3. The majority of examiners within those laboratories attempted to deconvolute the mixture. Within these labs, certain examiners achieved relatively high GIM and AT scores even with a highly ambiguous and difficult mixture dataset (1:1:1 ratio 3-person mixture, Figures 3 and 5). The success of these examiners indicates that deconvoluting complex 3-person mixtures is possible with a high degree of accuracy and precision, giving hope that doing so could one day be the norm and not the exception.

3 Discussion

3.1 Quantifying variability in mixture interpretation

The boxplots of the GIM and AT scores for the six mixtures provide a snapshot on the state of DNA mixture interpretation in the forensic DNA community. The inclusion of an assumed known reference DNA profile has a marked positive effect on interpretability, increasing both GIM and AT scores such that the two-person Mixture 2 has the best metric results with respect to both median and IQR scores. Likely peak height, cited in the survey as the most influential factor in interpretation, plays a significant role in quality of interpretation. Mixtures 1 and 4 display higher GIM and AT results than Mixture 3, whose average peak height results is just barely over the study-provided stochastic threshold, and was intentionally generated to determine the change in variation with sub-optimal peak heights and reflect difficult casework samples.

Results generally indicate that the two-person DNA mixtures given were considered interpretable by a majority of examiners (ability to generate a profile for each contributor). Mixtures 1, 2, and 4 have median GIM and AT scores of at least 0.8 for all examiners (Tables 2 and 3). A majority of examiners appear to have more difficulty with Mixture 3, a mixture with low peak height values and a median score of approximately 0.3. However, there are examiners able to fully deconvolute Mixture 3, indicating that while low peak heights are a challenge, they are not insurmountable.

In comparison, three-person mixtures are difficult, with a majority of laboratories not considering the datasets for interpretation. The majority of examiners that interpreted Mixture 5, a 4:1:1 three-person mixture with a reference profile provided, only reported back the genotype of the reference profile without further attempting to interpret the rest of the contributors in the mixture, thus giving a median GIM and AT score of 0.33. With Mixture 6, the majority of laboratories and examiners did not attempt to interpret the mixture, with a median GIM score of 0, indicating that at least half of all examiners marked all loci as 'Inconclusive' or uninterpretable. Although a majority of the examiners did not attempt to deconvolute either Mixtures 5 or 6, there are examiners that succeed in interpreting and fully deconvoluting the mixture into individual contributors with high precision and accuracy. Thus, while interpreting a 3-person mixture is clearly a complex challenge, it is not an insurmountable one. Interpreting difficult three-person mixtures is achievable, even with Mixture 6's low peak heights and with three contributors at equal ratio.

3.2 Capturing the relative interpretation state for each participating laboratory

A major goal of this study was to identify the variability within individual participating laboratories. Determining the areas of interpretation difficulty, such as those with large variation or mixtures that an examiner will not interpret, reveals possible areas for protocol and training improvements. When new changes are implemented into a laboratory, the GIM and AT metrics can be recalculated among examiners and compared to previous GIM and AT scores and determine whether the changes decreased variation and increased accuracy.

We found that overall, among the larger participating laboratories, a strong correlation (Spearman's coefficient > 0.9) exists between the median GIM and AT scores. We consider these laboratories that are consistently able to deconvolute difficult mixtures with high precision and accuracy and minimized variations as the current upper threshold of DNA interpretation. Hence, long-term goals for other participating laboratories should include improving median scores and shrinking IQR scores to be closer to those of high-performing laboratories and to be more internally consistent.

A laboratory's IQR number is helpful in indicating whether perhaps training can minimize variability, with a smaller range indicating stronger adherence to and/or clearer protocols. Since we found no (strong) correlation between scores and years of experience (Spearman's coefficient), we posit the range of scores found within a lab are not due to differing levels of experience, but to other factors such as in-house training and quality control. Therefore, another laboratory goal is to decrease the range of scores from each constituent examiner. These metrics can then be recalculated among examiners when protocol changes are implemented.

Furthermore, laboratories often have a tight IQR in one mixture, but a much larger range in another. Likewise, a separate laboratory may experience the exact opposite in their IQR scores for the same two mixtures. Hence, the areas of consensus (and the lack thereof) differ by laboratory and potentially point to specific areas of ambiguities or difficulties unique to each lab and/or protocol.

4 Conclusion

Since the results of a DNA interpretation in casework often have profound and long-lasting repercussions in the criminal justice system, the interpretation should be as objective, reproducible, and error-free as possible. Like all forensic fields, the quality of DNA interpretation is dependent on a number of factors, including the sample, examiner, and laboratory. This study provided six carefully curated DNA mixtures of varying difficulty to 189 participating examiners and 55 laboratories. The study attempted to provide a snapshot on the interpretation variability in the forensic DNA community.

The results suggest that there is significant intra- and inter-laboratory variation. They also suggest that two-person mixtures with signal peaks above stochastic threshold are generally interpretable, while three-person mixtures are currently beyond the scope or protocol limits for most participating examiners. The results highlight the impact of a reference profile and of strong peak heights in the interpretability of a mixture. There are, however, laboratories and participants that were able to interpret the difficult three-person mixtures and resolve genotypes for each contributor, even under very challenging conditions with nearly equivalent contributor ratios.

Table 2: Data for GIM scores in percentage for all laboratories with five or more examiners. Individual scores were computed per mixture for every examiner in the lab, and the overall median score and interquartile range (IQR) are reported. Median and IQR scores are reported for all combined examiners in a large lab, as well as all examiners in the study.

	Mixture 1		Mixture 2		Mixture 3		Mixture 4		Mixture 5		Mixture 6	
	Median	IQR	Median	IQR	Median	IQR	Median	IQR	Median	IQR	Median	IQR
Lab A	91.5	13.1	100.0	0.0	92.7	3.8	87.5	13.6	33.3	40.4	0.0	0.0
Lab B	83.3	6.7	100.0	6.7	30.4	11.5	76.7	7.6	59.5	35.1	29.7	34.2
Lab C	94.2	8.7	100.0	0.1	46.1	47.3	97.5	4.4	54.4	57.7	16.9	78.0
Lab D	90.0	9.7	86.7	3.3	0.0	0.0	78.3	11.7	33.3	0.0	0.0	0.0
Lab E	76.3	26.2	93.9	5.4	17.9	18.0	86.3	6.2	63.9	21.8	0.0	0.0
Lab F	64.5	29.2	91.9	5.1	18.0	1.4	69.9	2.1	20.0	11.1	0.0	0.0
Lab G	37.5	11.1	96.7	27.2	12.1	90.3	64.1	9.3	33.3	3.9	11.3	81.1
Lab H	82.8	11.7	100.0	8.0	0.0	29.8	71.7	34.1	33.3	4.6	0.0	0.0
Lab I	85.8	0.4	100.0	50.0	0.0	37.7	92.0	43.2	33.3	0.0	0.0	0.0
Lab J	70.0	7.4	93.3	13.3	82.8	0.0	52.7	35.8	21.7	12.0	17.4	83.3
Lab K	79.3	5.0	98.7	26.7	0.0	0.0	60.0	20.9	0.0	24.4	0.0	0.0
Lab L	85.8	4.1	100.0	0.0	7.6	33.5	92.3	19.8	47.9	7.2	0.0	0.0
Lab M	66.9	23.1	90.0	14.0	18.0	40.0	72.6	21.3	44.6	29.0	11.3	14.1
All Large Labs	84.5	16.5	98.3	11.7	18.0	59.8	78.3	22.0	33.3	24.8	0.0	11.3
All Examiners	85.0	15.5	99.9	10.4	30.4	75.0	81.7	18.9	33.3	35.0	0.0	11.3

This study quantified variations using novel metrics and provides forensic laboratories a framework to track the effects of any protocol changes to improve interpretation and minimize variability. More in-depth analysis of the study results is needed to identify sources of errors, whether they are lab-specific or community-wide. In this way, the discussion of common errors and metrics from the best-practice labs can be transferred to the rest of the forensic DNA community. The ability to do so can only be helped by the sharing of protocols and techniques within the DNA community [38, 39, 30]. Thus, there is a need for resources and feedback that reach beyond the scope of individual labs, such that successful methods in one lab can become prevalent methods in the general community.

In addition, significant advancements have been made in the forensic DNA community regarding DNA mixture interpretation since the submission of the DNA interpretation in 2014. Probabilistic DNA interpretation software programs are more commonly available and are more widely in use by the forensic community, likely aiding in the interpretation of more difficult mixtures. Thus, the authors anticipate that a second round of a large-scale study may find a further reduction in variability, possible correlated with increased use of automated tools. However, it should be noted that software assistance is not a substitution for mixture interpretation [34], nor is it the subject of this paper.

Table 3: Data for AT scores in percentage of total possible for all labs with five or more examiners. Individual scores were computed per mixture for every examiner in the lab, and the overall median score and interquartile range (IQR) are reported. Median and IQR scores are reported for all combined examiners in a large lab, as well as all examiners in the study.

	Mixture 1		Mixture 2		Mixture 3		Mixture 4		Mixture 5		Mixture 6	
	Median	IQR	Median	IQR	Median	IQR	Median	IQR	Median	IQR	Median	IQR
Lab A	88.3	13.3	100.0	0.0	93.3	13.3	85.0	8.3	33.3	34.4	0.0	0.0
Lab B	85.0	3.3	100.0	6.7	30.0	6.7	75.0	5.0	60.0	34.4	28.9	28.9
Lab C	94.1	8.3	100.0	0.0	66.7	35.0	98.3	7.1	79.4	65.8	38.9	94.0
Lab D	90.0	5.0	86.7	3.3	0.0	0.0	78.3	10.0	33.3	0.0	0.0	0.0
Lab E	80.0	33.3	98.3	3.3	20.0	20.0	93.3	6.7	73.3	28.9	0.0	0.0
Lab F	65.0	28.3	100.0	0.0	20.0	1.7	80.0	0.0	20.0	11.4	0.0	0.0
Lab G	41.7	6.7	96.7	23.3	13.3	98.3	63.3	18.3	33.3	6.7	13.3	92.2
Lab H	83.3	10.0	100.0	6.7	0.0	33.3	71.7	33.3	33.3	4.4	0.0	0.0
Lab I	86.7	1.7	100.0	50.0	0.0	60.0	91.7	46.7	33.3	0.0	0.0	0.0
Lab J	68.3	13.3	93.3	13.3	93.3	3.3	60.0	40.0	22.2	13.3	20.0	100.0
Lab K	80.0	5.0	98.3	26.7	0.0	0.0	60.0	20.0	0.0	24.4	0.0	0.0
Lab L	86.7	4.6	100.0	0.0	8.3	42.9	94.2	20.8	58.3	11.4	0.0	0.0
Lab M	68.3	20.8	90.0	14.6	20.0	42.1	70.0	25.0	45.0	31.4	13.3	16.7
All Large Labs	85.0	16.7	100.0	11.7	20.0	68.3	80.0	23.3	33.3	33.3	0.0	13.3
All Examiners	85.0	16.7	100.0	6.7	36.7	81.7	85.0	21.7	33.3	37.8	0.0	18.9

5 Materials and Methods

The study was conducted between 2013-2014, and initialized by generating six mixture samples, with four being a mixture of two DNA sources, and the remaining two being a mixture of three DNA sources. Each sample was analyzed, and the electropherogram files obtained. The *.fsa* files, along with a questionnaire, and standardized worksheets to record results, were then sent to forensic laboratories, primarily from the United States, for voluntary participation in the study. Fifty-five laboratories with 189 examiners returned completed questionnaires and worksheets.

5.1 Preparation of Samples

Samples were taken from buccal swabs of 14 unrelated individuals, incubated at 56°C, extracted, and purified with QIAGEN BioRobot EZ1 Advanced XL™ with Investigator Card. The estimated con- centration of DNA present from each contributor sample was determined by quantifying with the Applied Biosystems® Plexor® HY (Promega) quantification kit and 7500 HID instrument. DNA quantities were targeted at 1 ng DNA. The sample DNA was then amplified using the Applied

Biosystems® Geneamp™ PCR System 9700 and separated by capillary electrophoresis on the Applied Biosystems® 3130xl Genetic Analyzer™ with the Identifiler Plus® (ID+) or PowerPlex® 16 Hot Start (PP16 HS) amplification kits. Single source profiles were generated for each of the 14 contributors in order to serve as a key for the mixtures.

The 14 individual profiles were populated into NIST's Virtual Mixture Maker [40] to develop hypothetical 2- and 3-person mixtures. The program performs a pairwise comparison of STR profiles in a dataset and calculates the number of loci possessing 1-6 alleles in all possible mixtures. The median allelic overlap from the 2- and 3-person mixtures were selected for the study and used in the mixture generation as follows.

Accurate assessment of the single-source sample concentrations allowed for the appropriate selection of DNA target quantities to be used in order to generate mixtures at the desired ratios. Single-source samples (except two samples) were normalized to approximately 0.1ng/μL in a final volume of 500 μL of TE buffer. For these two samples, the concentrations were adjusted by using their peak height values as an concentration estimate after analysis in the Applied Biosystems® 3130xl Genetic Analyzer™ due to larger peak heights than expected.

Two and 3-person mixtures were generated using ID+ and PP16 HS amplification kits. Single-source samples were amplified at a target of 0.4ng with the PP16 HS kit. The 3130xl CE instrument was run with each respective kit according to the kit manufacturer's instructions using 28 reaction cycles. The resulting *.fsa* files were analyzed utilizing the GeneMapper® IDX™ software (versions 1.0.1/1.1).

A total of six mixtures were generated, including four 2-person mixtures and two 3-person mixtures. All mixtures were then quantified with Plexor® HY, followed by amplification in triplicate and analysis on the 3130xl CE Genetic Analyzer. Peak height response (signal intensity) was used to determine the ratio of one mixture to the other. All peak heights for a given contributor to a mixture were summed and then compared with the sum of peak heights from the other contributors to estimate a ratio of contributors. If the ratio needed to be adjusted, the concentrations were adjusted based on the peak height response of the unshared alleles.

5.2 Examiner Participation

Participants in the study were solicited via forensic conference presentations, the *Crime Lab Minute* newsletter from the American Society of Crime Lab-

oratory Directors, and direct solicitation to DNA Technical Leaders across American forensic laboratories. Each examiner was provided a questionnaire, DNA mixture data, and a response worksheet to record their interpretation analysis. The worksheet allowed for an interpreted profile to be entered for each contributor in a mixture. The analysts were required to determine the number of contributors per mixture. Laboratories were requested to have each individual examiner complete the interpretation and submit their own interpretation.

One hundred and eighty- nine individual examiners from 55 laboratories returned completed interpretation worksheets in 2013-2014, providing a snapshot on interpretation during that time frame. Laboratories were primarily from the United States, with a few laboratories participating from Canada, the United Kingdom and New Zealand. The U.S. labs were categorized within the study as either local (i.e. affiliated with a city), state, or federal, with local labs being the most numerous.

Examiners were sent the *.fsa* files from either the ID+ or PP16 kit based on the kit routinely used by the laboratory and participating examiner. In addition to the six *.fsa* files, each examiner was sent an Excel® spreadsheet with instructions in which to enter their interpretation and comments, as well as stochastic and analytical thresholds that were to be used as part of the interpretation. Fields were present in the spreadsheet to enter data for each mixture on the NOC, estimated mixture proportion, genotype for each locus in a contributor profile, and any additional comments, such as the type of interpretation analysis performed on the locus, interpretation model used, final statistic if generated, and any other comments.

Also submitted to each examiner for completion was a study survey. A participant questionnaire was included and requested the examiner's education level, years of experience, the most influential factor in interpretation according to the examiner (such as peak height or NOC), any formal training received, typical caseload at their respective laboratory, length of time to interpret each mixture, and a qualitative assessment of the mixture difficulty.

5.3 Metrics

In order to understand and measure variability among examiners and between laboratories collectively, a method was needed to quantify and compare the genotypes generated from a DNA mixture interpretation. Because interpretations are often complex and include a large range of possible solutions (obligates, CPI/CPE, etc.), the issue of comparing them is non-trivial.

Addressing this issue necessitates defining a metric that quantifies the effectiveness of an interpretation. More specifically, a high quality interpretation has the following characteristics: it is able to correctly identify the number of contributors in the mixture, deconvolutes the genetic profile of each contributor, and correctly identifies the implied genotype at each locus, excluding extra genotypes at each locus. These characteristics can be categorized as "accuracy" (the correct genotype is included at a particular locus for a contributor) and "precision" (minimizing the number of genotypes included at a particular locus for a contributor), two qualities that have previously been identified as crucial to a high quality interpretation [36]. Although each generated genotype of an interpretation is not strictly speaking an independent trial, we nevertheless use the terms "accuracy" and "precision" in a general, not scientific, sense to describe our metrics. A high quality interpretation is both accurate, including the correct answer (genotype), and precise in giving only the correct answer (minimal genotypes). In practice, interpretations can vary widely in both characteristics.

A two-pronged metric was developed to quantify both characteristics. The Allelic Truth (AT) score is solely concerned with accuracy, and the Genotype Interpretation Metric (GIM) score is solely concerned with precision. These two complementary scores reveal inter- and intra- laboratory variation on the quality of DNA mixture interpretation, and provide a way to zero in on unusually low scor- ing results. These metrics are also a more detailed way to provide labs and individual examiner's specific feedback in potential training and benchmarking scenarios.

5.4 The Allelic Match (AM) Score

The Allelic Match (AM) score measures an interpretation's accuracy, and is broken down into three subscores: the Allelic Truth (AT), Allelic Falsehood (AF), and Inconclusive (INC) scores. They respectively score the number of alleles correctly and incorrectly interpreted, as well as the number loci that are deemed inconclusive (not interpreted).

We use the GIM with the following definitions. In forensic DNA analysis using the ID+ amplification kit, fifteen autosomal loci are typed as comparison points and are denoted L = {D8S1179, D21S11, D7S820, CSF 1PO, D3S1358, THO1, D13S317, D16S539, D2S1338, D19S433, vWA, TPOX, D18S51, D5S818, FGA}. We denote the set of possible alleles (covering at least 95% of the population) for a locus $l \in L$ as A_l. For example, $A_{TPOX} = \{i \mid 6 \leq i \leq 13\}$. We also

define an augmented set of alleles as $A^t_l = A_l \cup$ 'any^t.

A *combination* c_l for a locus l is denoted as a tuple (a, b) where $a \in A_l$, $b \in A^t_l$, e.g. $c_{TPOX} = (8, 11)$ or $(8, any)$. If $a, b \in A_l$, then a, b are ordered such that $a \leq b$. An *interpretation* for a locus is a set of combinations given for a locus l $i_l \in A_l*$, where A_l* is the Kleene star on A_l. If $i_l = E$, it is interpreted as "inconclusive". An example for $l = TPOX$ is $i_{TPOX} = \{(8, 8), (8, 11), (11, 11)\}$.

A *profile* P is a set of interpretations for all fifteen loci, i.e. $P = \{i_l \; \forall l \in L\}$, that describe a single individual's DNA. By definition, a DNA *mixture* M has more than one individual's DNA, and its *deconvolution* D_M is denoted as a set of profiles, with the number of contributors (NOC) determined by the examiner, i.e. $D_M = \{P_j \mid 1 <= j <= NOC\}$. Since every examiner E was given six mixtures to interpret, a full *response* $R_E = \{D_k \mid 1 <= k <= 6\}$ is a set of six mixture interpretation data sets.

A mixture may have multiple genotypes per locus, but only one genotype is the true or correct answer that accurately reflects the DNA of the contributing individuals at each locus. We represent each *contributor* C_k as a set of alleles $C_k = \{c_l* \mid c_l* \text{is the correct combination } c_l, \forall l \in L\}$. Hence, a mixture M with N contributors has a true deconvolution $T_M = \{C_k \mid 1 \leq k \leq N\}$. An examiners NOC is correct if NOC $= |T_M|$. Note that instead of having multiple combinations or INC possible as with other profiles, a contributor has only one combination per locus.

We can now define the allelic match score AM for each interpretation i of a locus as a tuple of three subscores: AT, AF, and INC, i.e. $AM_i = (AT_i, AF_i, INC_i)$.

Given a mixture and a contributor, the true combination of a locus l is denoted $c*_l = (x*, y*)$, where $x*, y* \in A_l$. We score a single combination $c_l = (x, y)$ as

$$AM(c_l) = (AT(c_l), AF(c_l), INC(c_l))$$

$$\begin{cases} (2, 0, 0), & \text{if } (x = x*, y = y*) \\ (1, 1, 0), & \text{if } (x* = x \text{ and } y! = y*) \text{ or } (x = y* \text{ and } y! = x*) \text{ or } (y = x* \text{ and } x! = y*) \text{ or } (y = y* \text{ and } x! = x*) \\ (1, 0, 0), & \text{if } (x = x* \text{ or } x = y*, y = \text{'}any^t \\ (0, 1, 0), & \text{if } (x! = x* \text{ and } x! = y*, y = \text{'}any) \\ (0, 2, 0), & \text{if } (x! = x*, x! = y*, y! = x* \text{ and } y! = y*) \\ (0, 0, 2), & \text{if } c_l = \text{'}inc^t \end{cases} \quad (1)$$

We define a total ordering on the tuple $AM = (AT, AF, INC)$ such that for two AM scores $AM_1 = (AT_1, AF_1, INC_1)$ and $AM_2 = (AT_2, AF_2, INC_2)$, $AM_1 < AM_2$ if $(AT_1 < AT_2)$ or if $(AT_1 = AT_2$ and $AF_1 > AF_2)$. In other words, an AM score is larger than another if its AT score is higher; if the AT scores are equal, then the higher AM score is the one with the smaller AF score. For an interpretation i_l, its Allelic Match score $AM(i_l)$ is calculated as the maximal score of all its combination, i.e. $AM(i_l) = AM(c^t_l)$ such that $c^t_l \in i_l$ and $\forall c_l \in i_l$, $AM(c_l) <= AM(c^t_l)$.

Hence, one can assign an aggregate AM score for every profile, deconvolution and response by summing individual AM scores, applying vector addition to the AM tuple, i.e. for $AM_1 = (AT_1, AF_1, INC_1)$, $AM_2 = (AT_2, AF_2, INC_2)$, $AM_1 + AM_2 = (AT_1 + AT_2, AF_1 + AF_2, INC_1 + INC_2)$. For a profile P, its score is calculated $AM(P) = (AT(P), AF(P), INC(P)) = \sum_{l \in L} AM(i_l)$. The highest possible AT, AF or INC score for a profile of fifteen loci is thirty, since each locus is scored for its two alleles. Similarly, the highest possible AT, AF or INC score is sixty for a two-person mixture and ninety for a three-person mixture.

5.5 The Genotype Interpretation Metric

While the AM score does not penalize entering multiple genotype combinations at each locus in an interpretation, clearly having fewer combinations is preferable to more. In addition to measuring accuracy with the AM score, we also measure precision with the GIM score, which measures the number of genotype combinations generated at each locus.

The GIM score compares the number of combinations in the interpretation against the total number of combinations ($C_{str)}$ available at each locus, calculated from published allele National Center for Biotechnology Information (NCBI) frequency values averaged across African American, Caucasian, and Hispanic groups to cover 99.5% of the (US?) population. At each locus, the most precise interpretation is a single two-allele genotype combination for each donor, which receives a perfect GIM score of 1. An uninterpretable or inconclusive locus, locus dropout, or "Any, Any", is the Inconclusive ("INC") label, which receives a GIM score of 0. Because a combination with an 'Any' does not reduce the possible genotype combinations, the GIM score is reduced by half. Thus, an interpretation that is given with one Any and one allele call is given a score of 0.5, for attempting to determine half the genotype. Combinations that reduce the possible number of genotypes are then scored and generally score between 0.5 and 1. Those with fewer genotype combinations, such as an unrestricted 8, 11, and 12, will score higher than those with additional

genotype combinations provided, such as and unrestricted 8, 11, 12, *and* 13. If given a non- INC locus interpretation $i_l = c_k$, we partition the set of k combinations into those containing an 'any' and those without:

$i_l = C_a \cup C_{wa}$, where $C_a = \{(a, b)|a \in A_l \text{ and } b = \text{'}any^t\}$ and $C_{wa} = \{(a, b)|a, b \in A_l\}$. We measure its GIM score as

$$GIM(i_l) = \begin{cases} 1, & \text{if } C_a = \emptyset \text{ and } |C_{wa}| = 1 \\ 0, & \text{if } i_l = \text{'}INC^t \\ \square^{1-\left|\frac{C_{wa}}{C_{str}}\right|}_{2^{|C_a|}}, & \text{otherwise} \end{cases}$$

As with the AM score, we can assign an aggregate GIM score for every locus, contributor profile, and mixture profile (all contributors in a mixture), deconvolution and response by summing individual GIM scores: the GIM score of a profile $GIM(P) = {}_{l \in L} GIM(i_l)$ is the sum of GIM scores across all loci, the score of a deconvolution $GIM(D_M) = {}_{P_j \in D_M} GIM(P_j)$ is the sum of its profile scores, and the score of the entire response $GIM(R) =$

${}_{D_M \in R} GIM(D_M)$ is the sum of all its mixture deconvolution scores.

5.6 Acknowledgements

We acknowledge the help of Dr. Brenda Held, Cammi Strong, Miles Paca, Douglas White, and Karl Mereus, who assisted in data collection, processing and initial analysis. We also thank the National Institute of Justice for providing funding for the study through an agreement with the Defense Forensic Science Center. The opinions or assertions contained herein are the private views of the authors, and do not reflect the views of the Department of the Army or the Department of Defense. Names of commercial manufacturers or products included are incidental only, and inclusion does not imply endorsement by the authors, DFSC, US Army Criminal Investigation Command, OPMG, Department of Army, or Department of Defense. Unless otherwise noted, all figures, diagrams, media, and other materials used in this manuscript are created by the respective authors and contributors of the research.

References

[1] David L Duewer, Margaret C Kline, Janette W Redman, Pamela J Newall, and DJ Reeder. NIST mixed stain studies #1 and #2: interlaboratory comparison of DNA quantification practice and short tandem repeat multiplex performance with multiple-source samples. *Journal of Forensic Science, 46(5):1199—1210, 2001.*

[2] Mark A Jobling and Peter Gill. Encoded evidence: DNA in forensic analysis. *Nature Reviews Genetics*, 5(10):739, 2004.

[3] SWGDAM interpretation guidelines for autosomal STR typing by forensic DNA testing laboratories, 2010.

[4] SWGDAM interpretation guidelines for autosomal STR typing by forensic DNA testing laboratories, 2017.

[5] Bruce Budowle, Anthony J Onorato, Thomas F Callaghan, Angelo Della Manna, Ann M Gross, Richard A Guerrieri, Jennifer C Luttman, and David Lee McClure. Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. *Journal of Forensic Sciences*, 54(4):810– 821, 2009.

[6] Max M Houck. *Professional Issues in Forensic Science*. Academic Press, San Diego, 2015.

[7] Bruce Budowle, Maureen C Bottrell, Stephen G Bunch, Robert Fram, Diana Harrison, Stephen Meagher, Cary T Oien, Peter E Peterson, Danielle P Seiger, Michael B Smith, et al. A perspective on errors, bias, and interpretation in the forensic sciences and direction for continuing advancement. *Journal of Forensic Sciences*, 54(4):798–809, 2009.

[8] Itiel E Dror and Greg Hampikian. Subjectivity and bias in forensic DNA mixture interpretation. *Science and Justice*, 51(4):204–208, 2011.

[9] Itiel E Dror. Cognitive forensics and experimental research about bias in forensic casework. *Science and Justice*, 52(2):128–130, 2012.

[10] David H Kaye. The design of "the first experimental study exploring DNA interpretation". *Science and Justice*, 52(2):126–127, 2012.

[11] Ate Kloosterman, Marjan Sjerps, and Astrid Quak. Error rates in forensic DNA analysis: definition, numbers, impact and communica- tion. *Forensic Science International: Genetics*, 12:77–85, 2014.

[12] David R Paoletti, Travis E Doom, Carissa M Krane, Michael L Raymer, and Dan E Krane. Empirical analysis of the STR profiles resulting from conceptual mixtures. *Journal of Forensic Science*, 50(6):JFS2004475–6,

2005.

[13] David J Balding and John Buckleton. Interpreting low template DNA profiles. *Forensic Science International: Genetics*, 4(1):1–10, 2009.

[14] Peter Gill, Jonathan Whitaker, Christine Flaxman, Nick Brown, and John Buckleton. An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. *Forensic Science International*, 112(1):17–40, 2000.

[15] Hannah Kelly, Jo-Anne Bright, James Curran, and John Buckleton. The interpretation of low level DNA mixtures. *Forensic Science International: Genetics*, 6(2):191–197, 2012.

[16] Corina CG Benschop, Hinda Haned, Tanja JP de Blaeij, Alexander J Meulenbroek, and Titia Sijen. Assessment of mock cases involving complex low template DNA mixtures: a descriptive study. *Forensic Science International: Genetics*, 6(6):697–707, 2012.

[17] Carll Ladd, Henry C Lee, Nicholas Yang, and Frederick R Bieber. Interpretation of complex forensic DNA mixtures. *Croatian Medical Journal*, 42(3):244–246, 2001.

[18] John S Buckleton, James M Curran, and Peter Gill. Towards understanding the effect of uncertainty in the number of contributors to DNA stains. *Forensic Science International: Genetics*, 1(1):20–28, 2007.

[19] Mark W Perlin, Jennifer M Hornyak, Garett Sugimoto, and Kevin WP Miller. TrueAllele® Genotype identification on DNA mixtures containing up to five unknown contributors. *Journal of Forensic Sciences*, 60(4):857–868, 2015.

[20] Vince L Pascali and Sara Merigioli. 'Stochastic' effects at balanced mixtures: A calibration study. *Forensic Science International: Genetics*, 8(1):113–125, 2014.

[21] Christine A Rakay, Joli Bregu, and Catherine M Grgicak. Maximizing allele detection: effects of analytical threshold and DNA levels on rates of allele and locus drop-out. *Forensic Science International: Genetics*, 6(6):723–728, 2012.

[22] Roberto Puch-Solis, Lauren Rodgers, Anjali Mazumder, Susan Pope, Ian Evett, James Curran, and David Balding. Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters. *Forensic Science International: Genetics*, 7(5):555–563, 2013.

[23] Benôıt Leclair, Chantal J Frégeau, Kathy L Bowen, and Ron M Four-

ney. Systematic analysis of stutter percentages and allele peak height and peak area ratios at heterozygous STR loci for forensic casework and database samples. *Journal of Forensic Science*, 49(5):968–980, 2004.

[24] Jo-Anne Bright, James M Curran, and John S Buckleton. Investigation into the performance of different models for predicting stutter. *Forensic Science International: Genetics*, 7(4):422–427, 2013.

[25] Jo-Anne Bright, Duncan Taylor, James M Curran, and John S Buckleton. Developing allelic and stutter peak height models for a continuous method of DNA interpretation. *Forensic Science International: Genetics*, 7(2):296–304, 2013.

[26] Jo-Anne Bright, Duncan Taylor, Simone Gittelson, and John Buckleton. The paradigm shift in DNA profile interpretation. *Forensic Science International: Genetics*, 31:e24–e32, 2017.

[27] Yolanda Torres, Inmaculada Flores, Victoria Prieto, Manuel López- Soto, Marıa José Farfán, Angel Carracedo, and Pilar Sanz. DNA mix- tures in forensic casework: a 4-year retrospective study. *Forensic Science International*, 134(2-3):180–186, 2003.

[28] TM Clayton, JP Whitaker, R Sparkes, and P Gill. Analysis and interpretation of mixed forensic stains using DNA STR profiling. *Forensic Science International*, 91(1):55–70, 1998.

[29] Mark W Perlin and Alexander Sinelnikov. An information gap in DNA evidence interpretation. *PLOS one*, 4(12):e8327, 2009.

[30] L Prieto, H Haned, A Mosquera, M Crespillo, M Alemañ, M Aler, F Alvarez, C Baeza-Richer, A Dominguez, C Doutremepuich, et al. Euroforgen-NoE collaborative exercise on LRmix to demonstrate standardization of the interpretation of complex DNA profiles. *Forensic Science International: Genetics*, 9:47–54, 2014.

[31] Margaret C Kline, David L Duewer, Janette W Redman, and John M Butler. NIST Mixed Stain Study 3: DNA quantitation accuracy and its influence on short tandem repeat multiplex signal intensity. *Analytical chemistry*, 75(10):2463–2469, 2003.

[32] David L Duewer, Margaret C Kline, Janette W Redman, and John M Butler. NIST mixed stain study 3: signal intensity balance in commercial short tandem repeat multiplexes. *Analytical chemistry*, 76(23):6928–6934, 2004.

[33] Margaret C Kline, David L Duewer, Janette W Redman, and John M Butler. Results from the NIST 2004 DNA quantitation study. *Journal of Forensic Science*, 50(3):1–8, 2005.

[34] John Butler, MC Kline, and MD Coble. NIST Interlaboratory studies involving DNA mixtures (MIX05 and MIX13): Variation observed and lessons learned. *FSI Genetics*, 10.1016/j.fsigen.2018.07.024.

[35] Steven Rand, Marianne Schürenkamp, and Bernd Brinkmann. The GEDNAP (German DNA profiling group) blind trial concept. *International Journal of Legal Medicine*, 116(4):199–206, 2002.

[36] S Rand, M Schürenkamp, C Hohoff, and B Brinkmann. The GEDNAP blind trial concept part ii. Trends and developments. *International Journal of Legal Medicine*, 118(2):83–89, 2004.

[37] Thomas D Gautheir. Detecting trends using spearman's rank correlation coefficient. *Environmental Forensics*, 2(4):359–362, 2001.

[38] Christine S Tomsey, Michael Kurtz, Barbara Flowers, Jeffrey Fumea, Beth Giles, and Sarah Kucherer. Case work guidelines and interpretation of short tandem repeat complex mixture analysis. *Croatian Medical Journal*, 42(3):276–280, 2001.

[39] Peter Gill, Charles H Brenner, John S Buckleton, Angel Carracedo, M Krawczak, WR Mayr, Niels Morling, M Prinz, Peter M Schneider, and BS Weir. DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures. *Forensic Science International*, 160(2-3):90–101, 2006.

[40] David Duewer, Ceceilia Crouse, and John M. Butler. New tools to aid work with STR profile mixtures: mixSTR and Virtual MixtureMaker.
https://strbase.nist.gov/pub_pres/mixSTRposterNIJ2005.pdf