CONFIRMING LONG-RUN DESCRIPTIONS OF NOISE ACQUIRED FROM FORENSICALLY RELEVANT DNA LABORATORY PIPELINES

By

QHAWE BHEMBE

A capstone submitted to the

Graduate School-Camden

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the Degree of Master of Science

Graduate Program in Forensic Science

Written under the direction of

Dr. Catherine Grgicak

And Approved by

_____

Catherine M. Grgicak, Ph.D.

_____

Jason G. Linville, Ph.D.

_____

Kimberlee S. Moran, MSc.

Camden, New Jersey

May 2021

CAPSTONE ABSTRACT

Confirming Long-Run Descriptions of Noise Acquired From Forensically Relevant DNA Laboratory Pipelines

By QHAWE BHEMBE

Capstone Director:

Dr. Catherine Grgicak

In forensic DNA applications, characterizing DNA signal to noise resolution is needed to establish effective and reasonable analytical thresholds (AT). Studies have shown the significance of understanding the behavior of baseline noise such that it can be effectively modeled or can be used to compute an analytical threshold that minimizes Type I and II detection error. Previous studies on noise have described electropherogram noise as well-described by normal, log-normal and Gamma distributions, but there still exist differences of opinion on which distribution class to use.

PROVEDIt single source and mixture samples amplified using the Powerplex® Fusion 6C kit were used for noise characterization. First, we determine whether a normal or log-normal distribution class best fits the noise data. We also ascertain whether noise distributions are significantly different between colors and between loci.

Our findings demonstrate that the log-normal distribution fits the noise data better than the commonly employed normal-class. In addition, noise peak heights were dependent on both dye and locus. Lastly, noise peaks showed an increase in peak height with increase in injection time, suggesting there may be two sources of noise: that originating from instrumentation and that from amplification.

ACKNOWLEDGEMNTS

I would like to thank Dr. Grgicak for affording me the opportunity to work under her guidance and supervision and her continued support as I delve more deeply in the field of forensic DNA analysis. I would also like to extend my sincere thanks to my colleagues at the office, especially Nidhi Sheth for her support and encouragement, and to my committee members Dr. Linville and Ms. Moran for taking the time to review and comment upon this work.

Table of Contents

Table of Figures

List of Tables

1. Introduction

Since it was first described in 1985, forensic DNA analysis has been one of forensic science's most powerful tools, namely for its ability to identify victims, exonerate individuals, provide investigative leads, or connect a perpetrator to the scene[1]. It has been described as the gold standard in forensic science [2, 3]. Human identification using DNA was first described by Alec Jeffreys in [4] where Variable Number of Tandem Repeats (VNTRs), also known as *mini satellites*, allowed for near-perfect individualization. Since the publication of Jeffreys [4], the domain has migrated from using VNTRs to short tandem repeat (STR) genetic markers since they are conducive to PCR amplification due to their relatively short lengths. Not only are STRs viable forensic markers due to their ability to be effectively amplified, their relatively small molecular weight (i.e., base-pair size) is preferable since biological evidence often contains DNA that is degraded or inhibited [5], rendering DNA fragment lengths that are only tens of base pairs. Since forensically relevant STR units typically range from three to five [6, 7], the target range is small, making for relatively efficient PCR, thereby improving the sensitivity of the DNA pipeline.

Evaluation of short tandem repeat (STR) regions using PCR, capillary electrophoresis and laser-induced fluorescence is the prevailing method by which laboratories identify individuals. In many instances, however, samples submitted to the DNA Units contain low-copy numbers of DNA from an unknown number of unknown contributors. The result is data containing low signal-to-noise, high allele dropout rates and substantial levels of signal from interfering contributors; thus, obtaining a metric that effectively summarizes the likelihood a suspect or donor is a contributor to an evidentiary sample can become an arduous task.

Subsequently, laboratories apply filters and thresholds to the designated peaks. Typically, laboratories implement filters to remove likely artifact peaks such as pull-up and -A, while analytical thresholds (AT) are constant values applied across the sample, color or locus to effectively delineate allele signal from noise. The process of setting an analytical threshold requires a scientific and analytical approach where the lab must strike a balance between optimizing the number of true peaks detected while minimizing the rate of false detection [8].

Laboratory decisions such as PCR cycle number and injection parameters significantly influence the information content in the electropherogram as they significantly influence rates of allele dropout. Here, allele dropout is defined as the probability that allele, *a*, did not survive the pre-PCR steps and is, therefore, not present in the PCR tube for amplification. Notably, there are no standards that govern the signal processing choices or filtering protocols; rather laboratories must set up their own operational parameters, suggesting large-scale datasets are required.

With today's technology, the forensic laboratory can reach a limit of detection (LoD) of 1-copy. Notably, an LoD of 1-copy does not imply that all STRs from a single individual will be detected with confidence; indeed, previous work describes the forensic DNA system as one where any number of extracted DNA fragments [5, 9, 10] of allele, *a,* may be transferred to the amplification tube. Thus, if the DNA count is initially of low-copy number, then the step of fractionating the extract volume into one that is amplified and one that is stored can lead to zero copies of allele *a* surviving the PCR steps. This results in 'total allele drop-out' wherein no modification to PCR cycle, load volume or injection time can improve this signal loss. Contrast that with estimations of allele non-detection rates, which is another Type II error produced during DNA processing. In this

instance, the amplifiable DNA fragment consisting of STR *a* is present in the tube, is amplified, but not detected. Unlike the other category of drop-out, the source of this non-detection is the result of low signal-to-noise which can be corrected by increasing laboratory parameters such as PCR cycle number, injection times, injection voltage, and so on.

Issues associated with the interpretation of complex STR signal are well documented and have been highlighted in reports written by the National Academy of Science [2] and, more recently, the President's Council of Advisors on Science and Technology. [3].

The complexity of forensic samples has compelled the forensic community to implement algorithmic solutions for interpretation. Probabilistic genotyping systems, generally, output the likelihood ratio (LR), which has become the prevailing means of communicating the strength of DNA evidence and is the ratio of the probabilities of the evidence given two mutually exclusive hypotheses.

$$LR = \frac{Pr(E|H_p)}{Pr(E|H_d)} \quad (1)$$

Where LR is the likelihood ratio, Pr is the probability, E is the Evidence, H_p is the prosecution's hypothesis and H_d is the defense hypothesis. In equation 1, the likelihood ratio (LR) is calculated using the probability of the evidence, E given the prosecution's hypothesis that the person of interest (POI) is a contributor divided by the defense's hypothesis that some random person (i.e., someone else other than the POI) contributed to the evidence. A likelihood ratio greater than one favors the prosecution's hypothesis while a LR less than one favors the defense's hypothesis.

Despite significant advances over the last two decades, production of a fully objective and automated DNA interpretation pipeline has yet to be developed. Previous studies have shown that the LR can be sensitive to assumptions regarding the number of contributors and the probability of dropout and drop-in [11, 12]. Factors such as PCR settings and differences between allele frequency databases also impact the LR computed using a continuous method [13, 14], while other work has demonstrated LR outcomes differ between semi-continuous and continuous systems. Despite these studies, comparisons between continuous probabilistic systems are uncommon in the literature, though some examples using small in-house datasets do exist. For example, Morimoto et al. [15] compared the continuous system Kongoh to another continuous system, EuroForMix, and demonstrated that for most high-template simple mixtures, the LR outcomes were similar; however differences in LRs from each model were obtained for more complex mixtures, wherein the authors attributed the variation in outputs as "differences in the computational principle of estimating peak height variances". Other reports of inter-software comparisons in the scientific literature recommend the use of multiple software to test one item of evidence [16], though the work used limited datasets and did not replicate the runs. Notably, the authors of [17] disagree with this recommendation to use multiple probabilistic systems in [16], demonstrating that consensus regarding this issue has not been reached.

Work by the authors of [18] using CEESIt, a computational framework that evaluates LR and LR-distribution for continuous models, has demonstrated that seemingly minor modifications to a single probabilistic framework can lead to distinct LR evaluations for some mixtures, while McNevin et al. provide commentary on the study described in [19] and suggest additional studies that evaluate the LRs garnered from distinct,

independent laboratories on the same mixture are still needed to address the remaining concerns on mixture interpretation outlined by PCAST [20]. To ensure full independence from the developer and facile publication-to-publication comparisons, a searchable, freely accessible dataset of forensically relevant signal is a necessity.

The work by Swaminathan et al. [18] demonstrate that it is necessary to rigorously test any new or updated versions of a probabilistic genotyping system to confirm that modifications to any model improve inference, while the work or commentary in [15, 16, 19, 20], among others, demonstrates that expanded inter-software comparisons are justified. We note, however, that burdens associated with the developmental validation demonstrating that new or updated algorithms are better or, at least, concordant with previous results are not the responsibility of the operations laboratories, though they should be able to effectively compare the results reported in the scientific literature. As it is common for developers to test new technologies on in-house data containing, potentially, vastly different information contents, it is challenging for independent parties to compare results across publications. If developers, however, gained access to a common database of samples, publication-to-publication comparisons of the impact of these signal-processing choices on downstream interpretation could result. Further, providing easily accessible forensic DNA data to the broader scientific community affords the added boon of engaging scientists outside of the forensic realm to apply their own expertise to solve these grand challenges in a cost-effective manner. As such, ongoing efforts to develop sizable mixture interpretation databases that can be utilized either by independent parties or the developers themselves are necessary to alleviate unnecessary burdens on forensic laboratories when differences between underlying algorithms exist. Continued growth in

the numbers and complexity of algorithmic solutions is expected, making curation of a public, large-scale, forensically relevant DNA database a necessity.

1.1 Databases in Forensic Science

Forensic DNA interpretation practice has undergone rapid transformation over the past few decades due to innovations in statistical computing and more fundamentally, the availability of data through databases such as CODIS [21] and PROVEDIt [22]. Databases and data are the cornerstone of scientific and forensic inquiry as they allow for the storage and dissemination of valuable forensically relevant information.

The enactment of the DNA Identification Act in 1994 revolutionized forensic DNA with the establishment of the Combined DNA Index System (CODIS); a platform that electronically shares DNA STR profiles of convicted offenders with crime laboratories who are able to generate a forensic DNA profile from a no-suspect case.

Another type of commonly encountered forensically relevant database consists of the allele frequencies of the forensic STRs found within various populations, which are referred to as the 'populations databases'. Though many publications describing allele frequencies across multifarious populations have been produced [23-25], the National Institute of Standards and Technology (NIST) maintains a population database [26] for the most common US populations for public use.

The third type of forensically relevant database of interest to the domain of forensic DNA analysis is that of the research databases consisting of tens of thousands of DNA mixtures. Unlike the other two database types, these databases consist of complex DNA signal garnered from many contributors, with unknown quantities of DNA in any mixture ratio; thus, the signal is not high-fidelity and can, therefore, be used to test the multifarious software systems and procedures claiming to effectively compute the weight-of-evidence

against a person of interest. The largest database of this type is named PROVEDIt for Project Research Openness for Validation with Experimental Data [22], which is available on http://www.lftdi.com.

The database was generated over a 4 year period with its first publication in 2018 [22], and at the time contained over 25,000 short tandem repeat (STR) profiles of varying DNA mass and quality. The dataset includes 1- to 5- person DNA samples, amplified with targets ranging from 1 to 0.007 ng. In the case of multi-contributor samples, the contributor ratios ranged from equal parts of each contributor to mixtures containing up to 99 parts of one and 1 part of the other(s). Additionally, these profiles were generated using a variety of laboratory conditions from samples containing pristine; damaged (i.e., UV-Vis); enzymatically/sonically degraded; and inhibited DNA.

Since it was first described, PROVEDIt data has been used across myriad studies. The PROVEDIt database has already made a demonstrable impact on the forensic DNA community and was utilized by multiple parties to develop machine learning algorithms [27], develop *in silico* laboratories [10, 28], test validation software [29, 30], test number of contributor software [25], test model variants on inference [18], and test impacts of model parameterizations using different data sets [31]. Notably, it has been cited as a pertinent resource by the NIST authors of [32] and by the ISFG authors of [33], demonstrating its value. As an example, Kelly et al. [31] used some 20-single source PROVEDIt profiles amplified with the GlobalFiler kit, injected at 15s for 29 PCR cycles with DNA mass ranging between 0.08 to 0.5ng to parameterize the models of STRmix, a probabilistic genotyping system that interprets DNA evidence. The study also sought to ascertain whether the calibration parameters garnered from the PROVEDIt dataset could be adopted for casework in a laboratory setting that employs the same technology with the

view to demonstrate the robustness of probabilistic genotyping (PG). To do this, they used 71 PROVEDIt mixture samples (2 and 3-contributor mixture samples) amplified with the GlobalFiler multiplex kit and compared these 74 mixture samples that were generated under the same protocol by four laboratories participating in the study (i.e., amplified using GlobalFiler kit, for 29 PCR cycles using 3500 Genetic Analyzer).

In another study [34], PROVEDIt mixture samples amplified using the GlobalFiler multiplex kit were supplied to 174 participants from 42 laboratories to test different versions of the STRmix probabilistic software and test the consistency of subjectively assigning the number of contributors (NOC) to unknown. Specifically, a 4-person and 3-person PROVEDIt whole blood mixture (RD14-0003-44_45_46_47-1;1;4;1-M3a-0.105GF-Q0.8) and (RD14-0003-30_31_32-1;4;4-M2a-0.75GF-Q0.6) were used. Of the 174 participants, 162 of those assigned NOC=4, with 11 submissions assigning NOC=3 for the 4-person mixture. Similarly, 151 participants assigned NOC=3 with the remainder assigning NOC=4 for the 3-person mixture showing relatively high levels of reproducibility among participants but also demonstrating that consistent NOC assignments by subjective evaluation may be difficult to justify when the samples are complex.

In another study [35], 815 PROVEDIt profiles were used to complete a large-scale validation of NOCIt --a software that computes the posterior probability of the NOC given the evidence. As the true number of contributors increases, NOCIt outperformed traditional counting methods that rely on binary decisions regarding presence or absence. Notably, by using the more complex PROVEDIt samples (i.e., NOC=3 to 5), the demonstrated that a range of n may be required for forensic electropherograms with many peaks at each locus.

In another study conducted by Hannig et al. [36], the authors used 2-person and 3-person PROVEDIt data to compare two probabilistic systems that compute the likelihood ratio (LR). They wanted to calibrate their models and investigate the degree to which these models deviated from the likelihood ratio associated with the known ground truth. Their calibration curves revealed that one of the system's calibrations had a negative slope, suggesting that it was overstating the LR in support of the prosecution's hypothesis at higher LRs, while the other overstated the LR by only a maximum factor of 10 across LRs.

In yet a different study [37], the authors used 815 samples consisting of 100 single source samples and 666, 2-person mixture samples described in the supplementary material of [35] to ascertain the efficacy of a method for assigning NOC using a decision tree and compared its performance to three other methods i.e., NOCIt [35], a machine learning approach described in Benschop et al. [38] and a counting method. In brief, like similar databases containing pertinent information [39-41], the forensic domain is reliant on databases containing large numbers of samples. In particular, the PROVEDIt dataset meets this need by making available a large set of data of varying quality and complexities meeting the gap authored in the recent PCAST report [3]. In what follows, we describe a recent extension to the PROVEDIt database in the form of a search tool written in Visual Basic for Applications (VBA) that allows a database user to find raw data for a particular set of samples fitting within a set of pre-defined parameters.

1.2 Noise

In the second part of the work, we use samples found in the PROVEDIt database to validate work that has previously suggested that noise, and its corresponding analytical threshold, is best described and determined using a skewed distribution such as the log-normal class.

The implementation of an AT that is based on statistically sound characterization of the noise is imperative for effective operations. Early forensic DNA work on the subject [8, 42] borrows from the analytical chemistry literature [43] and suggests the AT be determined using the mean and standard deviation of run blanks, or negatives or signal in non-allele positions using the following equation:

$$AT = \mu + k\sigma \quad (2)$$

where AT is the analytical threshold, μ is the mean, *k i*s the numerical factor associated with the desired level of coverage and σ is the standard deviation. Various *k* values have been proposed, and a common choice is *k*=3; however, as described in [44] a *k*-factor of 3 suggests that,

$$P(S \geq AT | T_{c=0} = 0) \leq 10^{-2.801} \quad (3)$$

or the probability that signal, *S*, will exceed the AT will be less than 0.0016 or 1.6 in 1000 peaks given no DNA is present at PCR cycle, *c*=0. If each DNA electropherogram contained 100 noise peaks, then approximately 1 in every 7 electropherograms would be expected to render a noise peak exceeding threshold, impacting hundreds of cases per year. Thus, though *k*=3 likely works well for data where there is a single peak-to-sample relationship (e.g., atomic absorption or UV-VIS at a single wavelength), it is unlikely to provide tractable results for operational labs relying on manual interpretation strategies or probabilistic systems requiring the application of ATs. Other *k*-factors may be deemed more appropriate [43] such as *k*=4, *k*=5, and so on, each increasing the level of coverage against the false-detection of noise peaks. This is not surprising as an analytical threshold set using equation 1 linearly depends on the *k* value selected. For example, Table 1 shows exceedance probability for a range of *k* values.

Table 1:Analytical threshold (AT) for various k-factors for a Gaussian and log-normal noise model. Also shown is log base 10 of the probability that a single noise measurement exceeds that threshold and the log base 10 of the probability that at least one of 100 independent noise measurements exceeds the threshold, therein representing the probability that an electropherogram will exhibit noise exceeding threshold.

k	**AT**	**AT_{ln}**	**$\log(P_{meas})$**	**$\log(P_{profile})$**
3	$\mu+3\sigma$	$\exp(\nu+3\tau)$	-2.8697	-0.8697
4	$\mu+4\sigma$	$\exp(\nu+4\tau)$	-4.4993	-2.4993
5	$\mu+5\sigma$	$\exp(\nu+5\tau)$	-6.5426	-4.5426
6	$\mu+6\sigma$	$\exp(\nu+6\tau)$	-9.0059	-7.0059
7	$\mu+7\sigma$	$\exp(\nu+7\tau)$	-11.8928	-9.8928
8	$\mu+8\sigma$	$\exp(\nu+8\tau)$	-15.1764	-13.1764

The aforementioned treatment notably focuses on mitigating risk associated with false noise detection without considering the Type II error (i.e., allele drop-out). Thus, using similar concepts, the authors of [29] take the aforementioned logic a step further and describe a strategy that assigns the AT such that the sum across γ and $\propto$ is minimized, which represent Type I and Type II errors, respectively. Summing the probability that any noise peak or signal will be in exceedance of that AT when there is no DNA molecule present, and the probability that the signal will be less than AT when one amplifiable DNA copy is present arguably gives the best analytical threshold as it takes into consideration both Type I and II errors. Whatever statistical method is ultimately used to choose the AT, the need for the forensic domain to ensure these methods apply well to multitudes of chemistries and forensically relevant methods is of importance.

An electropherogram is classified as high fidelity if each allele's signal surpasses baseline noise levels. In the forensic context high-fidelity data, therefore, does not necessarily imply that full genetic information is available in the signal since allele non-detection or drop-out can also be driven by ineffective sensitivity levels and the propensity

of target molecules to remain in the extract tube during fractionation of the extract into the PCR tube [45].

As with all signal interpretation schemes reliant on thresholds to arbitrate whether a peak is or is not composed of signal -- e.g., fluorescence -- there is a trade-off between increasing the confidence of acquiring high-fidelity electropherograms and decreasing the probability with which noise will erroneously be detected. There are myriad laboratory decisions that influence the Type I and II signal detection error rate including PCR cycle number, load volumes, injection time, and high pass signal thresholds, also known as analytical thresholds (AT). Though the forensic DNA sphere has witnessed a surge in the development of probabilistic systems, there is no consensus regarding a standard continuous model, or whether applying the same standard model is recommended for all cases within the criminal justice system. Since the choice of model influences the probability calculation, it is possible that changes to the underlying model will result in differences in the LR. In previous work, it was demonstrated that typical forensic pipelines elicit allele peak intensities that comfortably exceed baseline noise levels [10, 29] for the commonly employed GlobalFiler™ pipeline. Since effective implementation of any threshold must necessarily work well in the long-run across multifarious assays and conditions, this work seeks to confirm that the method used to characterize noise as described in [29, 44] is applicable to a chemistry based on modified PCR cycling conditions, amplification reagents and at half-reaction volume.

2 Methodology

2.1 PROVEDIt

In the first part of this work, we developed the PROVEDIt DataSearch Sheet using Visual Basic for Applications (VBA). This script will be integrated into the PROVEDIt database [22] on www.lftdi.com to allow users to easily explore the myriad electropherograms generated using DNA mixtures, of varying DNA qualities and quantities. Access to the data is made publicly available in two formats: 1) CSVs containing the data exported from the peak detection software GeneMapper IDX v1.4 with and without artifacts removed; and 2) the raw data files (.hid or .fsa) catalogued over a series of 31 folders and organized by number of contributors (NOC) that constitute the mixture, the capillary electrophoresis platform, the injection time, the name of the assay, the PCR cycle conditions. Because the raw files are organized in such a way, we developed the classification scheme and search functionality to be able to locate samples on the PROVEDIt database which may be of value to a particular study, test or validation. The PROVEDIt samples were tagged and indexed with the following information highlighted in bold:

Project code	Whole Blood Mixtures/DNA extracts/SUDA
True NOC	1-5 Person mixtures
DNA mass	0.0078-1 ng
Kit	Globalfiler™, Identifiler™ Plus, Powerplex® HS 16, Powerplex® Fusion 6C
Quality Index	0-499
Injection time	5, 10, 15, 20, 24 and 25 s
Instrument	3130/3500

Although the PROVEDIt database is sequestered into a series of .zip folders, a user may use one or more of these tags on the search tool to filter for particular sample combinations. For example, downloading only mixed-inhibited or mixed-degraded

samples is not possible; rather the search tool requires the user to download all mixtures run on a particular platform and use the file path contained in this tool to locate samples of interest rather than having to sift through the run folders manually. In a similar vein, if the research application is one that requires tests on low- or high- template samples, even though these sample types are distributed amongst the run folders, this tool enables the user to trace and filter these specific sample types.

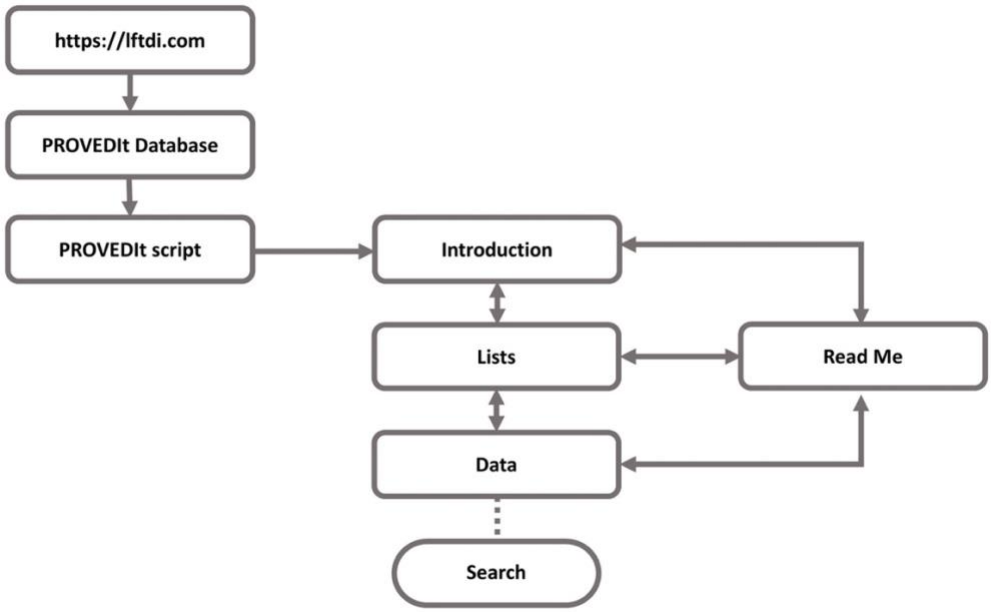


Fig. 1 A schematic representation outlining the steps needed to locate the sample file folder on the PROVEDIt database. 4 modules are contained within the script: Introduction, Read Me, Lists and Data module. The search function contained within the Data module provides an interactive platform where the user may filter information based on some or all of the tags used to index the samples.

In Fig. 1 is a schematic of the steps associated with locating a set of relevant raw files within the database. The introduction module serves as the starting point and provides a brief overview. The *List* module contains an overview of some of the pertinent information (associated with the sample name) the user may select to filter samples of interest. The *Data* module contains a summary of all the STR profiles found in the PROVEDIt database. It also contains the *Search* function. Upon clicking the 'search' button, the user may

complete a search by filtering through the data and making selections within a drop-down list. Clicking *Continue* generates an output workbook containing samples that match the search criteria.

2.2 Noise

2.2.1 Description of the Data

A subset of the PROVEDIt database was used to study background noise. For single source samples, varying DNA masses were targeted and amplified with the Powerplex® Fusion 6C (Promega) assay [6] for 29 cycles and injected into the Applied Biosystems® 3500 Genetic Analyzer following the manufacturer's recommended protocol [6]. Injection times of 5, 15 and 25 s were used as were the following target masses: 0.0078, 0.0156, 0.031, 0.125 and 0.25 ng. Mixture samples (i.e., 2-5 contributors) were amplified in a similar manner; but with different target masses ranging from 0.015, 0.03, 0.06, and 0.125 ng. The total target masses for mixtures, therefore, varied depending upon the contributor ratio and number of contributors, but typically ranged from 0.06 to 0.75 ng. Mixture samples were generated from 21 genotype combinations. Table 2 summarizes the samples used for this noise study. The Powerplex® Fusion 6C kit amplifies 27 STR loci across 6-color channels with high signal to noise.

Table 2: Summary of the Powerplex® Fusion 6C samples used for the noise study. The samples are composed of single source samples and mixture samples (i.e., 2-5 contributors).

				Sample Size per injection		
	Target mass	**PROVEDIt ID**	**Ratio**	**5 s**	**15 s**	**25 s**
1P	0.0078-0.25			329	337	334
2P	0.03-0.75	40_41	1;1	24	22	20
		42_43	1;4			
		44_45	1;9			
3P	0.045-0.75	30_31_32	1;1;1	22	21	24
		46_47_48	1;4;1			
		49_50_29	1;4;4			
4P	0.06-0.75	32_33_34_35	1;1;1;1	35	41	36
		33_34_35_36	1;1;2;1			
		36_37_38_39	1;1;4;1			
		37_38_39_40	1;1;9;1			
		40_41_42_43	1;2;2;1			
		44_45_46_47	1;4;4;1			
		48_49_50_29	1;4;4;4			
		50_29_30_31	1;9;9;1			
5P	0.075-0.75	30_31_32_33_34	1;1;1;1;1	39	35	37
		31_32_33_34_35	1;1;2;1;1			
		33_34_35_36_37	1;1;2;4;1			
		35_36_37_38_39	1;1;2;9;1			
		36_37_38_39_40	1;1;4;1;1			
		43_44_45_46_47	1;4;4;4;1			
		48_49_50_29_30	1;9;9;9;1			

2.2.1.1 Data Analysis

Microsoft excel CSV files for each of the profiles were downloaded for all 3 injection conditions. The data had previously been filtered with the removal of pull-up, dye blob, spikes and minus A as described in [22]. All peaks originating from pull up or minus a – i.e., incomplete adenylation -- were filtered. For peaks to be classified as pull up, the peak in question had to be in the same position ($\pm$0.3 bases) as the allelic peak in another

color channel and have a peak height of 5% or less of the allelic peak. Further, if a peak fell between two adjacent allelic peaks in different dyes and had a 'plateau-like' shape, then the peak in question was classified as complex pull up. A peak was determined to be minus A if it was one base shorter ($\pm$0.3 bases) than the allelic peak [46]. Ground truth information about the samples was known; that is, genotypes of the individuals who contributed to the samples was provided with the data, which made the signal positions of noise and stutter positions within the electropherogram known. Allelic signal, off-ladder and signal originating from stutter positions (i.e., N-1 and N+1) were removed. Amelogenin and the STRs found on the Y- chromosome, DYS391, DYS570 and DYS576 were excluded from further analysis. Once the noise data were prepared, a series of statistical evaluations were conducted.

Fig. 2 is a visualization of four representative STR loci of a single-source sample. More specifically, it depicts the peak heights associated with the sample when amplified using 0.25 ng of DNA, where the y-axis represents peak height in relative fluorescent units (rfu) and the x-axis depicts the base pair size of the amplified STR fragment. In this representation, off-ladder (OL) peaks are not depicted. As shown in this figure, electropherograms typically render the following two main features:

1. The peak height of allele peaks are significantly greater than that of stutter or noise.
2. Stutter peaks are significantly larger than noise.

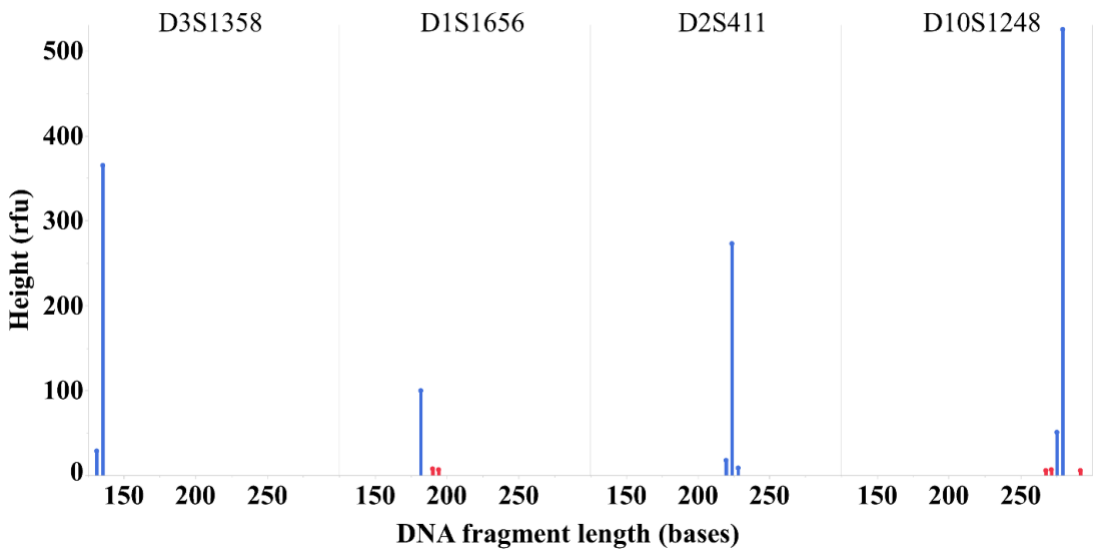


Fig. 2 An example of an electropherogram depiction showing allelic peaks and stutter peaks (both forward and reverse) in blue and noise peaks in red for 4 loci within the blue color channel. Noise peaks appear randomly within the loci. For example, D3S1358 and D2S411 contain no noise peaks while D1S1656 and D10S1248 exhibited 2 and 3 noise peaks, respectively.

If EPG data is taken to be a composition of noise; allele and stutter; and the positions of allele and stutter are known, then it follows that filtering the signal categorized as allele and stutter from the total signal will result in data that can only originate from noise. It is these data we explore below.

2.2.2 **Color channel effects on noise**

Noise peak height measurements were grouped by color to explore if noise is independent of color. If not, it would suggest that any lower-bound signal threshold – i.e., AT --used in forensic DNA analysis should be determined on a per color basis rather than on a profile basis. To accomplish this, however, we must first determine which statistical test to use, which requires an examination as to whether noise follows a normal distribution.

Noise peak heights for 334 single source profiles injected for 25 s were grouped by color and graphed using JMP®PRO v15.0 to generate 5 histograms corresponding to noise measured across the different dyes. Each was qualitatively investigated for normality by subjective evaluation of normal quantile plot, which would indicate that the data is normally distributed if the points follow a diagonal line.

In cases where the data are not normally distributed, hypothesis tests based on non-parametric statistics are employed. To statistically test the hypothesis that the data are normal, we apply the Anderson-Darling test with a decision threshold of 0.05. To test if the median noise values are significantly different between colors, we apply the Kruskal Wallis test with a *p*-value threshold of 0.05.

2.2.3 **Noise by locus**

Next, we delve further and investigate whether noise peak measurements are independent of locus. As before, the noise data were sorted on a per-locus basis and boxplots for noise measurements per locus across the different color channels were plotted to visually inspect the data. We further zoom in per locus to investigate a distribution class that better describes noise. Normal quantile plots were used to visually analyze for normality. Fitted normal and log-normal distributions were also graphed on JMP®PRO v15.0 to investigate noise class distribution on a per locus basis.

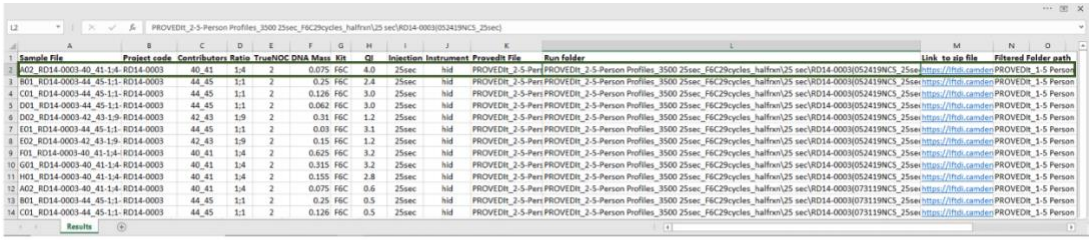
For statistical testing, we evaluated the null hypothesis that the data followed a normal distribution. The Anderson-Darling test was used to test for the goodness of fit for both the normal and log-normal distributions and the Akaike Information Criterion (AIC) was used to select the distribution class that best described the data. The assumption of the Akaike Information Criterion is that a class distribution with the smallest AIC value is a better model. To test the null hypothesis that noise is independent of locus, we used the Kruskal

Wallis test and permutation test in JMP®PRO v15.0. The permutation test randomizes the data points between the loci and the test gives a *p*-statistic which can be used to accept or reject the null hypothesis.

3 Results

3.1 PROVEDIt

The PROVEDIt database consists of many samples that contain signal garnered from samples from many contributors of low- and high- template and of varying quantity. In its current state, the data is sequestered in multiple .zip files and the development of this searchable tool makes it possible to locate the run folder containing samples of interest. This tool makes use of pertinent information associated with the sample name and applies the tags i.e., project code, TrueNOC, DNA mass, Kit, Quality Index, Injection time and Instrument to filter samples.



Sample File	A02_RD14-0003-40_41-1;4-M3S30-0.075F6C-Q4.0_01.25sec.hid
Project code	RD14-0003
Contributors	40_41
Ratio	1;4
True NOC	2
DNA mass (ng)	0.075
Kit	Powerplex ® Fusion 6C (F6C)
Quality Index (Q.I.)	4.0
Injection time	25 s
Instrument	3500 Genetic Analyzer (hid)
PROVEDIt file (.zip file)	PROVEDIt_2-5-Person Profiles_3500 25sec_F6C29cycles_halfrxn.zip
Run folder	PROVEDIt_2-5-Person Profiles_3500 25sec_F6C29cycles_halfrxn\25 scc\RD14-0003(052419NCS_25scc)
Link to zip folder	https://lftdi.camden.rutgers.edu/repository/PROVEDIt_2-5-Person Profiles_3500 25sec_F6C29cycles_halfrxn.zip
Filtered folder path	PROVEDIt_1-5 Person CSVs Filtered/PROVEDIt_1-5-Person CSVs Filtered-3500_F6C29cycles_halfrxn/2-5-Persons/25sec

Fig. 3 shows an example worksheet of search results generated by the PROVEDIt search tool. Column L (run folder) shows the file path of the .zip folder containing the samples of interest once downloaded from the PROVEDIt website. Highlighted in green (row 2) is an example of a sample that is split according to the different columns of the worksheet and we further zoom in on the sample to show the information contained in each column. The sample name is listed in column A, with the number of contributors listed in column E.

In Fig. 3, a sample worksheet generated by the PROVEDIt searchable tool is shown, and contains a summary of the 2-Person mixture PROVEDIt amplified with the Powerplex® Fusion 6C kit, injected at 25s. Below the sample worksheet is an example using sample file (A02_RD14-0003-40_41-1;4-M3S30-0.075F6C-Q4.0_01.25sec.hid); describing the pertinent information contained in each column associated with the naming convention of the samples. Here, the metadata associated with each sample file is broken down into 14 columns. As many as 7 of the 14 are used in filtering samples within the interface of the PROVEDIt tool. The organization of the output sheet is similar to the parent tool with the addition of two extra columns containing the run folder name (column L) and Hyperlink to that folder (column M). Once the run compressed .zip folder containing the

samples of interest is downloaded, the user can use the file path (column L) provided in the sheet to navigate through the subfolders.

3.2 Noise

A substantive body of analytical chemistry literature [47-49] and forensic DNA analysis [8, 44, 50] discuss the ways in which noise could be used to determine the high-pass signal threshold demarcating the line between confidently labelling a peak as containing allele or stutter signal. However, the forensic EPG can be described as a vector of triples containing information on allele, base-pair size and height. The number of possible allele positions is innumerable; however, barring any mutations, the complete set of alleles that can describe a sample can be defined by the number of common STRs found in the population-at-large. The common forensically relevant STR alleles are well studied [26, 51]. In the case of Powerplex® Fusion 6C, the list of common STR alleles for the 27 STR loci are available in the manufacturer's 'bin file' [6]. Of the noise locations studied in this work, we focus on the non-zero noise peaks.

3.2.1 Color channel effects on noise

Numerous studies have evaluated the impact of the color channel or fluorophore wavelength on the intensity of the noise. For example, the authors of [44] evaluated baseline noise for 643 single source STR profiles from 60 donors amplified with the Identifiler™ Plus assay and 303 single source STR profiles from 48 donors amplified with the Powerplex® 16 HS kit. Samples from both kits had a target masses ranging from 0.0078-0.25 ng. The Identifiler™ Plus kit samples were injected for 10 s for 29 cycles while the Powerplex® 16 HS samples were injected for 32 cycles. For this study, the authors performed the G-test of independence to investigate whether a noise description by dye was sufficient and their findings were a confirmation to the existing literature that noise

is dependent on fluorophore channel. The authors then sought to investigate whether a per locus noise description was justified, and once again performed the G-test of independence to evaluate the null hypothesis that noise is independent of locus. This hypothesis testing revealed that noise is dependent on locus. According to the investigated noise class distribution for this study, the log-normal distribution provided a better description of noise over the Gaussian distribution.

Similarly, Krane et al. in [42] evaluated noise for 50 STR runs from a negative and positive controls and a reagent blank run on a 310 Genetic Analyzer using the Profiler® Plus assay. Results from this study suggested that noise from the negative and positive controls followed a normal Gaussian distribution, though the conclusion was made by subjective evaluation of a histogram rather than a full statistical analysis of the data. In contrast, [8] showed that a log-normal distribution was a better fit for the negative samples. In every study, regardless of kit or instrument, the intensity of noise was found to be dependent upon the color channel for which the STR is detected. In this work, we continue the work in [44] and also evaluate the noise independence on color for the PowerPlex® Fusion 6C kit, a next-generation STR assay recently introduced to operations. In Fig. 4, **a**) is a boxplot of the noise peaks quantized by the 337 samples amplified at 0.0078, 0.0156, 0.31, 0.125, and 0.25 ng and injected for 25 s. In this figure, we notice the following features: Firstly, the medians of the noise peaks between the colors are relatively unchanged, ranging from 7-13.

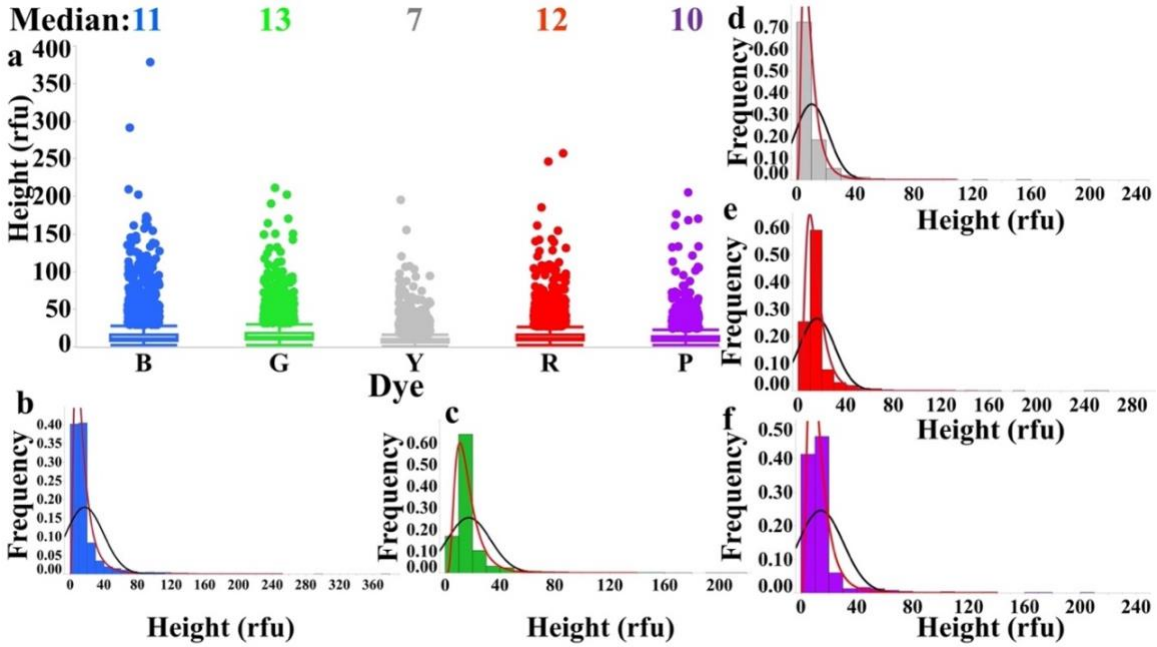


Fig. 4 ***a)*** *Boxplots for the single source samples injected for 25 s showing noise peak heights across color channels*: [■] *blue dye,* [■] *green dye,* [■] *yellow dye,* [■] *red dye and* [■] *purple dye. The yellow dye was changed to grey for contrasts.* ***b-f)*** *Histograms showing the noise distribution for the different color channels;* [■] *blue,* [■] *green,* [■] *yellow,* [■] *red and* [■] *purple respectively. Fitted normal (black) and fitted log-normal (red) distributions were plotted to identify a class distribution that better describes the data. The log-normal distribution visually provides a better fit.*

Despite the similarity in median, the distributions between color are different from one another. This is obvious, for example, when comparing the blue to yellow color channel, where the blue channel has a markedly larger number of noise peaks in the hundreds of RFUs, while the yellow channel does not. Indeed, the number of noise peaks greater than 100 RFU was at its maximum -- i.e., 195 rfu -- in the yellow channel and at its maximum -- i.e., 378 rfu -- in the blue channel. To quantitatively test the hypothesis that the noise within color channel are of the same distribution, we apply statistical hypothesis testing. Before doing so, however, we must first examine the distribution-class to estimate whether the data follow a normal distribution. The justification for analyzing baseline noise per dye is well documented in literature [44, 52] which lays the foundation for hypothesis testing of this study.

To evaluate baseline noise independence between channels, the Kruskal Wallis test was performed. This test statistic does not assume central tendency associated with normality, but instead, the data are ranked from smallest to largest. The resultant probability that the test-statistic is as large or larger than the one obtained with this data was $<10^{-4}$, indicating that the null hypothesis of independence can be rejected. As previously expressed, a significant number of the noise peaks are larger than expected which brings to bear the importance of describing noise using a systems' view, as opposed to an instrumental view.

Interpretation of forensic DNA analysis occurs after a series of steps that impact downstream interpretation. Generally, the DNA pipelines includes extraction, amplification, electrophoresis, detection, analysis and interpretation. In this instance interpretation connotes the processes by which forensic scientists determine the likelihood ratio (LR), which is a ratio of probabilities, i.e., probability of observing the evidence given the suspect contributed over the probability that they did not contribute questioned material. This can be accomplished in a binary way-classifying the evidence as being explained or not explained by a set of genotype combination, or in a continuous fashion by assigning the probability of the evidence given a certain genotype combination. Whether the interpretation utilizes a set of heuristics or uses a continuous approach, the misclassification of noise as a potential allele requiring evaluation necessarily increases the number of genotype combinations requiring consideration.

Fig 5 is an illustrative example of the potential negative impact of a false noise detection on a simple case assuming manual or binary interpretation is performed. As previously stated, the weight of evidence in the context of DNA interpretation is given within the likelihood ratio framework. The way in which the likelihood ratio (LR) is

computed involves making an assumption on the number of contributors (NOC). Several methods have been devised for assigning the NOC and they include the Maximum Allele Count (MAC), which counts the number of peaks crossing AT, divides by two and rounds up. In this method the NOC is estimated by equating the minimum number of people that might explain the evidence to the actual number. It is under this NOC assumption that the LR is calculated.

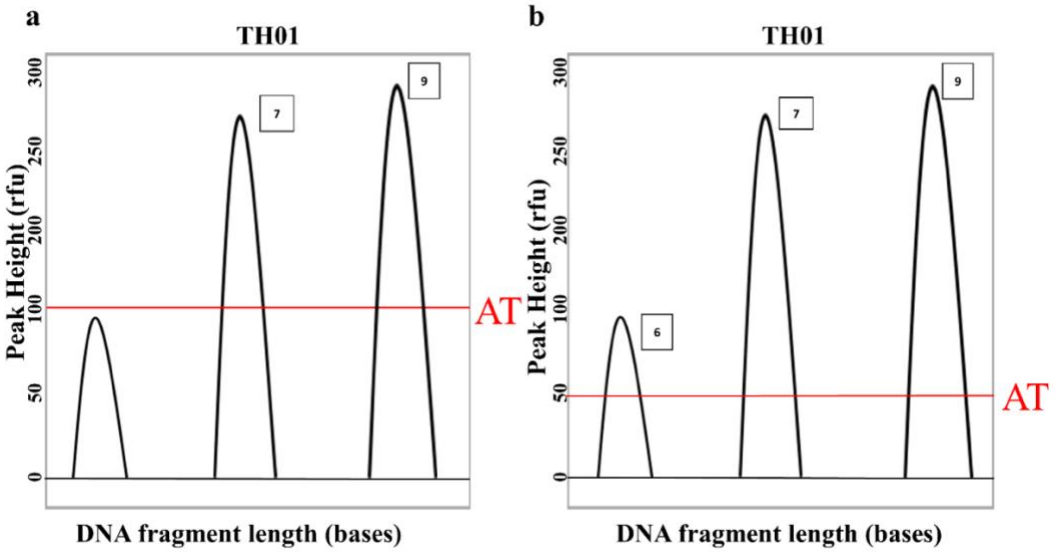


Fig. 5a. Depiction of an electropherogram of the Th01 locus with the AT set at 100 rfu while b) has an AT set at 50 rfu. The figure shows the false detection of noise in b where the AT is set nearer to the baseline.

In Fig. 5**a**, only 2 peaks (7 and 9) would be considered for interpretation as the allele 6 falls below AT=100 rfu; meaning NOC=1. As depicted in Fig. 5**b**, setting the AT closer to baseline – i.e., 50 rfu – results in an NOC=2 as the noise peak (allele 6) was falsely classified as containing allelic signal, illustrating the effects of signal misinterpretation on inference.

Though effects of information content on forensic inference has not been systematically studied on a large scale, a National Institute of Standards and Technology

(NIST) 2013 mixture study [32] demonstrated there is a great deal of variation in interpretation outcomes across the United States. In one scenario, 100 laboratories participated in mixture analysis of a complex mixture containing ≥ 3 contributors. Interestingly, when the mixture was compared against suspects A, B, and C, 6 laboratories excluded suspect C, 3 included A & B while suspect C was neither included or excluded, 21 stated A, B, nor C could be included/excluded, and 70 laboratories included all suspects and provided statistics that communicated the weight of evidence, wherein the match statistic for suspect C ranged from 1 in 9 to 1 in 344,000. The authors of the study suggested that if laboratories use the same AT, then one would expect to see similar results [32], however, implementing the same AT across laboratories is unlikely to lead to consistency since signal intensity differences between laboratory pipelines exist and are dependent upon factors including PCR cycle number, injection time, and the capillary electrophoresis platform itself [53].

Peak height is dye dependent as observed in [44]. It is interesting to note that some dyes have noise peak heights in the hundreds. This may in part be explained by the definition of noise we provided for this study as we only defined stutter as signal residing in the N-1 and N+1 positions. This notion is supported by the findings of [44], where they discovered that the removal of noise peaks in the N-2 position reduced the largest noise peaks. Given that our DNA mass ranged from 0.0078-1 ng, it is not surprising to observe these larger noise peak heights. Visual inspection of the histograms in Fig 4 shows a higher probability of noise peaks fall below 40 rfu. This finding mimics an earlier probabilistic study of noise [44] where most of their noise peaks were also under 40 rfu.

In Fig 4 **b-f** we show a visual comparison of the distribution of baseline noise measurement using histograms for the different dyes. The fitted normal and log normal

distributions are compared. Qualitative analysis shows that the log normal distribution provides a better fit to the data than the normal distribution. An Anderson Darling goodness of fit test was conducted and the Akaike information criterion (AICc) was used to select the best class distribution. The model with the lowest AICc was presumed to be a better fit.

Table 3 The Akaike information criterion (AICc) of the Anderson-Darling test. The log-normal distribution has the minimum AICc, which provides a better description for the noise distribution than the normal assumption across all color channels.

Dye	Distribution	AICc	BIC	Anderson-Darling Statistic	p-value
B	Log-normal	20009	20021	69	<0.0001
	Normal	25250	25262	428	<0.0001
G	Log-normal	20736	20748	67	<0.0001
	Normal	25020	25032	386	<0.0001
Y	Log-normal	15255	15267	68	<0.0001
	Normal	19546	19557	368	<0.0001
R	Log-normal	24385	24397	72	<0.0001
	Normal	29776	29788	471	<0.0001
P	Log-normal	10071	10082	36	<0.0001
	Normal	12797	12808	237	<0.0001

In summary, the lognormal distribution of baseline noise is a better class of distribution, which is consistent with the findings of the G-test of independence where the authors of [44] asserted that baseline noise from the Identifiler® Plus and Powerplex® kits followed a log-normal distribution. Interestingly the log-normal class is distinct from the Gaussian assertions of [8] and assumptions of [54], though explained by the subjective assessment that formed the basis of the former's conclusion and unique implementation of the noise model in the latter.

3.2.2 Noise distribution per locus

Once we established that noise peak heights are dependent on channel, we investigate whether noise ought to be characterized on a per locus basis. To accomplish this, we studied the distribution of noise on a per locus basis within the same dye. Visual inspection of the boxplots for all loci in Fig 6 highlights the differences in noise peak height measurements. Take for example the blue channel which has six loci. The distribution of noise from one locus is visually different to another. The D10S1248 has a maximum noise peak height of 378 rfu while D2S441 has its maximum at 46 rfu. This trend is similar throughout all color channels.

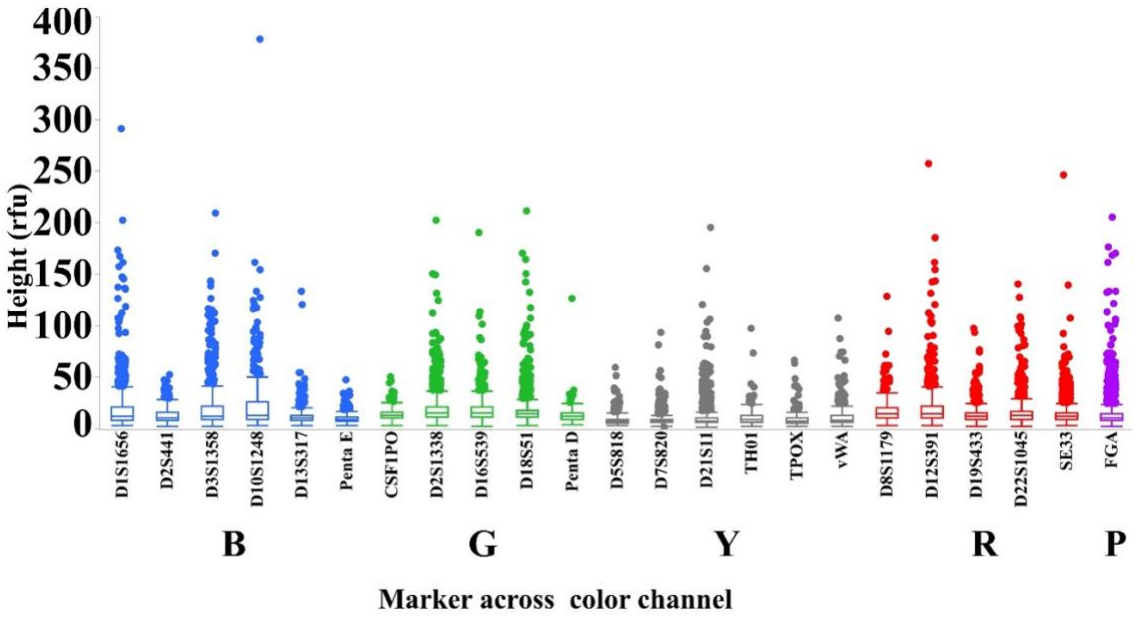


Fig. 6: Boxplots of noise peaks, separated by locus, of single source samples injected for 25 s. The blue channel had the highest non-zero noise peaks with maximum height at 378 rfu. The colors indicate the fluorescent dye color at that locus: [■] *blue,* [■] *green,* [■] *yellow,* [■] *red dye and* [■] *purple. The yellow dye was changed to grey for contrast.*

Histograms and Normal quantile plots were used to evaluate which distribution class best describes the noise data within each locus. Fig. 7**a** and 7**b** show histograms for the D1S1656 and D31S1358 loci, respectively as representative loci. A fitted normal (in

black) and log-normal (red) distributions is also shown. In both plots it seems unlikely that a normal distribution will be a better descriptor of the noise at each locus than the log-normal class. The Anderson-Darling test was used to evaluate this quantitatively by testing the null hypothesis that the data are normally or log-normally distributed. Based on the AICc, a log-normal distribution is a better fit for noise data even on a per locus level. To evaluate the null hypothesis that baseline noise measurements are independent of locus, a permutation test in JMP®15 was performed. This randomization test of independence gave a p-value of 10^{-4}, which led to the rejection of the null hypothesis that noise is interchangeable between loci within a color channel.

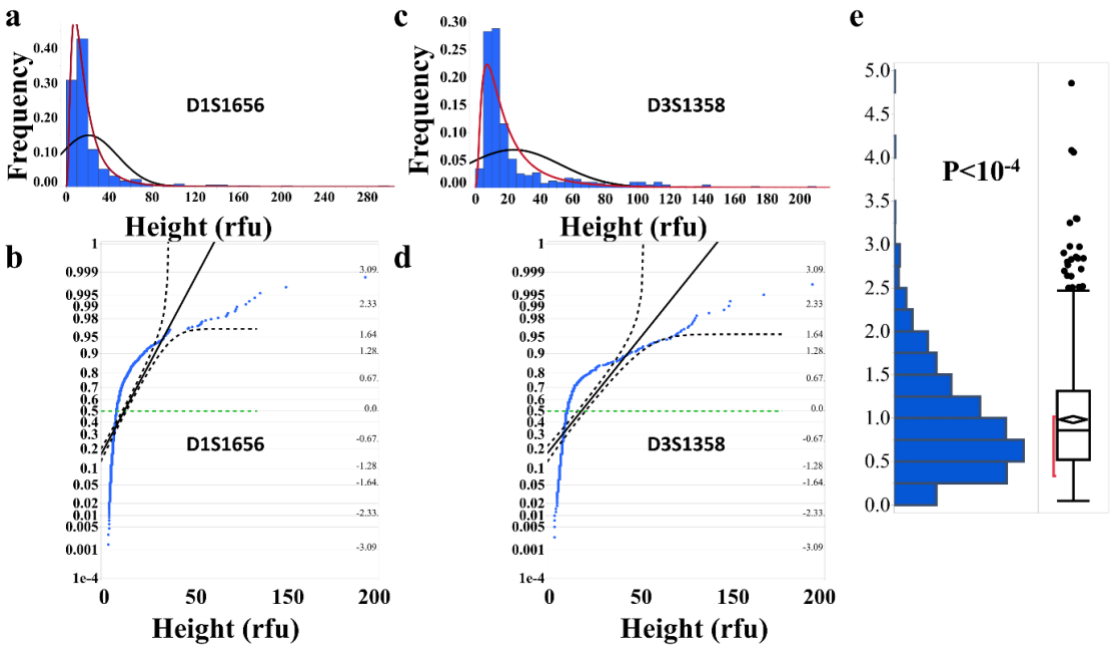


Fig. 7 ***a*** *Histogram for the noise distribution at the D1S656 locus (within the blue dye). A fitted normal (black) and a fitted log-normal (red) distribution were plotted to identify which best explains the data.* ***b*** *A normal quantile plot for the D1S656 locus. The noise data points (in blue) do not fall along the diagonal line (black), suggesting the data is not normal.* ***c*** *shows a histogram of noise contained in the D3S1358 locus. The fitted normal (black) and log-normal (red) distributions are also shown.* ***d*** *depicts a normal quantile plot for D3S1358 locus.* ***e*** *shows a simulated distribution from the blue dye permutation test and the corresponding p-value demonstrating noise is not interchangeable between loci, where the test statistic is the F-value, and is shown on the y-axis.*

In Fig 7, **e** is shown a simulated distribution from the permutation test, using the F-value as the test statistic. The small p-value lends to a rejection of the null hypothesis, suggesting baseline noise measurements are locus dependent; that is, they are characteristic to that locus and cannot be interchanged between loci despite being detected using the same dye channel for detection. These findings corroborate the findings in [44] as they also found that noise is also locus-dependent for the Identifiler^{TM} Plus and Powerplex® kits, explaining the decision of [54] to model noise on a per-locus basis.

Bregu et. al [8] reported an interesting finding regarding the behavior of noise as the injection time increased. In the study, they reported that there were no significant differences observed with increasing the injection time. For the current study, we explore this notion for the PowerPlex® Fusion 6C kit. In Fig. 8 **a,** injection protocols for single source samples across all dyes were compared. Contrary to the study cited above, visual inspection of the boxplots shows an increase in noise peak height with increase in injection time. Quantitatively, we used the Kruskal Wallis test to evaluate the null hypothesis that there was no significant difference with an increase in injection time. The reported *p*-value ($p<10^{-4}$) compels us to reject the null hypothesis and assume that the alternative hypothesis is true.

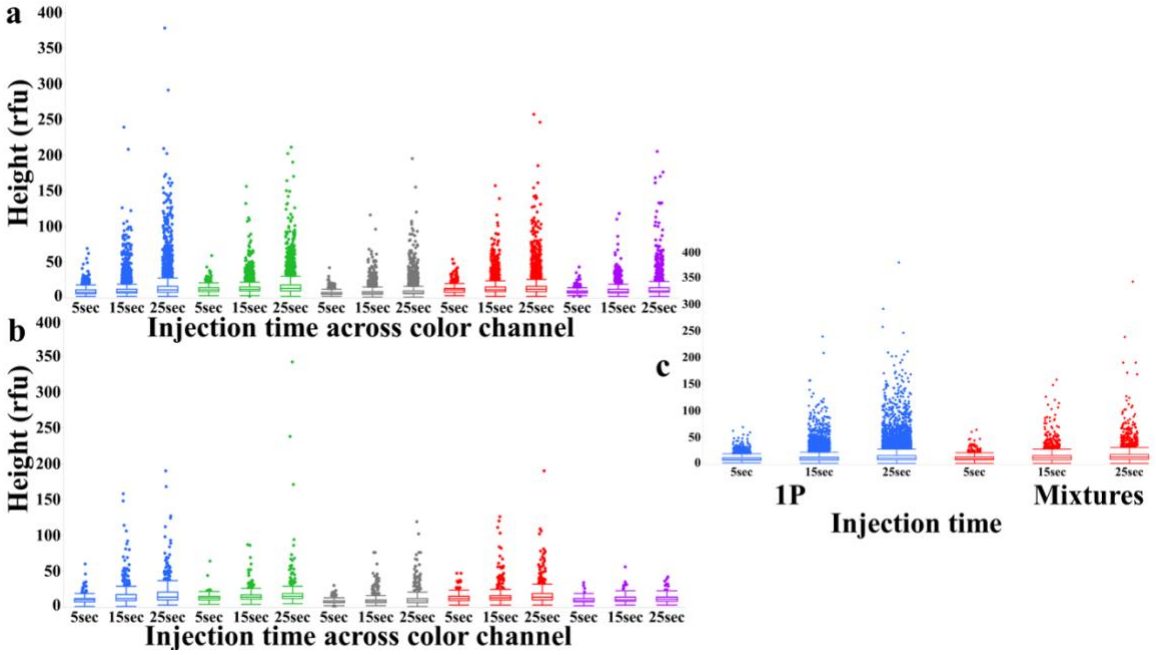


Fig. 8 ***a*** *shows boxplot for noise peaks across the different dyes for the 3 injection protocols (5, 15 and25s).* ***b*** *shows boxplots for all mixture samples injected for 5, 15 and 25 s across the different dyes.* ***c*** *shows boxplots for noise peak height measurements for single source and mixture samples compared side by side for all 3 injection times. Single source samples have higher noise peaks compared to mixture samples.*

In Fig. 8 **b)**, we performed a similar analysis with mixture samples injected at 5, 15 and 25 s as with the single source samples. We observed a similar trend where an increase in injection time resulted in increases in peak heights. In Fig. 8 **c)** we provide a side-by-side comparison of single source samples and mixture samples with respective noise peak heights across the different injection times. The boxplots show an interesting feature where single source samples have higher noise peaks compared to mixture samples. Once again, we statistically evaluated this using the Kruskal Wallis test, and a *p*-value less than 10^{-4} was obtained, suggesting that there was a significant difference in noise peak height between 1-person samples and mixture samples.

4 Conclusion

In the current study, we evaluated the noise distribution for the Powerplex® Fusion 6C kit on a per color channel basis and we observed that noise peaks were different between dyes. In our analysis, the log-normal distribution was a better fit to describe the data. We also demonstrated that noise peak heights were not independent of locus. Noise studies are important in forensics as there is a strong need to optimize our test methods to garner some consistency in our interpretation. The authors of [18] showed the discrepancy in changing the parameters of four models they were testing. In that study, a model where noise had been described as normal has a LR falling below 10^{-7} while a log-normal distribution of noise gave a LR>1. This further reiterates the point that understanding the behavior of noise is important to make sound interpretation protocols whether the method uses analytical thresholds or probabilistic genotyping systems.

Overall, we have confirmed the work by previous studies [44, 50] that a log-normal distribution is a better model in characterizing baseline noise. We further observe a noise dependence on a locus basis, therein justifying locus-dependence inference procedures.

References

[1] J. M. Butler, "Advanced Topics in Forensic DNA Typing: Methodology," (in English), *Advanced Topics in Forensic DNA Typing: Methodology*, Book pp. 1-680, 2012. [Online]. Available: <Go to ISI>://WOS:000321699600024.

[2] C. Natl Res, "Strengthening Forensic Science in the United States: A Path Forward," (in English), *Strengthening Forensic Science in the United States: a Path Forward*, Book pp. 1-328, 2009. [Online]. Available: <Go to ISI>://WOS:000344871900014.

[3] P. s. C. o. A. o. S. a. Technology, *Report to the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-comparison Methods. References*. Executive Office of the President of the United States, President's Council …, 2016.

[4] A. J. Jeffreys, V. Wilson, and S. L. Thein, "Hypervariable 'minisatellite' regions in human DNA," *Nature*, vol. 314, no. 6006, pp. 67-73, 1985-03-01 1985, doi: 10.1038/314067a0.

[5] S. Karkar, L. E. Alfonse, C. M. Grgicak, and D. S. Lun, "Statistical modeling of STR capillary electrophoresis signal," (in English), *Bmc Bioinformatics*, Article; Proceedings Paper vol. 20, p. 12, Dec 2019, Art no. 584, doi: 10.1186/s12859-019-3074-0.

[6] M. G. Ensenberger *et al.*, "Developmental validation of the PowerPlex (R) Fusion 6C System," (in English), *Forensic Science International-Genetics*, Article vol. 21, pp. 134-144, Mar 2016, doi: 10.1016/j.fsigen.2015.12.011.

[7] J. M. Butler, "Genetics and Genomics of Core Short Tandem Repeat Loci Used in Human Identity Testing," *Journal of Forensic Sciences*, vol. 51, no. 2, pp. 253-265, 2006-03-01 2006, doi: 10.1111/j.1556-4029.2006.00046.x.

[8] J. Bregu, D. Conklin, E. Coronado, M. Terrill, R. W. Cotton, and C. M. Grgicak, "Analytical Thresholds and Sensitivity: Establishing RFU Thresholds for Forensic DNA Analysis,," *Journal of Forensic Sciences*, vol. 58, no. 1, pp. 120-129, 2013-01-01 2013, doi: 10.1111/1556-4029.12008.

[9] R. G. Cowell, "A unifying framework for the modelling and analysis of STR DNA sample arising in forensic casework," 2018-02-27T13:05:16 2018.

[10] K. R. Duffy, N. Gurram, K. C. Peters, G. Wellner, and C. M. Grgicak, "Exploring STR signal in the single- and multicopy number regimes: Deductions from an in silico model of the entire DNA laboratory process," (in English), *Electrophoresis*, Article vol. 38, no. 6, pp. 855-868, Mar 2017, doi: 10.1002/elps.201600385.

[11] C. C. G. Benschop, H. Haned, L. Jeurissen, P. D. Gill, and T. Sijen, "The effect of varying the number of contributors on likelihood ratios for complex DNA mixtures," *Forensic Science International: Genetics,* vol. 19, pp. 92-99, 2015-11-01 2015, doi: 10.1016/j.fsigen.2015.07.003.

[12] J. A. Bright, J. M. Curran, and J. S. Buckleton, "The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation," (in English), *Forensic Science International-Genetics,* Article vol. 12, pp. 208-214, Sep 2014, doi: 10.1016/j.fsigen.2014.06.009.

[13] D. Taylor, J.-A. Bright, J. Buckleton, and J. Curran, "An illustration of the effect of various sources of uncertainty on DNA likelihood ratio calculations," *Forensic Science International: Genetics,* vol. 11, pp. 56-63, 2014-07-01 2014, doi: 10.1016/j.fsigen.2014.02.003.

[14] J. A. Bright, K. E. Stevenson, J. M. Curran, and J. S. Buckleton, "The variability in likelihood ratios due to different mechanisms," (in English), *Forensic Science International-Genetics,* Article vol. 14, pp. 187-190, 2015, doi: 10.1016/j.fsigen.2014.10.013.

[15] S. Manabe, C. Morimoto, Y. Hamano, S. Fujimoto, and K. Tamaki, "Development and validation of open-source software for DNA mixture interpretation based on a quantitative continuous model," *PLOS ONE,* vol. 12, no. 11, p. e0188183, 2017-11-17 2017, doi: 10.1371/journal.pone.0188183.

[16] E. Alladio *et al.*, "DNA mixtures interpretation - A proof-of-concept multi-software comparison highlighting different probabilistic methods' performances on challenging samples," (in English), *Forensic Science International-Genetics,* Article vol. 37, pp. 143-150, Nov 2018, doi: 10.1016/j.fsigen.2018.08.002.

[17] D. A. Taylor, J. S. Buckleton, and J.-A. Bright, "Comment on “DNA mixtures interpretation – A proof-of-concept multi-software comparison highlighting different probabilistic methods’ performances on challenging samples” by Alladio et al.," *Forensic Science International: Genetics,* vol. 40, pp. e248-e251, 2019-05-01 2019, doi: 10.1016/j.fsigen.2019.02.022.

[18] H. Swaminathan, M. O. Qureshi, C. M. Grgicak, K. Duffy, and D. S. Lun, "Four model variants within a continuous forensic DNA mixture interpretation framework: Effects on evidential inference and reporting," (in English), *Plos One,* Article vol. 13, no. 11, p. 23, Nov 2018, Art no. e0207599, doi: 10.1371/journal.pone.0207599.

[19] J.-A. Bright *et al.*, "Internal validation of STRmix™ – A multi laboratory response to PCAST," *Forensic Science International: Genetics,* vol. 34, pp. 11-24, 2018-05-01 2018, doi: 10.1016/j.fsigen.2018.01.003.

[20] D. Mcnevin, K. Wright, J. Chaseling, and M. Barash, "Commentary on: Bright et al. (2018) Internal validation of STRmix™ – a multi laboratory response to

PCAST, Forensic Science International: Genetics, 34: 11–24," *Forensic Science International: Genetics,* vol. 41, pp. e14-e17, 2019-07-01 2019, doi: 10.1016/j.fsigen.2019.03.016.

[21] B. Budowle, T. R. Moretti, A. L. Baumstark, D. A. Defenbaugh, and K. M. Keys, "Population data on the thirteen CODIS core short tandem repeat loci in African Americans, US Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians," (in English), *Journal of Forensic Sciences,* Article vol. 44, no. 6, pp. 1277-1286, Nov 1999. [Online]. Available: <Go to ISI>://WOS:000085830600024.

[22] L. E. Alfonse, A. D. Garrett, D. S. Lun, K. R. Duffy, and C. M. Grgicak, "A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt," *Forensic Science International: Genetics,* vol. 32, pp. 62-70, 2018-01-01 2018, doi: 10.1016/j.fsigen.2017.10.006.

[23] Y. Li *et al.,* "Allele frequency of 19 autosomal STR loci in the Bai population from the southwestern region of mainland China," *ELECTROPHORESIS,* vol. 36, no. 19, pp. 2498-2503, 2015-10-01 2015, doi: 10.1002/elps.201500129.

[24] M. Ozeki and K. Tamaki, "Allele frequencies of 37 short tandem repeat loci in a Japanese population," *Legal Medicine,* vol. 15, no. 6, pp. 342-346, 2013-11-01 2013, doi: 10.1016/j.legalmed.2013.08.006.

[25] S. Y. Yoo *et al.,* "A large population genetic study of 15 autosomal short tandem repeat loci for establishment of Korean DNA Profile Database," *Molecules and Cells,* vol. 32, no. 1, pp. 15-19, 2011-07-01 2011, doi: 10.1007/s10059-011-2288-4.

[26] C. R. Hill, D. L. Duewer, M. C. Kline, M. D. Coble, and J. M. Butler, "U.S. population data for 29 autosomal STR loci," *Forensic Science International: Genetics,* vol. 7, no. 3, pp. e82-e83, 2013-05-01 2013, doi: 10.1016/j.fsigen.2012.12.004.

[27] M. A. Marciano and J. D. Adelman, "PACE: Probabilistic Assessment for Contributor Estimation-A machine learning-based assessment of the number of contributors in DNA mixtures," (in English), *Forensic Science International-Genetics,* Article vol. 27, pp. 82-91, Mar 2017, doi: 10.1016/j.fsigen.2016.11.006.

[28] R. G. Cowell, "Computation of marginal distributions of peak-heights in electropherograms for analysing single source and mixture STR DNA samples," *Forensic Science International: Genetics,* vol. 35, pp. 164-168, 2018-07-01 2018, doi: 10.1016/j.fsigen.2018.04.007.

[29] K. C. Peters, H. Swaminathan, J. Sheehan, K. R. Duffy, D. S. Lun, and C. M. Grgicak, "Production of high-fidelity electropherograms results in improved and consistent DNA interpretation: Standardizing the forensic validation process," (in English), *Forensic Science International-Genetics,* Article vol. 31, pp. 160-170, Nov 2017, doi: 10.1016/j.fsigen.2017.09.005.

[30] O. Hansson, P. Gill, and T. Egeland, "STR-validator: An open source platform for validation and process control," *Forensic Science International: Genetics,* vol. 13, pp. 154-166, 2014-11-01 2014, doi: 10.1016/j.fsigen.2014.07.009.

[31] H. Kelly *et al.*, "A sensitivity analysis to determine the robustness of STRmix (TM) with respect to laboratory calibration," (in English), *Forensic Science International-Genetics,* Article vol. 35, pp. 113-122, Jul 2018, doi: 10.1016/j.fsigen.2018.04.009.

[32] J. M. Butler, M. C. Kline, and M. D. Coble, "NIST interlaboratory studies involving DNA mixtures (MIX05 and MIX13): Variation observed and lessons learned," *Forensic Science International: Genetics,* vol. 37, pp. 81-94, 2018-11-01 2018, doi: 10.1016/j.fsigen.2018.07.024.

[33] M. D. Coble *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," (in English), *Forensic Science International-Genetics,* Article vol. 25, pp. 191-197, Nov 2016, doi: 10.1016/j.fsigen.2016.09.002.

[34] J.-A. Bright *et al.*, "STRmix™ collaborative exercise on DNA mixture interpretation," *Forensic Science International: Genetics,* vol. 40, pp. 1-8, 2019-05-01 2019, doi: 10.1016/j.fsigen.2019.01.006.

[35] C. M. Grgicak, S. Karkar, X. Yearwood-Garcia, L. E. Alfonse, K. R. Duffy, and D. S. Lun, "A large-scale validation of NOCIt?s a posteriori probability of the number of contributors and its integration into forensic interpretation pipelines," (in English), *Forensic Science International-Genetics,* Article vol. 47, p. 14, Jul 2020, Art no. 102296, doi: 10.1016/j.fsigen.2020.102296.

[36] J. Hannig, S. Riman, H. Iyer, and P. M. Vallone, "Are reported likelihood ratios well calibrated?," *Forensic Science International: Genetics Supplement Series,* vol. 7, no. 1, pp. 572-574, 2019-12-01 2019, doi: 10.1016/j.fsigss.2019.10.094.

[37] M. Kruijver *et al.*, "Estimating the number of contributors to a DNA profile using decision trees," (in English), *Forensic Science International-Genetics,* Article vol. 50, p. 11, Jan 2021, Art no. 102407, doi: 10.1016/j.fsigen.2020.102407.

[38] C. Benschop, A. Backx, and T. Sijen, "Automated estimation of the number of contributors in autosomal STR profiles," (in English), *Forensic Science International Genetics Supplement Series,* Article vol. 7, no. 1, pp. 7-8, Dec 2019, doi: 10.1016/j.fsigss.2019.09.003.

[39] C. G. Son, "Database of mRNA gene expression profiles of multiple human organs," *Genome Research,* vol. 15, no. 3, pp. 443-450, 2005-02-14 2005, doi: 10.1101/gr.3124505.

[40] T. J. Lambert, "FPbase: a community-editable fluorescent protein database," *Nature Methods,* vol. 16, no. 4, pp. 277-278, 2019-04-01 2019, doi: 10.1038/s41592-019-0352-8.

[41] R. A. Vanbogelen, P. Sankar, R. L. Clark, J. A. Bogan, and F. C. Neidhardt, "The gene-protein database ofEscherichia coli: Edition 5," *Electrophoresis,* vol. 13, no. 1, pp. 1014-1054, 1992-01-01 1992, doi: 10.1002/elps.11501301203.

[42] J. R. Gilder, T. E. Doom, K. Inman, and D. E. Krane, "Run-specific limits of detection and quantitation for STR-based DNA testing," (in English), *Journal of Forensic Sciences,* Article vol. 52, no. 1, pp. 97-101, Jan 2007, doi: 10.1111/j.1556-4029.2006.00318.x.

[43] H. Kaiser, "QUANTITATION IN ELEMENTAL ANALYSIS .2," (in English), *Analytical Chemistry,* Editorial Material vol. 42, no. 4, pp. A26-&, 1970, doi: 10.1021/ac60286a027.

[44] U. J. Monich, K. Duffy, M. Medard, V. Cadambe, L. E. Alfonse, and C. Grgicak, "Probabilistic characterisation of baseline noise in STR profiles," (in English), *Forensic Science International-Genetics,* Article vol. 19, pp. 107-122, Nov 2015, doi: 10.1016/j.fsigen.2015.07.001.

[45] P. Gill *et al.*, "DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures," *Forensic Science International,* vol. 160, no. 2-3, pp. 90-101, 2006-07-01 2006, doi: 10.1016/j.forsciint.2006.04.009.

[46] J. M. Butler, *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers*. Elsevier Science, 2005.

[47] H. Kaiser, "QUANTITATION IN ELEMENTAL ANALYSIS," (in English), *Analytical Chemistry,* Editorial Material vol. 42, no. 2, pp. A24-+, 1970, doi: 10.1021/ac60284a022.

[48] L. A. Currie, "Limits for qualitative detection and quantitative determination. Application to radiochemistry," *Analytical Chemistry,* vol. 40, no. 3, pp. 586-593, 1968-03-01 1968, doi: 10.1021/ac60259a007.

[49] L. A. Currie, "Detection and quantification limits: origins and historical overview," (in English), *Analytica Chimica Acta,* Article vol. 391, no. 2, pp. 127-134, May 1999, doi: 10.1016/s0003-2670(99)00105-1.

[50] U. J. Moenich *et al.*, "A Signal Model for Forensic DNA Mixtures," in *48h Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, Nov 02-05 2014, LOS ALAMITOS: Ieee Computer Soc, in Conference Record of the Asilomar Conference on Signals Systems and Computers, 2014, pp. 429-433. [Online]. Available: <Go to ISI>://WOS:000370964900077. [Online]. Available: <Go to ISI>://WOS:000370964900077

[51] J. M. Butler, R. Schoske, P. M. Vallone, J. W. Redman, and M. C. Kline, "Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations," *J Forensic Sci,* Research Support, Non-U S Gov't vol. 48, no. 4, pp. 908-11, 2003.

[52] J. M. Butler, *Advanced Topics in Forensic DNA Typing: Interpretation : Interpretation*.

[53] E. L. R. Butts, M. C. Kline, D. L. Duewer, C. R. B. Hill, J. M. Butler, and P. M. Vallone, "NIST validation studies on the 3500 Genetic Analyzer," *Forensic Science International: Genetics Supplement Series,* vol. 3, no. 1, pp. e184-e185, 2011-12-01 2011, doi: 10.1016/j.fsigss.2011.08.092.

[54] H. Swaminathan, C. M. Grgicak, M. Medard, and D. S. Lun, "NOCIt: A computational method to infer the number of contributors to DNA samples analyzed by STR genotyping," (in English), *Forensic Science International-Genetics,* Article vol. 16, pp. 172-180, Mar 2015, doi: 10.1016/j.fsigen.2014.11.010.