Forensic Science International: Genetics 32 (2018) 62–70

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Short communication

A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt

Lauren E. Alfonse[a], Amanda D. Garrett[a], Desmond S. Lun[b,c,d], Ken R. Duffy[e], Catherine M. Grgicak[a,b,f,*]

[a] Biomedical Forensic Sciences Program, Boston University School of Medicine, United States
[b] Center for Computational and Integrative Biology, Rutgers University, United States
[c] Department of Computer Science, Rutgers University, Camden, United States
[d] Department of Plant Biology and Pathology, Rutgers University, New Brunswick, United States
[e] Hamilton Institute, Maynooth University, Ireland
[f] Department of Chemistry, Rutgers University, Camden, United States

ARTICLE INFO

Keywords:
Forensic DNA
PROVEDIt
STRs
Human identification
STR database

ABSTRACT

DNA-based human identity testing is conducted by comparison of PCR-amplified polymorphic Short Tandem Repeat (STR) motifs from a known source with the STR profiles obtained from uncertain sources. Samples such as those found at crime scenes often result in signal that is a composite of incomplete STR profiles from an unknown number of unknown contributors, making interpretation an arduous task. To facilitate advancement in STR interpretation challenges we provide over 25,000 multiplex STR profiles produced from one to five known individuals at target levels ranging from one to 160 copies of DNA. The data, generated under 144 laboratory conditions, are classified by total copy number and contributor proportions. For the 70% of samples that were synthetically compromised, we report the level of DNA damage using quantitative and end-point PCR. In addition, we characterize the complexity of the signal by exploring the number of detected alleles in each profile.

1. Introduction

Amplification of multiple short tandem repeat (STR) fragments, also known as microsatellites, followed by capillary electrophoresis and laser-induced fluorescence detection is the chief technique by which forensic and clinical laboratories identify the presence of a contributor to unknown biological specimens [1,2]. Forensic laboratories conduct human identity testing by examining hypervariable STR regions obtained from evidence to assess the likelihood that a person-of-interest (POI) was a contributor to the biological material, while clinical laboratories examine STR profiles to evaluate chimerism after hematopoietic stem cell transplantation [3] or to quality check cell lines within the laboratory [4]. Other applications of multiplex-STR amplification include parentage and kinship testing [5,6], and examinations of human population diversity [7].

The analysis of forensic samples is particularly challenging as little is known about the condition of, or complexity associated with, the biological material present on an evidentiary item. Any number of contributors could have deposited any number of cells on the substrate from which the sample was collected. Furthermore, the forensic sample may be obtained from environments that promote DNA degradation [8] or PCR inhibition [9,10], leading to electropherograms (EPGs) that exhibit a downward trend in peak heights as the molecular weight of the amplicon increases [11]. The result is an STR profile that may consist of incomplete signal from any number of unknown contributors. This signal is further obfuscated by the presence of noise [12,13] and stutter, a PCR-based artefact. Stutter occurs as a result of strand slippage during PCR which produces an echo peak at a fixed known distance from the allelic peak [14]. Stutter artefacts are typically one STR unit larger or smaller than the biological allele and comprise a significant portion of the total signal [15–17], complicating interpretation of mixed source traces. For complex samples, this combination of issues in conjunction with a range of post-processing EPG signal filtering and a variety of inference techniques can contribute to contradictory inference [18,19].

As suggested in [20] and [21], a large dataset would play a critical role in demonstrating the foundational validity and robustness of new or existing DNA identity testing technology. It would facilitate new techniques in satisfying admissibility standards that require scientific evidence to have earned widespread acceptance across all sample types

[*] Corresponding author at: Department of Chemistry, Science Building Rm 306C, Rutgers University, 315 Penn St, Camden, NJ, 08102, United States.
E-mail address: c.grgicak@rutgers.edu (C.M. Grgicak).

http://dx.doi.org/10.1016/j.fsigen.2017.10.006
Received 14 July 2017; Received in revised form 7 September 2017; Accepted 20 October 2017
Available online 24 October 2017

before they are presented to the trier of fact. Additionally, all burgeoning technologies [22–24] need to be benchmarked against current ones, making the availability of a readily accessible, universal testing set crucial.

Producing a large set of DNA profiles garnered from cellular admixtures that are meticulously processed such that artificial dilution effects [17,25] are minimized and all input masses are represented is an involved endeavor requiring the extraction, quantification, amplification, and electrophoresis of thousands of samples. Meeting that need, here we describe the PROVEDIt (Project Research Openness for Validation with Empirical Data) dataset: a database containing over 25,000 STR profiles garnered from numerous sample types that contain DNA of varying quality and quantity. The collection of profiles includes one to five person mixtures of varying contributor ratios amplified with DNA target masses ranging from 0.007 to 1 ng. The profiles were generated using 144 laboratory conditions from samples containing untreated; UV-damaged; enzymatically/sonically degraded; and inhibited DNA. To the best of our knowledge, this constitutes the largest, most comprehensive, condition-dependent STR mixture database for purposes of human identity testing.

The statistical comparison of evidentiary or unknown profiles to profiles obtained from known persons is typically accomplished within a likelihood ratio (LR) framework [26–28]. This approach compares the probability of the data given two hypotheses: that a POI contributed to the profile; and that the POI did not contribute. The LR framework has recently replaced traditional means of determining evidential strength, becoming the prevailing mechanism by which the strength of the evidence is communicated [16]. Where traditional methods involve binary, manual interpretation, the adoption of the LR framework has resulted in movement towards utilizing probabilistic models to evaluate unknown and evidentiary profiles [29–35]. The complex nature of low-template, compromised, multi-contributor DNA signal elicits the need for continued research, examination, and development in this arena [36–38].

For example, probabilistic-based interpretation systems typically require an assumption on the number of contributors (NOC) [39,40]. As a result, work on NOC estimation and the NOC assumption [18,41–43] has catalyzed the development of methods that manage this limitation [44–47]. Moreover, proposed interpretation schemes do not all use the same underlying probabilistic model. Model choice governs the probabilistic computation, and different models can result in different LRs, particularly for complex samples [48,49]. Thus, there is interest in exploring the impact of model variations on the inference [20]. Despite advances, the interpretation of whole signal remains challenging. Recent work in the development of probabilistic peak detection algorithms based on a Bayesian framework [50,51], the application of artificial neural networks [52], and the development of methods that do not rely upon signal thresholds [33,53,54] demonstrates that improvement in the field of human identification is forthcoming.

The primary challenge here is the development of methods that utilize all of the information contained within the data. Given that solutions will likely require interdisciplinary approaches and perspectives, a comprehensive dataset, such as the one introduced here, is necessary to foster research and development in this realm. In what follows, we first describe the dataset: the number and range of complexity of the samples by providing an overview of sample types; the number of contributors contained within the samples; and measures of DNA degradation and stutter artefacts. We then illustrate its value by considering a forensically relevant problem as an exemplar: we explore the impact of various conditions and sample types on the ability to detect the alleles from which the signal arose.

2. Materials and methods

2.1. Ethics

The procedures used to acquire the source materials for the database were in accordance with the ethical standards of the Institutional Review Board: Boston University School of Medicine Protocol Number H-31941.

2.2. Defining the dataset

The collection of profiles was generated over a four-year period using 144 different laboratory conditions. We extracted and purified DNA using typical organic or silica-based purification techniques. Three commercially available STR multiplexes (PowerPlex® 16 HS, Identifiler® Plus, and GlobalFiler®) and two generations of capillary electrophoresis instruments were used to amplify and separate the STR fragments. PowerPlex® 16 HS and Identifiler® Plus co-amplify 15 STR loci plus the sex-determining Amelogenin locus, while GlobalFiler® simultaneously amplifies 21 autosomal STRs, 1 Y-STR, 1 Y indel, and Amelogenin. The chromosomal locations and genetic diversity of the STRs amplified are well-established [2]. The capillary electrophoresis instruments used were the 3130 and 3500 Genetic Analyzers, both of which are commonly employed in operational settings [55].

The samples contained DNA originating from one to five persons, and the amplification target masses ranged from 0.007 to 1 ng. In the case of multi-contributor samples, contributor ratios ranged from equal parts to mixtures containing 99 parts of one and one part of the other(s) (Supplementary Tables 1 and 2). In addition, in a large subset of the samples, the DNA was compromised prior to amplification by: enzymatic degradation; UV irradiation; sonication-induced degradation; or PCR inhibition with humic acid. Table 1 summarizes the number of samples generated under each condition. Detailed laboratory methods are available in Supplementary Methods.

The two- to five-person mixture profiles were constructed using 37 different genotype combinations (Supplementary Tables 1 and 2). For

Table 1
The number of STR single-source and mixture profiles generated under each laboratory condition. Mixtures were prepared using two methods as denoted under *Sample Type*: (1) DNA Extract Mixtures (DEM) were created by mixing DNA extracts to reach the specified ratio of major to minor contributor(s), and (2) Whole Blood Mixtures (WBM) were prepared by combining aliquots of whole blood from multiple contributors in the appropriate proportions prior to extraction. The STR multiplexes and the number of PCR cycles are listed in the second column, while the capillary electrophoresis instrument type and injection times are described in the next two columns. The laboratory conditions used to induce DNA damage, degradation, or PCR inhibition are specified. If additional damage was not induced in the laboratory environment, then the samples were classified as 'Untreated.'.

Sample Type	Kit (PCR cycle no.)	3130 (5, 10, 20 s)	3500 (5, 15, 25 s)	Untreated	DamagedUV	Degraded Enzymatic	Degraded Sonication	Inhibited Humic Acid
WBM	IDPlus (28 cycles)	x		1560	1512	4728	1152	1512
DEM	IDPlus (29 cycles)		x	3212				
WBM	GlobalFiler (29 cycles)		x	1560	1512	4728	1152	1512
DEM	PP16HS (32 cycles)	x		1024				

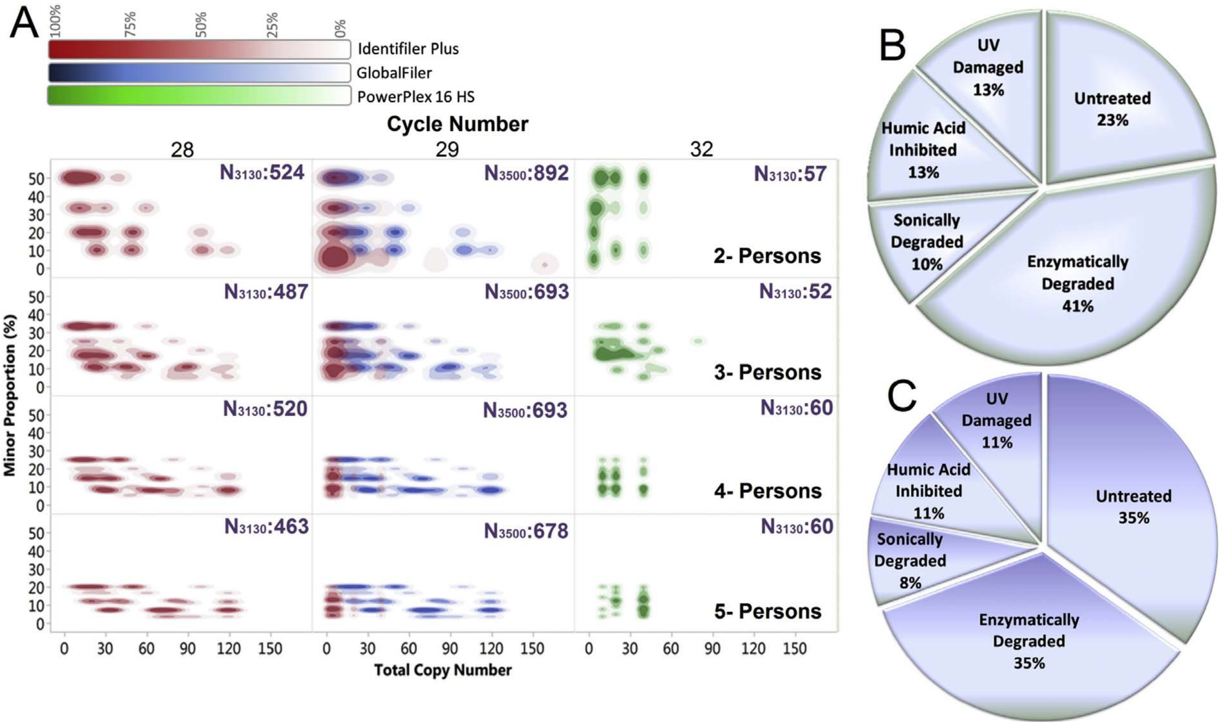

Fig. 1. Schematic depicting a summary of the sample types. (A) A density plot of the proportion of the minor contributor versus the total copy number, as per qPCR, for two- to five-person mixtures amplified with common human identity STR multiplexes, separated by the total number of PCR cycles utilized. The number of samples and the Genetic Analyzer type are also presented. (B) The percentage of samples processed on the 3130 Genetic Analyzer. (C) The percentage of samples processed on the 3500 Genetic Analyzer.

mixtures amplified with each of the three commercially available STR multiplexes, Fig. 1A provides a density plot of the proportion of the minor contributor versus the total copy number, separated by the number of PCR amplification cycles. Of the 5160 mixture samples, 76% contain a contribution from at least one individual of less than 20% of the total DNA content. Of 25,164 total samples, 17,808 were subjected to a condition that was expected to induce inefficiencies in amplification for some or all loci. Fig. 1B and C report the percentage of samples processed per condition on the 3130 and 3500 Genetic Analyzers, respectively. The database, therefore, is comprised of samples that range from simple to complex, and cumulatively represents the wide variety of data that might be encountered.

2.3. Description of the online resource

The dataset is available online at the *Laboratory for Forensic Technology Development and Integration* on lftdi.com. There are three versions of the data (Raw, Unfiltered, and Filtered), which are organized in folders under the *PROVEDIt* tab:

1) The raw data files (.hid and .fsa files) are catalogued in 24 groups based on the number of contributors in the sample, the instrument platform, the instrument injection time, the autosomal STRs targeted during amplification (i.e., kit type) as well as the number of PCR cycles.
2) Unfiltered CSV files are found in a folder labeled ‘UnFiltered’. These are exported CSV files from GeneMapper® *ID-X*. Details regarding the analysis settings are provided in the section *Peak Detection and Artefact Filtering,* below. These data contain the allele designation, basepair size and peak heights for all samples. Artefacts, such as pull-up, minus A and raised baseline are not filtered from these data.
3) Filtered CSV files are found in a folder labeled ‘Filtered’. These are the same CSV files described in 2), but with minus A and pull up artefacts removed. In the case of samples amplified with the GlobalFiler® Amplification kit, exotic stutters in the SE33 locus, which are located half a repeat unit away from the allele, are also filtered. Conditions for artefact removal are detailed in the section *Peak Detection and Artefact Filtering,* below.

In addition to the data, each folder contains a file with the known genotypes and a file explaining the sample naming convention, which is recapitulated in the Supplementary Methods. There exist two sets of samples, DNA Extract Mixtures (DEM) and Whole Blood Mixtures (WBM), and the naming conventions between them differ. The DEM sample set is designated with the project code RD12-0002, while the WBM sample names contain project code RD14-0003. The names function as an “answer key” and contain most of the pertinent information related to the sample, such as the true NOC, the sources of DNA, the ratios of each contributor, and the total mass of DNA amplified. Inquiries regarding submission of additional samples to the database may be addressed to the corresponding author.

2.4. Dataset quality control

Prior to inclusion in the database, each sample was evaluated and compared to the known genotypes of the biological source(s). If a sample exhibited indications of gross contamination, the sample was omitted from the dataset. Gross contamination was defined as the presence of at least two extraneous peaks with heights exceeding that of baseline noise that could not be accounted for as part of the set of alleles or stutter belonging to the known contributors to that sample. All other samples remain in the dataset, including those that contained potential allele drop-in, which results from the amplification of extraneous fragments of DNA not native to the extract. Peak height balance between heterozygous alleles within a locus for one-person samples was evaluated. In accordance with expectations [15,55], the peak height balance decreased as the average intensity of the peak decreased (Supplementary Fig. 1). However, the dataset also includes samples with anomalous or unusual results. Specifically, sample 41, which has a known genotype of (14, 15) at the D1S1656 locus, reproducibly exhibited a lower than expected signal intensity for the 14 allele (Supplementary Fig. 2).

2.5. Peak detection and artefact filtering

To create the Unfiltered and Filtered versions of the database, electropherograms were analyzed with GeneMapper® *ID-X* using Local Southern sizing at an analytical threshold (AT) of one Relative Fluorescent Unit (RFU). The genotype table for each sample was exported from GeneMapper® as a CSV file containing the allele, size, and height for all peaks.

Prior to the analysis and interpretation described here, filtered data were created. Artefacts were removed by employing CleanIt, an automated filtering procedure used to remove signal associated with: cross-talk or pull-up between EPG color channels and incomplete adenylation, also known as minus A. CleanIt is available on lftdi.com. Via that tool, artefacts can be automatically removed from electropherogram data in accordance with user-set values, and a new file containing filtered signal is generated. This file can then be utilized for downstream interpretation.

For purposes of the filtered data described here, three criteria were applied to filter pull-up: 1) the potential pull-up peak and the parent peak were labeled with different dyes; 2) the size of the potential pull-up peak fell within ± 0.6 base pairs (b.p.) of the size of the parent peak; and 3) the height of the pull-up peak divided by the height of the parent peak was < 6%. If all three criteria were fulfilled, the peak was classified as pull-up and removed from the data.

Similarly, CleanIt categorized peaks as complex pull-up if the following five criteria were met: 1) two sister alleles present at the same locus were one STR unit apart; 2) the two sister alleles were > 50% of each other in height; 3) the potential complex pull-up peak(s) appeared in a dye channel different from that of the sister alleles; 4) the size of the complex pull-up peak (in b.p.) fell between the sizes of the two sister alleles, ± 0.3 b.p.; and 5) the height of the complex pull-up peak (s) divided by the height of the shorter sister allele was < 6%. A peak was removed from the data if all five criteria were met. CleanIt categorized signal as minus A and removed it if a potential minus A peak was within −1 ± 0.6 b.p. of a plus A peak at the same locus and the height of the potential minus A peak divided by the height of the plus A peak was < 16%.

In the case of samples amplified with GlobalFiler®, stutter peaks at the SE33 locus which were 2 base pairs to the left of the known allele with a peak height less than 18% of the known allele were also filtered from the dataset. CleanIt does not filter exotic stutter, and thus this task was accomplished after CleanIt filtering was completed.

2.6. Exploring signal quality and complexity

2.6.1. Forward and reverse stutter

Stutter is a PCR artefact that is commonly observed in STR amplification. Forward and reverse stutter intensities were evaluated as a function of allele size and peak height, with the exception of peaks that could have been the result of a combination of forward and reverse stutter, which arises for genotypes whose alleles differ by two repeat units. Specifically, we computed the stutter peak height ratio, defined as the intensity of the peak in stutter position divided by the allele peak height, and we plot the stutter ratio against the number of repeat units.

2.6.2. Evaluating DNA damage and PCR inhibition with qPCR and STR contours

The majority of samples were quantified using the Quantifiler® Trio DNA Quantification Kit (Life Technologies) on the 7500 Real-Time PCR System using the manufacturer's recommended thermalcycling protocol and an external calibrator [56,57]. If Quantifiler® Trio was not available, the Quantifiler® Duo assay was utilized. The Quantifiler® Trio assay co-amplifies four targets: a small autosomal human target (80 b.p.), a large autosomal human target (214 b.p.), a human male target, and an internal PCR control (IPC). The concentration of each target was calculated by determining the cycle number at which the emitted fluorescence of the given target reached a defined threshold and comparing that cycle number to an external calibration curve. The amplification of the smaller fragment relative to the larger fragment provided information regarding the degree of DNA damage [58,59] and is a potential predictor of quality of the STR profile. The quality index (QI), which is the ratio of the concentrations of the small and large autosomal fragments, was computed for all samples run with the Quantifiler® Trio assay.

The degree of degradation or inhibition was also assessed using the contour of the STR signal, which was well modeled as exponential decay in fluorescence as a function of molecular weight:

$$H_l = Ae^{B\overline{sl}} \tag{1}$$

where H_l is the sum of the peak heights associated with the known genotypes at locus l, $\bar{s}$ is the average base pair size of the STR alleles at locus l, and A and B are the exponential parameters obtained for each sample using least squares regression. In extreme cases of decay, the highest molecular weight peaks may not reach detectable levels. If high molecular weight markers exhibit low peak heights due to degradation or inhibition of the PCR reaction, B will take a large negative value. In contrast, if there is good signal balance across all loci, indicating efficient PCR and high quality template DNA, B will be near zero.

2.6.3. Signal detection

The non-detection (ND) rate was computed as:

$$ND = 1 - \left(\frac{N_{PH \geq 1}}{N_{expected}}\right) \tag{2}$$

where $N_{PH \geq 1}$ is the number of peaks in known allele positions with height greater than or equal to one RFU, and $N_{expected}$ is the expected number of alleles in the profile as determined by the known genotypes. Though related, ND is not equivalent to the rate of allelic dropout since fluorescent signal from noise is observed at low fluorescent intensities [60]. The number of peaks within a locus that exceeded one RFU were counted and compared against the expected number given the known genotypes for each sample. In addition, the largest number of known allele positions containing RFU signal greater than zero within a locus was compared against the true number of contributors.

3. Results

3.1. Signal quality and complexity

By plotting the stutter ratio against the number of repeat units we found that, as previously described, stutter ratios increase with allele size or, more specifically, with the longest uninterrupted repeat number [15,61]. Further, as with the allele peak height balance, the stutter ratio is also affected by the template mass of DNA available at PCR initiation. Specifically, the stutter ratio for both reverse and forward stutter peaks increased as the intensity of the allele signal decreased for all loci (Supplementary Figs. 3–5), which may be attributed to slippage early in the PCR process compounded by noise effects [17]. The average ratios ranged from 3 to 12% and 1 to 4% for the reverse and forward stutter categories, respectively.

Compromised DNA samples often exhibit decreasing signal with increasing molecular weight wherein the rate of DNA decay depends on factors that include whether the sample is exposed to microorganisms, sunlight, unfavorable temperatures, or geochemical inhibitors found in the environment. In some operational settings, quantification of the DNA extract is a necessity [62] and may be used to guide downstream laboratory processing decisions. Quantification is accomplished using qPCR and the simultaneous detection of the per-cycle increase in the

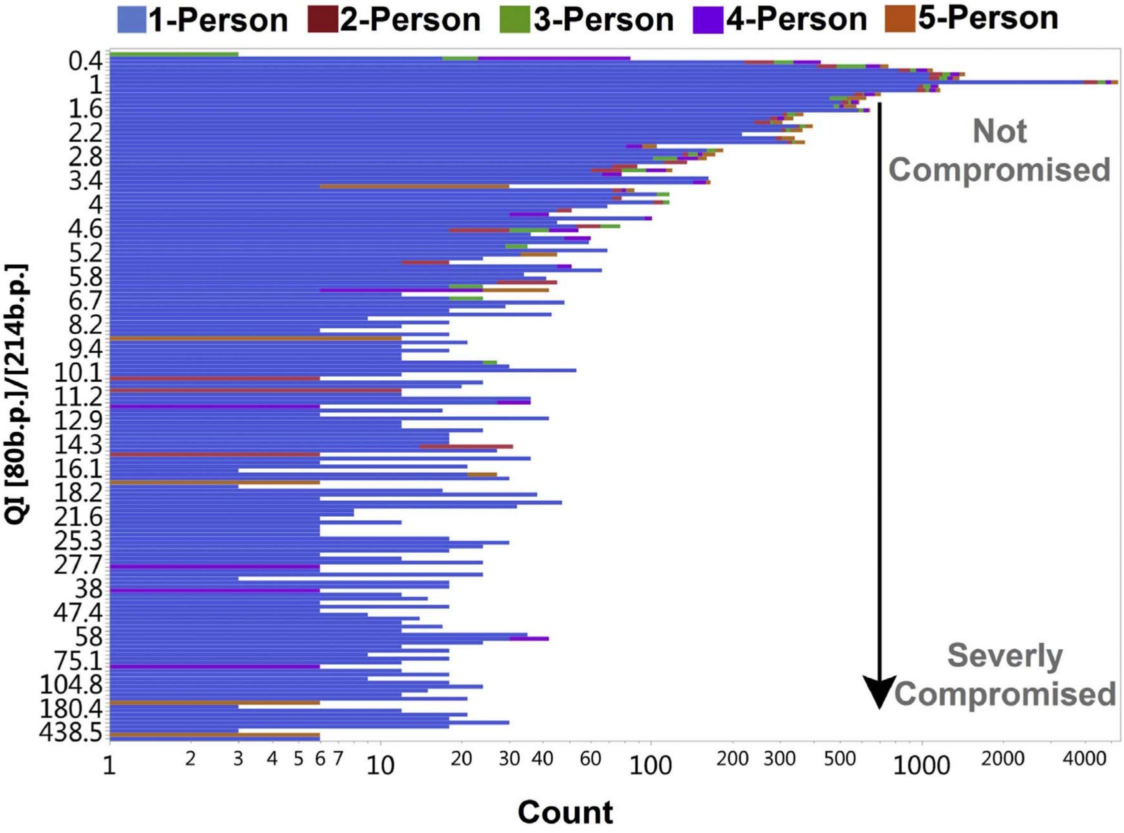

Fig. 2. The quality index (QI) obtained for WBM samples. The QI is a ratio of the concentrations of 80 and 214 b.p. fragments and may provide an indication of degradation, damage, or inhibition early in the DNA processing pipeline. In general, the QI increases as the DNA extract becomes more compromised. QIs ranged from less than 0.3–439. The majority of data represent samples originating from one person, but the full range of QIs from 'not compromised' to 'severely compromised' is also well represented in the complex mixtures of two to five contributors. In highly compromised samples, the 214 b.p. large autosomal fragment often does not reach detectable levels, rendering the QI value as indeterminate (Ind). There were 1207 samples for which indeterminate QI values were obtained, and those data are not included in this plot.

concentration of two distinct PCR amplicons of different length. Fig. 2 displays a histogram of the QI of the samples as determined by qPCR, demonstrating that the database consists of single-source and complex mixture profiles of varying states of decay. Specifically, 26% of the samples resulted in QI values suggesting moderate (QI = 1.5–4) and 12% severe (QI > 4) levels of degradation or inhibition. Supplementary Fig. 6 shows example qPCR data obtained for samples that demonstrate minor, moderate, and severe levels of amplification inefficiencies as measured by qPCR.

As illustrations of the different conditions of DNA and *B* values that may be encountered, Fig. 3A and B depict EPGs obtained from single-source samples that were untreated and severely damaged through sonication, respectively. Fig. 3C presents the QI value versus *B* calculated for each sample, separated by the amplification chemistry and the treatment protocol utilized. As expected, all untreated samples resulted in QI values near one and *B* parameters near zero. Samples that were subjected to conditions that degraded the DNA show that the QI and *B* parameter are correlated, suggesting that the QI metric can be used to predict the STR sloping pattern for these sample types. Interestingly, samples subjected to conditions intended to induce PCR inhibition show that while the qPCR QI metric demonstrates that PCR inhibitors affect the amplification of the large autosomal fragment, the STR profile presents only minor signs of inhibition. Spearman's ρ for QI and *B* was computed (Supplementary Table 3), with correlation being strong for enzymatically degraded, sonicated, and UV-damaged samples (ρ between −0.8656 and −0.6324), but less so for untreated and inhibited samples (ρ between −0.1382 and −0.0143).

3.2. *Signal detection rates*

Depending on the laboratory process, the condition of the sample, or the available template molecules, the rate of allele non-detection can be large and can substantively impact profile quality. In extreme cases, EPGs can contain partial STR profiles from many contributors where the signal from one or all of the contributors has been compromised in an unknown manner, resulting in inference challenges.

Given the known genotypes of each contributor within the sample, Fig. 4A plots the number of expected alleles versus the number of peaks observed in those allele positions. Notably, there is large overlap between the number of alleles observed in the three-, four- and five-person mixtures, suggesting that factors associated with allele loss and allele stacking can complicate downstream interpretation [63,64]. For example, Fig. 4B represents EPG signal from three representative loci obtained from a five-person mixture containing equal parts from each contributor, wherein no more than seven alleles were detected at any one locus, and the peak height ratios do not provide definitive evidence that five, rather than four, persons comprise this mixture. Fig. 4C depicts the samples that exhibit a maximum of two, four, six, or eight detected peaks in allele positions at any single STR locus as it relates to the rate of allele non-detection. As the rate of allele non-detection surpasses 0.1, many three- and four-person mixtures do not exhibit more than four detected alleles at any one locus. Similarly, at high rates of allele drop-out, some two-person mixtures do not contain loci with more than two peaks detected in an allele positon. Profiles that exhibited greater than four to five detects at a locus originated from three-, four-, and five-person samples. Similarly, profiles that exhibited seven to eight detects at a locus originated from four- or five-person mixtures.

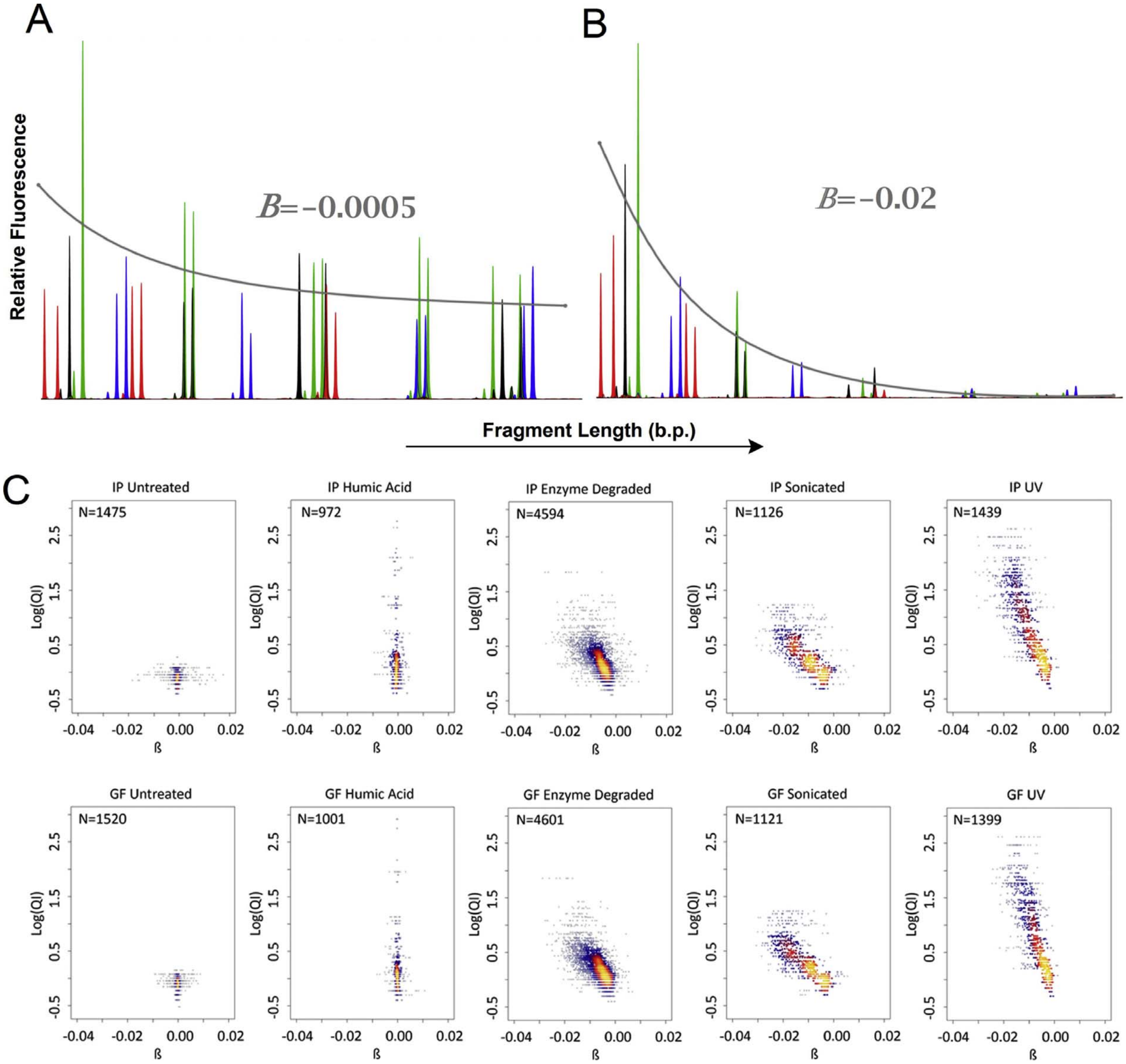

Fig. 3. The degree of sloping observed in the STR profile. The exponent in the decay in fluorescence as a function of molecular weight, *B*, ranged from 0.02 to −0.04. *B* values significantly below zero correspond to compromised samples and generally indicate some reduction in RFU signal as the length of the amplicons increase. (A) An electropherogram obtained from an untreated sample amplified at 0.25 ng (~40 copies). There is good intra-locus peak height balance across all heterozygous loci, and the total RFU signal is approximately equivalent across all loci labeled with the same dye, which is represented in the *B* value (−0.0005). (B) An electropherogram obtained from a sample treated with 30 sonication cycles amplified at 0.25 ng. The decrease in peak height as the fragment length increases is apparent and characteristic of the "sloping effect" observed in degraded profiles; this is represented by the highly negative *B* term obtained (−0.02). (C) The correlation between QI and *B*. The treatment protocol, PCR amplification kit, and number of profiles are noted for each plot. The plots are color-coded by density, where yellow and light purple represent areas of highest and lowest sample density, respectively. Correlation, measured using Spearman's Rho, is summarized in Supplementary Table 3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The mosaic plot of Fig. 4D summarizes the data for the two- to five-person mixture profiles and shows the proportion of samples from the known number of contributors versus the number of peaks detected in allele positions. Interestingly, 53%, 45%, and 52% of samples that exhibit five to six detected peaks at allele positions at a locus originated from three-person mixtures for the GlobalFiler®, Identifiler® Plus, and PowerPlex® 16 HS multiplexes, respectively. The remainder originated from either four- or five-person mixtures. Similarly, 44%, 37%, and 39% of samples that exhibit a positive detection count of seven or eight were from four-person mixtures, with the remainder originating from five-person samples. Despite the variety of five-person genotype combinations represented, no five-person mixture displayed more than eight detects at allele positions at any one locus (Supplementary Tables 1 and 2) for the samples in this database. This illustrates that the PROVEDIt database contains samples of sufficient breadth and complexity and contains numerous samples with high and low levels of false non-detections.

We note that the rate of non-detections reported in this work is dependent on a number of factors including the analytical threshold (AT) applied to the data and the peak detection parameters employed. If the AT is increased from 1 RFU, the rate of non-detection will also increase [17,65]; if the peak detection parameters such as smoothing are relaxed, peak detection may increase. In addition to the filtered data, the PROVEDIt database includes the raw files allowing for large-scale, inter-institutional studies that evaluate the effects of these conditions on detection and downstream inference.

4. Discussion

We have made available over 25,000 multiplex STR profiles, garnered from 144 laboratory conditions that range from one- to five-contributors at targets ranging from one to 160 copies of DNA. The

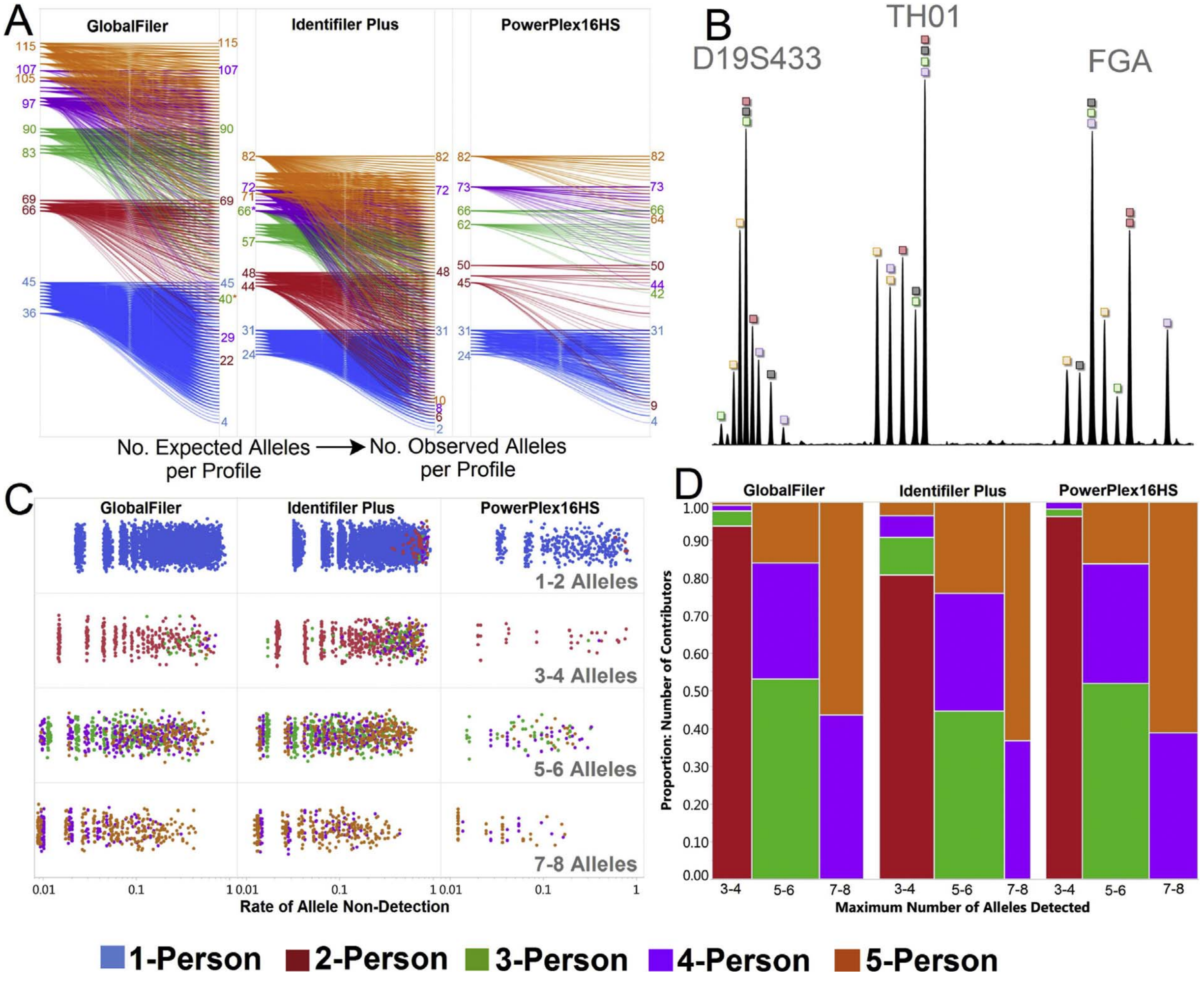

Fig. 4. Allele stacking and allele dropout in PROVEDIt profiles. In plots (A), (C), and (D) the data are color-coded by the known number of contributors and separated according to the PCR amplification kit utilized. (A) The expected number of alleles per profile, given the genotypes of the contributors in the mixture (left axis), connected to the number of times signal ≥1 RFU was detected in allele positions (right axis). (B) An EPG of three representative loci obtained for a five-person mixture containing equal parts from each person: (🟨) Contributor 1; (🟪) Contributor 2; (🟧) Contributor 3; (■) Contributor 4; and (🟥) Contributor 5. (C) The number of detected alleles at a locus for each sample against the rate of non-detection. The magnitude of the y-axis is not significant as the points were jittered for visualization purposes. (D) The proportion of samples originating from the known number of contributors versus the number of peaks ≥1 RFU at a locus for all samples. In no instance are greater than eight detections at allele positions observed at a locus, despite the presence of five-person genotype combinations in the database.

profiles are comprehensive in their genetic diversity and quality. We demonstrate that the database meets the needs of the human identification community in that it consists of sample types regularly encountered in operations. We illustrate the value of the database by showing that allele dropout and degradation effects are well represented. We demonstrate that PCR artefacts, such as stutter, are regularly encountered and are impacted by stochastic PCR or noise effects in the low-template regime.

Due to the persistence of STR profiling in clinical and forensic laboratories, these data are pertinent to public health, as well as criminal and international justice efforts. Further, from a global perspective, STR typing continues to be a crucial means by which decedents in mass graves, terrorist attacks, and mass disasters are identified [66]. As automated and/or probabilistic STR interpretation systems have begun to replace manual interpretation, thorough validation of new methodologies and comparison against existing ones is critical and necessitates the availability of a large-scale and open sample-set. We publish this database to facilitate advances in interpretation of these complex traces and foster continued multi-disciplinary development of approaches to evaluating STR signal.

5. Conclusion

The PROVEDIt database provides the forensic, clinical, and broader scientific community with a large-scale database that can be utilized for purposes of developing new or comparing existing interpretation or analysis strategies. The database is the largest, most comprehensive dataset of its kind and is openly available. In addition, the PROVEDIt data can be utilized as a benchmark against which new developments can be judged or for pedagogical pursuits. Most importantly, this resource fills the gap highlighted by a recent high-profile report [20] calling for large-scale studies that verify the use of computational procedures for purposes of human identity testing using STR signal obtained from mixed, possibly partial, sources.

Conflict of interest

None declared.

Acknowledgements

We thank Robin W. Cotton, John M. Butler, Mike Coble, Charlotte J. Word, Muriel Medard, Slim Karkar, and Harish Swaminathan for valuable discussions. We also thank Lauren Taranow and Kelsey C. Peters for help with sample preparation. This work was supported NIJ2011-DN-BX-K558, NIJ2012-DN-BX-K050, and ARO RIF W911NF-14-C-0096 awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice and the Department of Defense, Army Research Office, Rapid Innovation Fund, respectively. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not reflect those of the Department of Justice or Department of Defense.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2017.10.006.

References

[1] P. Gill, H. Haned, O. Bleka, O. Hansson, G. Dorum, T. Egeland, Genotyping and interpretation of STR-DNA: Low-template, mixtures and database matches-Twenty years of research and development, Forensic Sci. Int. Genet. 18 (2015) 100–117.

[2] J.M. Butler, Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, 2nd ed., Elsevier Academic Press, Burlington, 2005.

[3] S.A. Mousavi, M. Javadimoghadam, A. Ghavamzadeh, K. Alimoghaddam, A. Sayarifard, S.H. Ghaffari, et al., The relationship between STR-PCR chimerism analysis and chronic GvHD following hematopoietic stem cell transplantation, Int. J. Hematol.-Oncol. Stem Cell Res. 11 (2017) 24–29.

[4] C. Corral-Vazquez, R. Aguilar-Quesada, P. Catalina, G. Lucena-Aguilar, G. Ligero, B. Miranda, et al., Cell lines authentication and mycoplasma detection as minimun quality control of cell lines in biobanking, Cell Tissue Bank. (2017).

[5] G. Dørum, N. Kaur, M. Gysi, Pedigree-based relationship inference from complex DNA mixtures, Int. J. Legal Med. 131 (2017) 629–641.

[6] T. Egeland, K. Slooten, The likelihood ratio as a random variable for linked markers in kinship analysis, Int. J. Legal Med. 130 (2016) 1445–1456.

[7] N.M. Silva, L. Pereira, E.S. Poloni, M. Currat, Human neutral genetic variation and forensic STR data, PLoS One 7 (2012) e49666.

[8] R. Alaeddini, S.J. Walsh, A. Abbas, Forensic implications of genetic analyses from degraded DNA–a review, Forensic Sci. Int. Genet. 4 (2010) 148–157.

[9] R. Alaeddini, Forensic implications of PCR inhibition–A review, Forensic Sci. Int. Genet. 6 (2012) 297–305.

[10] M.E. Funes-Huacca, K. Opel, R. Thompson, B.R. McCord, A comparison of the effects of PCR inhibition in quantitative PCR and forensic STR analysis, Electrophoresis 32 (2011) 1084–1089.

[11] S. Vernarecci, E. Ottaviani, A. Agostino, E. Mei, L. Calandro, P. Montagna, Quantifiler® Trio Kit and forensic samples management: a matter of degradation, Forensic Sci. Int. Genet. 16 (2015) 77–85.

[12] U.J. Mönich, K. Duffy, M. Médard, V. Cadambe, L.E. Alfonse, C. Grgicak, Probabilistic characterisation of baseline noise in STR profiles, Forensic Sci. Int. Genet. 19 (2015) 107–122.

[13] J.R. Gilder, T.E. Doom, K. Inman, D.E. Krane, Run-specific limits of detection and quantitation for STR-based DNA testing, J. Forensic Sci. (2007) 2007.

[14] P.S. Walsh, N.J. Fildes, R. Reynolds, Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA, Nucleic Acids Res. 24 (1996) 2807–2812.

[15] J.-A. Bright, D. Taylor, J.M. Curran, J.S. Buckleton, Developing allelic and stutter peak height models for a continuous method of DNA interpretation, Forensic Sci. Int. Genet. 7 (2013) 296–304.

[16] F.R. Bieber, J.S. Buckleton, B. Budowle, J.M. Butler, M.D. Coble, Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion, BMC Genet. 17 (2016) 125.

[17] K.R. Duffy, N. Gurram, K.C. Peters, G. Wellner, 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, Electrophoresis 38 (2017) 855–868.

[18] C.C.G. Benschop, H. Haned, L. Jeurissen, P.D. Gill, T. Sijen, The effect of varying the number of contributors on likelihood ratios for complex DNA mixtures, Forensic Sci. Int. Genet. 19 (2015) 92–99.

[19] I.E. Dror, G. Hampikian, Subjectivity and bias in forensic DNA mixture interpretation, Sci. Justice 51 (2011).

[20] J.P. Holdren, E.S. Lander, W. Press, M. Savitz, W.M. Austin, C. Chyba, et al., Report To the President Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, (2016).

[21] M.D. Coble, J. Buckleton, J.M. Butler, T. Egeland, R. Fimmers, P. Gill, et al., DNA Commission of the International Society for Forensic Genetics: recommendations on the validation of software programs performing biostatistical calculations for forensic genetics applications, Forensic Sci. Int. Genet. 25 (2016) 191–197.

[22] X. Chen, J.C. Love, N.E. Navin, L. Pachter, M.J.T. Stubbington, V. Svensson, et al., Single-cell analysis at the threshold, Nat. Biotechnol. 34 (2016) 1111–1118.

[23] R.S. Just, L.I. Moreno, J.B. Smerick, J.A. Irwin, Performance and concordance of the ForenSeq™ system for autosomal and Y chromosome short tandem repeat sequencing of reference-type specimens, Forensic Sci. Int. Genet. 28 (2017) 1–9.

[24] G. Shin, S.M. Grimes, H. Lee, B.T. Lau, L.C. Xia, H.P. Ji, CRISPR–Cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis, Nat. Commun. 8 (2017) 14291.

[25] O. Hansson, T. Egeland, P. Gill, Characterization of degradation and heterozygote balance by simulation of the forensic DNA analysis process, Int. J. Legal Med. 131 (2017) 303–317.

[26] P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, Forensic Sci. Int. Genet. 7 (2013).

[27] C.D. Steele, M. Greenhalgh, D.J. Balding, Verifying likelihoods for low template DNA profiles using multiple replicates, Forensic Sci. Int. Genet. 13 (2014).

[28] J. Buckleton, J. Curran, A discussion of the merits of random man not excluded and likelihood ratios, Forensic Sci. Int. Genet. 2 (2008).

[29] H. Haned, P. Gill, K. Lohmueller, K. Inman, N. Rudin, Validation of probabilistic genotyping software for use in forensic DNA casework: definitions and illustrations, Sci. Justice 56 (2016) 104–108.

[30] J.A. Bright, D. Taylor, C. McGovern, S. Cooper, L. Russell, D. Abarno, et al., Developmental validation of STRmix (TM), expert software for the interpretation of forensic DNA profiles, Forensic Sci. Int. Genet. 23 (2016) 226–239.

[31] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, Validating TrueAllele® DNA mixture interpretation, J. Forensic Sci. 56 (2011).

[32] R. Puch-Solis, L. Rodgers, A. Mazumbder, S. Pope, I. Evett, J. Curran, et al., Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters, Forensic Sci. Int. Genet. 7 (2013) 555–563.

[33] H. Swaminathan, A. Garg, C.M. Grgicak, M. Medard, D.S. Lun, CEESIt A computational tool for the interpretation of STR mixtures, Forensic Sci. Int. Genet. 22 (2016) 149–160.

[34] R.G. Cowell, T. Graversen, S.L. Lauritzen, J. Mortera, Analysis of forensic DNA mixtures with artefacts, J. R. Stat. Soc. Ser. C Appl. Stat. 64 (2015).

[35] D.J. Balding, Evaluation of mixed-source, low-template DNA profiles in forensic science, Proc. Natl. Acad. Sci. U. S. A. 110 (2013).

[36] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Sci. Int. Genet. (2009) 2009.

[37] C.C.G. Benschop, C.P. van der Beek, H.C. Meiland, A.G.M. van Gorp, A.A. Westen, T. Sijen, Low template STR typing: effect of replicate number and consensus method on genotyping reliability and DNA database search results, Forensic Sci. Int. Genet. 5 (2011) 316–328.

[38] D. Taylor, J. Buckleton, Do low template DNA profiles have useful quantitative data? Forensic Sci. Int. Genet. 16 (2015) 13–16.

[39] M.W. Perlin, J.M. Hornyak, G. Sugimoto, K.W.P. Miller, TrueAllele® genotype identification on DNA mixtures containing up to five unknown contributors, J. Forensic Sci. 60 (2015) 857–868.

[40] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int. Genet. 7 (2013).

[41] I.W. Evett, S. Pope, Is it to the advantage of a defendant to infer a greater number of contributors to a questioned sample than is necessary to explain the observed DNA profile? Sci. Justice 54 (2014) 373–374.

[42] C.H. Brenner, Fairness in evaluating DNA mixtures, Forensic Sci. Int. Genet. 27 (2017) 186.

[43] S. Presciuttini, T. Egeland, About the number of contributors to a forensic sample, Forensic Sci. Int. Genet. 25 (2016) e18–e19.

[44] H. Swaminathan, C.M. Grgicak, M. Medard, D.S. Lun, NOCIt. A computational method to infer the number of contributors to DNA samples analyzed by STR genotyping, Forensic Sci. Int. Genet. 16 (2015) 172–180.

[45] M.A. Marciano, J.D. Adelman, PACE: Probabilistic Assessment for Contributor Estimation — a machine learning-based assessment of the number of contributors in DNA mixtures, Forensic Sci. Int. Genet. 27 (2017) 82–91.

[46] D. Taylor, J.-A. Bright, J. Buckleton, Interpreting forensic DNA profiling evidence without specifying the number of contributors, Forensic Sci. Int. Genet. 13 (2014) 269–280.

[47] K. Slooten, Accurate assessment of the weight of evidence for DNA mixtures by integrating the likelihood ratio, Forensic Sci. Int. Genet. 27 (2017) 1–16.

[48] T.W. Bille, S.M. Weitz, M.D. Coble, J. Buckleton, J.-A. Bright, Comparison of the performance of different models for the interpretation of low level mixed DNA profiles, Electrophoresis 35 (2014) 3125–3133.

[49] O. Bleka, C.C.G. Benschop, G. Storvik, P. Gill, A comparative study of qualitative and quantitative models used to interpret complex STR DNA profiles, Forensic Sci. Int. Genet. 25 (2016) 85–96.

[50] M. Woldegebriel, G. Vivó-Truyols, A new bayesian approach for estimating the presence of a suspected compound in routine screening analysis, Anal. Chem. 88 (2016) 9843–9849.

[51] M. Woldegebriel, G. Vivó-Truyols, Probabilistic model for untargeted peak detection in LC–MS using bayesian statistics, Anal. Chem. 87 (2015) 7345–7355.

[52] D. Taylor, D. Powers, Teaching artificial intelligence to read electropherograms, Forensic Sci. Int. Genet. 25 (2016) 10–18.

[53] M.W. Perlin, A. Sinelnikov, An information gap in DNA evidence interpretation, PLoS One (2009) 2009.

[54] M.W. Perlin, B. Szabady, Linear mixture analysis: a mathematical approach to resolving mixed DNA samples, J. Forensic Sci. 46 (2001) 1372–1378.

[55] J.A. Bright, S. Neville, J.M. Curran, J.S. Buckleton, Variability of mixed DNA profiles separated on a 3130 and 3500 capillary electrophoresis instrument, Aust. J. Forensic Sci. 46 (2014) 304–312.

[56] LifeTechnologiesCorporation, Quantifiler™ HP and Trio DNA Quantification Kits User Guide, (2017).

[57] C.M. Grgicak, Z.M. Urban, R.W. Cotton, Investigation of reproducibility and error associated with qPCR methods using Quantifiler® Duo DNA quantification kit, J. Forensic Sci. 55 (2010) 1331–1339.

[58] T. Kitayama, K. Fujii, H. Nakahara, N. Mizuno, K. Kasai, N. Yonezawa, et al., Estimation of the detection rate in STR analysis by determining the DNA degradation ratio using quantitative PCR, Legal Med. (Tokyo, Japan) 15 (2013) 1–6.

[59] W.R. Hudlow, M.D. Chong, K.L. Swango, M.D. Timken, M.R. Buoncristiani, A quadruplex real-time qPCR assay for the simultaneous assessment of total human DNA, human male DNA, DNA degradation and the presence of PCR inhibitors in forensic samples: a diagnostic tool for STR typing, Forensic Sci. Int. Genet. 2 (2008) 108–125.

[60] U.J. Monich, K. Duffy, M. Medard, V. Cadambe, L.E. Alfonse, C. Grgicak, Probabilistic characterisation of baseline noise in STR profiles, Forensic Sci. Int. Genet. 19 (2015) 107–122.

[61] C. Brookes, J.-A. Bright, S. Harbison, J. Buckleton, Characterising stutter in forensic STR multiplexes, Forensic Sci. Int. Genet. 6 (2012) 58–63.

[62] FBI, Quality Assurance Standards for Forensic DNA Testing Laboratories, (2009).

[63] J. Perez, A.A. mitchell, N. Ducasse, J. Tamariz, T. Caragine, Estimating the number of contributors to two-, three-, and four-person mixtures containing DNA, Croat. Med. J. 52 (2011) 314–326.

[64] T. Tvedebrink, On the exact distribution of the numbers of alleles in DNA mixtures, Int. J. Legal Med. 128 (2014) 427–437.

[65] C.A. Rakay, J. Bregu, C.M. Grgicak, Maximizing allele detection: effects of analytical threshold and DNA levels on rates of allele and locus drop-out, Forensic Sci. Int. Genet. 6 (2012) 723–728.

[66] L.G. Biesecker, J.E. Bailey-Wilson, J. Ballantyne, H. Baum, F.R. Bieber, C. Brenner, et al., DNA identifications after the 9/11 World Trade Center attack, Science 310 (2005) 1122–1123.