Forensic Science International: Genetics 38 (2019) 15–22

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Estimating number of contributors in massively parallel sequencing data of STR loci

Brian A Young[a,*], Katherine Butler Gettings[b], Bruce McCord[c], Peter M. Vallone[b]

[a] Nichevision Forensics, LLC., 526 South Main St., Akron, OH, 44311, USA
[b] U.S. National Institute of Standards and Technology, Biomolecular Measurement Division, 100 Bureau Drive, Gaithersburg, MD, 20899-8314, USA
[c] Department of Chemistry and Biochemistry and International Forensic Research Institute, Florida International University, University Park, Miami, FL, 33199, USA

ARTICLE INFO

Keywords:
Massively parallel sequencing
Number of contributors
Short tandem repeat

ABSTRACT

In recent years a number of computer-based algorithms have been developed for the deconvolution of complex DNA mixtures in forensic science. These procedures utilize likelihood ratios that quantify the evidence for a hypothesis for the presence of a person of interest in a DNA profile compared to an alternative hypothesis. Proper operation of these software systems requires an assumption regarding the total number of contributors present in the mixture. Unfortunately, estimates based on counting the number of alleles at a locus can be inaccurate due to the sharing and masking of alleles at individual loci. The effects of allele masking become increasingly severe as the number of contributors increases, rendering estimates about high-order mixtures uncertain. The accuracy of these estimates can be improved by increasing the number of STR markers in panels, and by using highly polymorphic markers. Increasing the number of STR markers from 13 to 20 (expanded CODIS panel) improves the accuracy of allele count-based estimation methods for low-order mixtures, but accuracy for high-order mixtures (> 3 contributors) remains poor due to allele masking. An alternative technique, massively parallel sequencing, holds great potential to improve the accuracy of the estimate of number of contributors due to its ability to detect sequence polymorphisms within alleles. This process results in an expansion of the number of alleles when compared to that obtained using capillary electrophoresis. Here, we show that the detection of these additional sequence-defined alleles in 22-marker panels improves number of contributor estimates in conceptual mixtures of 4 and 5 contributors.

1. Introduction

Often DNA samples recovered from a crime scene consist of mixtures of cells which may include the perpetrator, victim and an indeterminate number of additional individuals. The probability that a specific person of interest is included in the mixture can be calculated using the combined probability of inclusion (CPI) technique without assumptions about the total number of persons included in the mixture; although the assumed NOC should be specified in statements of inclusion or exclusion. However, the CPI calculation is valid only for simple mixtures and when DNA concentrations are sufficiently high that allele dropout is precluded [1]. Allele dropout can never be excluded with certainty. Therefore, alternatives to the CPI are being developed for improved mixture analysis. These alternatives include methods based on likelihood ratios (LR) including methods based on probabilistic genotyping [2,3].

In LR methods, the probability of the evidence is estimated given alternative hypotheses advanced by the prosecution and defense. However, several factors combine to obscure the true number of contributors in mixtures. These factors include allele masking and allele dropout, both of which act to reduce the number of alleles detected. In addition, stutter products of major contributor alleles can be confused with true alleles from minor contributors. Ambiguity of assignment of stutter can either increase or decrease the count of alleles. Because of these factors, the number of contributors to a mixture often cannot be known with certainty.

LR calculations require an assumption about the number of contributors (NOC) to a mixture [1]. Some mixture analysis software tools provide the user with automatic estimates of the NOC [4,5]. Alternatively, standalone software is available for estimating NOC using maximum likelihood [6], Bayesian [7,8] or machine learning [9] approaches. However, perhaps the most widely used method for estimating the number of contributors is the maximum allele count (MAC) method with a subjective assessment of peak heights [10–12]. The MAC

* Corresponding author at: NicheVision Forensics, LLC., 526 South Main St., Akron, OH, 44311, USA.
E-mail address: brian@nichevision.com (B.A. Young).

https://doi.org/10.1016/j.fsigen.2018.09.007
Received 27 June 2018; Received in revised form 12 September 2018; Accepted 24 September 2018
Available online 26 September 2018
1872-4973/

method identifies the minimum number of contributors necessary to explain the mixture. The true number of contributors is often larger than the minimum number due to allele masking and allele dropout [13–15]. Previous studies show that increasing the number of markers improves the accuracy of NOC estimates [16]; and that highly polymorphic markers are more informative [17].

Numerous studies have demonstrated the use of PCR-MPS methods for mixed DNA samples using a variety of sequencing platforms and forensic markers [18–23]. PCR-MPS (massively parallel sequencing) methods have the potential to improve the accuracy of NOC estimates over PCR-CE (capillary electrophoresis) methods. PCR-MPS methods can be highly multiplexed and currently-available forensic DNA analysis kits contain at a minimum the 20 autosomal STR loci in the expanded CODIS panel. The ForenSeq™ kit (Verogen™) includes 27 autosomal STR markers, while the Precision ID™ NGS kit (Applied Biosystems™) includes 31 autosomal STR markers and the PowerSeq™ Auto (Promega) includes 22 autosomal STR loci. Some of these kits also contain various numbers of Y-STR and X-STR markers. PCR-MPS methods are capable of typing STR loci based on nucleotide sequence rather than nucleotide length. Some loci are more polymorphic when alleles are defined based on sequence [24,25].

In this paper, we examine the potential advantages of PCR-MPS methods for estimating NOC. We compare the accuracy of NOC estimates derived from sequence-defined alleles and length-defined alleles in randomly-generated conceptual mixtures created from single-source sequence data derived from an earlier study [24]. Two specific allele counting methods are used to express the potential advantages of considering nucleotide sequence variation for estimating NOC. A direct comparison of various NOC estimation methods is not conducted. Rather, the focus is the effect of sequence variation on allele counts which is an important parameter in many different NOC estimation methods.

2. Materials and methods

2.1. Samples and sequencing

A total of 171 single-source samples were analyzed at 22 autosomal STR loci using a prototype Promega PowerSeq Auto/Y kit and an Illumina MiSeq sequencer as previously described [24]. DNA extracts from NIST population samples (n = 171) were selected to represent individuals of self-identified ancestry from three categories: African American (n = 66), Caucasian (n = 69), and Hispanic (N = 36). Twelve samples employed in the earlier study were excluded from this analysis either because CE concordance data were unavailable in [26], or because one or more allele was present below a minimum coverage of 10 reads.

2.2. Allele definitions

STR genotypes were generated for each of three allele definitions. Length-defined alleles were calculated from the number of nucleotides in the PCR amplicons. Allele lengths were converted to CE-equivalent sizes (aka allele numbers) according to standard ISFG sizing methods [27]. This allele definition is functionally equivalent to alleles generated by PCR-CE methods. Two different definitions were used for sequence-defined alleles: STR sequence and amplicon sequence. The STR sequence definition focuses on the nucleotide sequence of the STR locus proper. The genomic coordinates recommended by the International Society for Forensic Genetics (ISFG) DNA commission on minimal nomenclature requirements [28] and the STR sequence working group [29] were used to define the STR loci. The amplicon sequence definition utilized the nucleotide sequence of the entire PCR amplicon. To avoid false variation due to PCR-directed mutation, variation in the outermost 20 nucleotides were not considered in allelic diversity. STR locus sequences are a subset of whole amplicon sequences, with the difference being that whole-amplicon sequences include additional sequence diversity contributed by the amplicon regions flanking the STR locus. Whole-amplicon sequences include STR loci along with deletion/insertion polymorphisms (DIP) and SNP loci. Such multi-locus sequences are commonly termed haplotypes. For the purposes of this paper, variants under all three definitions are termed alleles, and are referred to as category 1, 2 and 3 alleles respectively.

2.3. Genotyping

Each of the 171 samples employed in this study was individually genotyped using MixtureAce™ software (NicheVision Forensics). Briefly, sequencer reads were classified according to locus using approximate matching of 20 nt sequence tags near the 5′ end of both forward and reverse reads. Approximate string matching is implemented in MixtureAce as a regular expression function that allowed for up to 2 mismatches. Reads corresponding to the GenBank bottom strand were reverse complemented and all reads within each locus cluster were sorted based on abundance of distinct sequences within the cluster. Abundance sorts were performed both on whole amplicon sequences, and on sequences of STR loci proper as delineated by ISFG. “Allele sequences were identified based on read count intensity above an analytical threshold of 10 reads. Stutter artifacts were filtered using the MixtureAce stutter filter.” All genotypes were manually reviewed for exceptionally high stutter missed by the stutter filters. Raw FASTQ-format files generated by the sequencing instrument were uploaded and genotyping results were output in CSV (comma separated values) format. The final CSV files contained exactly two alleles for each of the three allele definitions and for each of 22 autosomal STR loci. That is, homozygotes by any allele definition were represented by two identical alleles such that all sample profiles included 44 alleles. Genotyping accuracy was confirmed by comparing allele numbers for all 171 genotypes to published allele numbers generated by PCR-CE methods on the same samples [26]. Sequencing accuracy was confirmed by comparing STR locus sequences with published sequences for five positive controls. The positive controls included in the sequence data were NIST standard reference materials 2391c components A, B, C, E, and F [https://www-s.nist.gov/srmors/view_cert.cfm?srm=2391C]. These positive controls were not included in the 171 samples used for conceptual mixture construction.

2.4. Creation of conceptual mixtures

Genotypes of all 171 samples by all three allele definitions were combined into a single CSV file and loaded into R as a single data frame for bioinformatic analysis. Conceptual mixtures were constructed by bioinformatically combining individual sample genotypes to create all possible combinations without duplication using a custom R script. All but 10,000 combinations were randomly deleted from each sample space of all possible combinations using a random number generating function to assure that the remaining combinations were randomly spread across the sample space. This procedure was performed for each of the three allele definitions such that 10,000 random genotype combinations were generated for 2, 3, 4 and 5-way combinations of alleles for each of the three allele categories.

3. Results and discussion

Each of the 171 samples included in this study contained 22 PCR-targeted genomic regions yielding 44 PCR amplicons detected as sequencer reads. Amplicons (i.e. sequencer reads) were typed by each of three allelic categories. Thus, a total of 7524 chromosome segments were target-sequenced, and 22,572 allelic types were generated across the three allele categories. A total of 244, 396 and 462 distinct alleles were observed for allele categories 1, 2 and 3 respectively. PCR targets yielding identical types by one allelic category are homozygous by that allelic category but may be heterozygous by another allelic category.

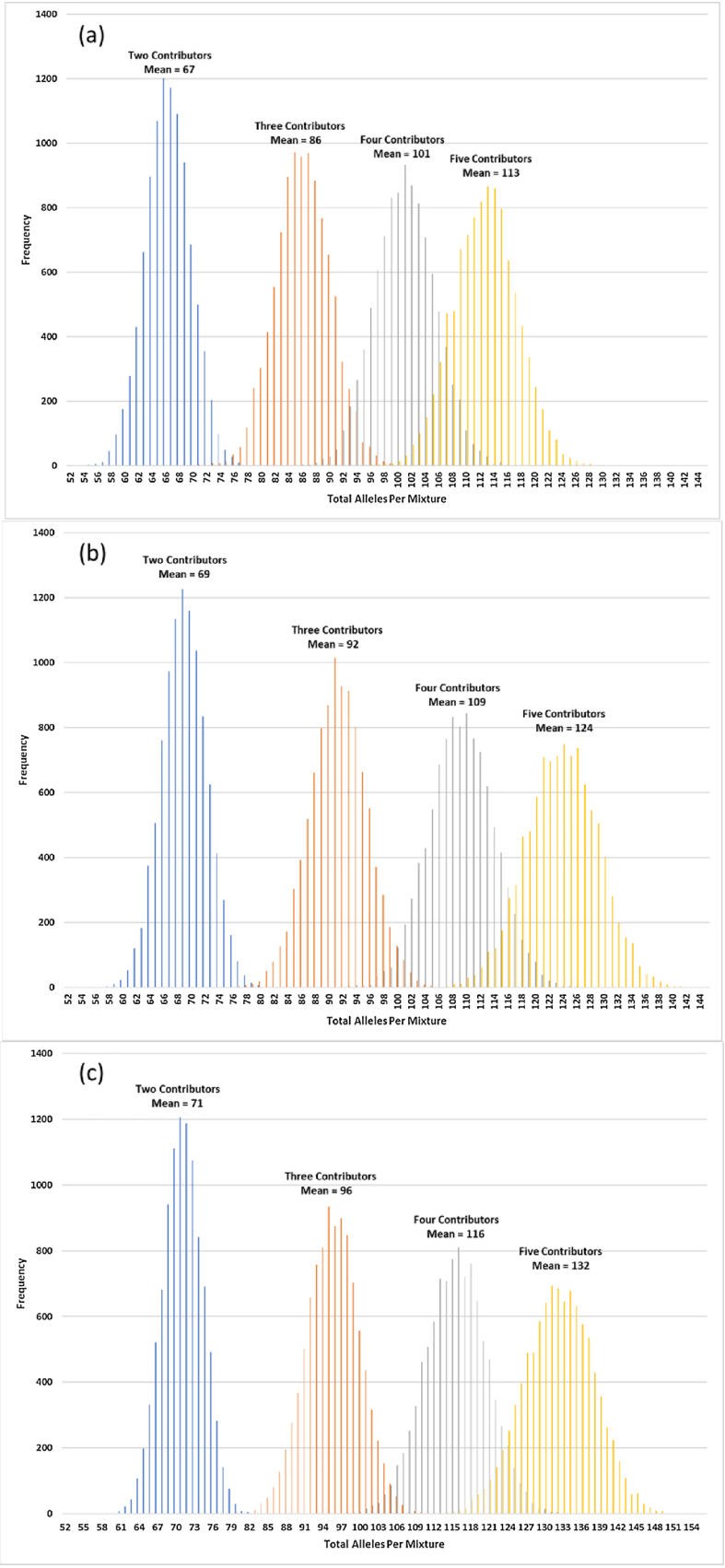

Fig. 1. Total distinct alleles per mixture when alleles are based on a) amplicon length, b) STR locus sequence or c) amplicon sequence. Each histogram represents 10,000 conceptual mixtures, and the means are displayed above each histogram.

Table 1
Rates at which the total allele count in conceptual mixtures containing N contributors are more likely under hypotheses that the number of contributors is either N − 1 or N + 1. Likelihoods were calculated using Normal distribution models.

Parameter	Allele Category											
	Category 1 Alleles				Category 2 Alleles				Category 3 Alleles			
True NOC	2	3	4	5	2	3	4	5	2	3	4	5
% More Likely Under N − 1	0.00%	0.25%	3.30%	7.58%	0.00%	0.14%	2.45%	6.81%	0.00%	0.03%	1.80%	5.27%
% More Likely Under N + 1	0.24%	2.57%	8.75%	0.00%	0.054%	1.15%	5.09%	0.00%	0.055%	0.83%	4.75%	0.00%
Total Error	0.24%	2.82%	12.05%	7.58%	0.054%	1.28%	7.54%	6.81%	0.055%	0.87%	6.55%	5.27%
Improvement	–	–	–	–	0.19%	1.54%	4.51%	0.77%	0.19%	1.96%	5.50%	2.31%

For purposes of direct comparison, 10,000 profiles were randomly selected for each of the four mixture levels. The number of theoretically possible mixed profiles increases exponentially with the number of contributors to the mixture. Thus, randomly selecting 10,000 samples from all 14,535 possible 2-way combinations of 171 samples represents a more complete sampling of the possible combinations than randomly selecting 10,000 samples from all 1.15×10^9 possible 5-way combinations of the same set of 171 samples. However, each of the 10,000-member distributions closely fit a Gaussian distribution, and by the law of large numbers 10,000 samples were sufficient to converge on the true means and standard deviations of the distribution of number of distinct alleles per mixture (data not shown).

3.1. Total distinct alleles per conceptual mixture

The total number of distinct alleles observed in mixed sample profiles has been a feature used for PCR-CE data when assigning NOC, e.g. [15]. The total alleles method exploits the fact that the total number of distinct alleles observed per mixture increases with increasing number of mixture contributors. At least for low-order mixtures, the distributions of the number of alleles do not overlap greatly (Fig. 1). While powerful for discriminating 2-contributor from 3-contributor mixtures, the discriminatory power of the total allele count feature decreases with increasing number of contributors to the mixture due to increasing overlap of the distributions [14,15]. Category 2 and 3 alleles improved the power of discrimination over category 1 alleles by the total alleles method.

An expected error rate calculation was used to quantify the increased power of discrimination, where expected error was defined as true N-contributor mixtures that would be classified as an N-1 or an N + 1 contributor mixture based on the total number of alleles observed and using the observed distributions of total alleles (Fig. 1). The classification decision rule was to assign a mixture to the NOC category exhibiting the highest likelihood for the observed number of alleles.

Likelihoods were calculated from data modeled as Normally distributed using the observed means and standard deviations. All fits of the data to Normal distributions were highly significant. Error rates due to allele masking decreased for category 2 and 3 alleles relative to category 1 alleles (Table 1). This trend confirms that sequence-defined alleles can provide better estimates of the number of contributors to a mixture when using total alleles as a discriminator.

3.2. Total distinct alleles per locus

The number of distinct alleles per locus is another feature of mixed DNA samples commonly used to estimate the number of contributors [1]. The maximum allele count (MAC) method is one method for assigning the number of contributors based on per-locus allele counts [10–12]. Central tenets of the method are that the number of contributors is best indicated by loci exhibiting the highest number of alleles, and that the minimum number of contributors that can explain the mixture is indicated by the rule set shown in Eq. (1)[1]. MAC method determinations can only indicate the minimum number of contributors to a mixture because the amount of allele masking (and allele dropout in casework samples) is usually unknown. Casework samples may also exhibit allele dropin which could inflate the minimum NOC estimate. While the numbers of contributors to the conceptual mixtures in this study is always known, the convention of expressing MAC method results in terms of the minimum NOC was preserved.

$$NOC \geq \begin{Bmatrix} integer\left(\frac{allele\ count}{ploidy}\right) when\ mod\left(\frac{allele\ count}{ploidy}\right) = 0 \\ integer\left(\left(\frac{allele\ count}{ploidy}\right) + 1\right) when\ mod\left(\frac{allele\ count}{ploidy}\right) \neq 0 \end{Bmatrix} \quad (1)$$

3.3. NOC accuracy for length- and sequence-based alleles

When the MAC method is applied to locus profiles, the minimum NOC required to explain the mixture is indicated by the loci exhibiting the largest number of alleles. Users may optionally consider peak heights when counting distinct alleles. The objective of this research was to measure the impact of various allele definitions on the accuracy of NOC estimates. A modification of the MAC method was used as a measure of accuracy where NOC determinations in conceptual mixtures are considered accurate when the minimum NOC corresponds to the known number of contributors in the mixture, and inaccurate otherwise. Peak heights were not considered in this definition of accuracy. Locus-specific allele counts for MAC accuracies for two-contributor mixtures are shown in Table 2 (see Supplementary Tables 1 through 4 for data at all mixture levels). When considering entire profiles, NOC determinations were 100% accurate for both two- and three-contributor conceptual mixtures for all allele categories (Fig. 2). This high rate of accuracy for low-order mixtures is consistent with previously published studies using expanded CODIS panels [17].

NOC accuracy based on category 1 alleles declines rapidly due to allele masking in mixtures of four and five contributors. Accuracy based on category 2 and 3 alleles declines less rapidly in these mixtures, indicating the value of these allele categories for the analysis of higher-order mixtures. This pattern suggest that category 3 alleles are only slightly more informative than category 2 alleles for assigning NOC by the MAC method. The major portion of increased informativeness comes from sequence variability within the STR locus.

3.4. Informativeness of loci

While category 2 and 3 alleles appear more informative for NOC designation than category 1 alleles, not all loci are equally informative. To measure the informativeness of loci for NOC, a correct assignment rate (CAR) metric was used. The CAR metric is defined as the

[1] Mod is the modulo(n, m) function which returns the remainder of division of n by m.

Table 2
Distinct allele counts by marker and allele type in 10,000 random two-contributor conceptual mixtures of 171 genotypes by each of three allele categories. Mixture counts are reported as the number of mixed samples in 10,000 two-contributor simulations exhibiting four, three, two or one allele. Percent change is expressed as the difference in counts of observed STR category 2 and category 3 alleles relative to category 1 alleles. Locus accuracy = the percentage of simulations for which the number of observed distinct alleles indicated the correct number of contributors. For two-contributor mixtures, accuracy is expressed as the percentage of mixtures exhibiting three or four distinct alleles.

		Allele Category							Allele Category			
Marker	Alleles Obs.	Category 1	Category 2		Category 3		Marker	Alleles Observed	Category 1	Category 2		Category 3	
		Mixture Count	Mixure Count	Change	Mixture Count	Change			Mixture Count	Mixture Count	Change	Mixture Count	Change
CSF1PO	4	1,300	1,393	0.93%	1,427	1.27%	D2S441	4	1,763	2,961	11.98%	3,088	13.25%
CSF1PO	3	5,070	5,103	0.33%	5,084	0.14%	D2S441	3	5,421	5,154	−2.67%	5,158	−2.63%
CSF1PO	2	3,379	3,276	−1.03%	3,261	−1.18%	D2S441	2	2,661	1,806	−8.55%	1,691	−9.70%
CSF1PO	1	251	228	−0.23%	228	−0.23%	D2S441	1	155	79	−0.76%	63	−0.92%
Locus Accuracy		**63.70%**	**64.96%**	**–**	**65.11%**	**–**	**Locus Accuracy**		**71.84%**	**81.15%**	**–**	**82.46%**	**–**
D10S1248	4	1,768	1,810	0.42%	1,810	0.42%	D3S1358	4	1,603	4,852	32.49%	4,852	32.49%
D10S1248	3	5,396	5,419	0.23%	5,419	0.23%	D3S1358	3	5,275	4,342	−9.33%	4,342	−9.33%
D10S1248	2	2,705	2,649	−0.56%	2,649	−0.56%	D3S1358	2	2,947	783	−21.64%	783	−21.64%
D10S1248	1	131	122	−0.09%	122	−0.09%	D3S1358	1	175	23	−1.52%	23	−1.52%
Locus Accuracy		**71.64%**	**72.29%**	**–**	**72.29%**	**–**	**Locus Accuracy**		**68.78%**	**91.94%**	**–**	**91.94%**	**–**
D12S391	4	4,325	7,158	28.33%	7,158	28.33%	D5S818	4	1,055	1,082	0.27%	5,000	39.45%
D12S391	3	4,549	2,568	−19.81%	2,568	−19.81%	D5S818	3	4,747	4,727	−0.20%	4,123	−6.24%
D12S391	2	1,085	270	−8.15%	270	−8.15%	D5S818	2	3,848	3,841	−0.07%	836	−30.12%
D12S391	1	41	4	−0.37%	4	−0.37%	D5S818	1	350	350	0.00%	41	−3.09%
Locus Accuracy		**88.74%**	**97.26%**	**–**	**97.26%**	**–**	**Locus Accuracy**		**58.02%**	**58.09%**	**–**	**91.23%**	**–**
D13S317	4	1,877	1,906	0.29%	4,107	22.30%	D7S820	4	2,433	2,372	−0.61%	4,487	20.54%
D13S317	3	5,108	4,999	−1.09%	4,688	−4.20%	D7S820	3	5,488	5,490	0.02%	4,379	−11.09%
D13S317	2	2,811	2,894	0.83%	1,165	−16.46%	D7S820	2	1,953	2,010	0.57%	1,065	−8.88%
D13S317	1	204	201	−0.03%	40	−1.64%	D7S820	1	126	128	0.02%	69	−0.57%
Locus Accuracy		**69.85%**	**69.05%**	**–**	**87.95%**	**–**	**Locus Accuracy**		**79.21%**	**78.62%**	**–**	**88.66%**	**–**
D16S539	4	1,757	1,757	0.00%	4,005	22.48%	D8S1179	4	2,093	4,724	26.31%	4,737	26.44%
D16S539	3	5,478	5,478	0.00%	4,842	−6.36%	D8S1179	3	5,319	4,450	−8.69%	4,439	−8.80%
D16S539	2	2,580	2,580	0.00%	1,111	−14.69%	D8S1179	2	2,464	814	−16.50%	812	−16.52%
D16S539	1	185	185	0.00%	42	−1.43%	D8S1179	1	124	12	−1.12%	12	−1.12%
Locus Accuracy		**72.35%**	**72.35%**	**–**	**88.47%**	**–**	**Locus Accuracy**		**74.12%**	**91.74%**	**–**	**91.76%**	**–**
D18S51	4	4,298	4,436	1.38%	4,436	1.38%	FGA	4	4,069	4,092	0.23%	4,092	0.23%
D18S51	3	4,736	4,644	−0.92%	4,644	−0.92%	FGA	3	4,751	4,730	−0.21%	4,730	−0.21%
D18S51	2	936	892	−0.44%	892	−0.44%	FGA	2	1,138	1,136	−0.02%	1,136	−0.02%
D18S51	1	30	28	−0.02%	28	−0.02%	FGA	1	42	42	0.00%	42	0.00%
Locus Accuracy		**90.34%**	**90.80%**	**–**	**90.80%**	**–**	**Locus Accuracy**		**88.20%**	**88.22%**	**–**	**88.22%**	**–**
D19S433	4	2,596	2,630	0.34%	2,630	0.34%	PentaD	4	4,041	3,928	−1.13%	4,206	1.65%
D19S433	3	4,931	4,919	−0.12%	4,919	−0.12%	PentaD	3	4,865	4,945	0.80%	4,776	−0.89%
D19S433	2	2,290	2,268	−0.22%	2,268	−0.22%	PentaD	2	1,062	1,094	0.32%	987	−0.75%
D19S433	1	183	183	0.00%	183	0.00%	PentaD	1	32	33	0.01%	31	−0.01%
Locus Accuracy		**75.27%**	**75.49%**	**–**	**75.49%**	**–**	**Locus Accuracy**		**89.06%**	**88.73%**	**–**	**89.82%**	**–**
D1S1656	4	4,754	5,682	9.28%	5,682	9.28%	PentaE	4	4,967	4,999	0.32%	4,999	0.32%
D1S1656	3	4,409	3,718	−6.91%	3,718	−6.91%	PentaE	3	4,276	4,246	−0.30%	4,246	−0.30%
D1S1656	2	807	578	−2.29%	578	−2.29%	PentaE	2	728	726	−0.02%	726	−0.02%
D1S1656	1	30	22	−0.08%	22	−0.08%	PentaE	1	29	29	0.00%	29	0.00%
Locus Accuracy		**91.63%**	**94.00%**	**–**	**94.00%**	**–**	**Locus Accuracy**		**92.43%**	**92.45%**	**–**	**92.45%**	**–**
D21S11	4	3,494	5,775	22.81%	5,775	22.81%	TH01	4	1,521	1,521	0.00%	1,945	4.24%
D21S11	3	4,900	3,650	−12.50%	3,650	−12.50%	TH01	3	5,242	5,242	0.00%	5,150	−0.92%
D21S11	2	1,540	561	−9.79%	561	−9.79%	TH01	2	2,970	2,970	0.00%	2,662	−3.08%
D21S11	1	66	14	−0.52%	14	−0.52%	TH01	1	267	267	0.00%	243	−0.24%
Locus Accuracy		**83.94%**	**94.25%**	**–**	**94.25%**	**–**	**Locus Accuracy**		**67.63%**	**67.63%**	**–**	**70.95%**	**–**
D22S1045	4	1,684	1,684	0.00%	1,734	0.50%	TPOX	4	1,063	1,063	0.00%	1,397	3.34%
D22S1045	3	5,087	5,087	0.00%	5,059	−0.28%	TPOX	3	4,562	4,562	0.00%	4,449	−1.13%
D22S1045	2	3,038	3,038	0.00%	3,016	−0.22%	TPOX	2	3,879	3,879	0.00%	3,660	−2.19%
D22S1045	1	191	191	0.00%	191	0.00%	TPOX	1	496	496	0.00%	494	−0.02%
Locus Accuracy		**67.71%**	**67.71%**	**–**	**67.93%**	**–**	**Locus Accuracy**		**56.25%**	**56.25%**	**–**	**58.46%**	**–**
D2S1338	4	5,014	7,067	20.53%	7,098	20.84%	VWA	4	2,526	4,119	15.93%	4,119	15.93%
D2S1338	3	4,354	2,682	−16.72%	2,655	−16.99%	VWA	3	5,575	4,757	−8.18%	4,757	−8.18%
D2S1338	2	625	248	−3.77%	244	−3.81%	VWA	2	1,841	1,084	−7.57%	1,084	−7.57%
D2S1338	1	7	3	−0.04%	3	−0.04%	VWA	1	58	40	−0.18%	40	−0.18%
Locus Accuracy		**93.68%**	**97.49%**	**–**	**97.53%**	**–**	**Locus Accuracy**		**81.01%**	**88.76%**	**–**	**88.76%**	**–**

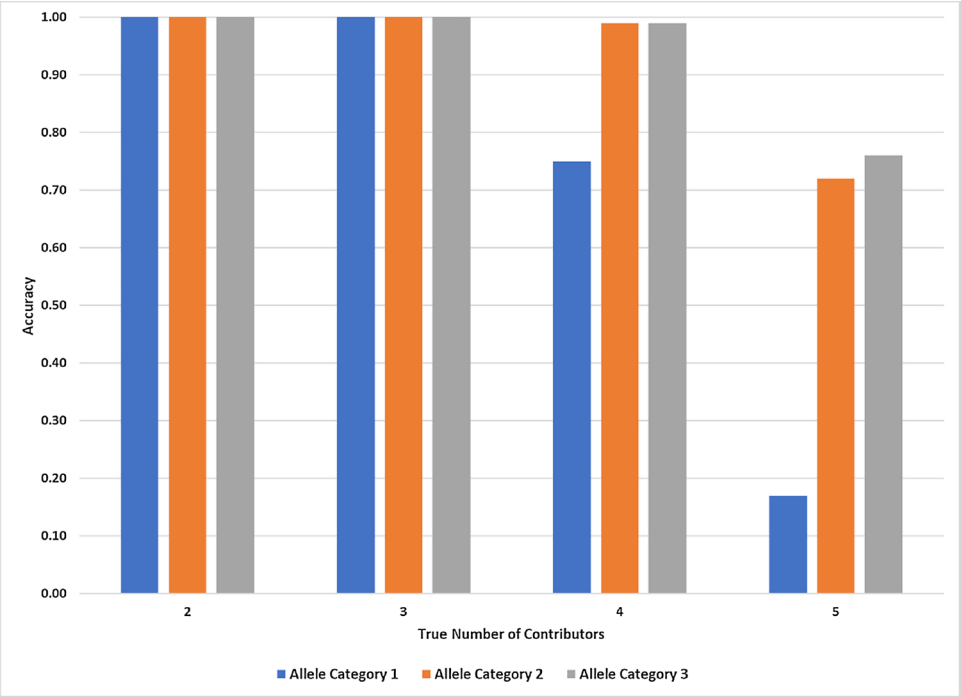

Fig. 2. Percent accuracy in designating the number of contributors to conceptual mixtures containing two-, three-, four- and five-contributors using the maximum allele count (MAC) method. Assignments were considered accurate when the minimum NOC indicated by the MAC method was equivalent to the true NOC. Accuracy was calculated for each of three allele category definitions: 1) amplicon length, 2) nucleotide sequence of the STR locus, 3) nucleotide sequence of whole amplicons. Each accuracy percentage was calculated from 10,000 random mixtures.

Table 3
Average accuracy in estimating the number of contributors (NOC) by the maximum allele count (MAC) method as a function of allele category and level of mixture. NOC assignments were considered accurate when the minimum NOC indicated by the MAC method was equivalent to the true NOC.

Number of Contributors	Allele Category	Average Accuracy Across All Loci
2	**1**	77.06%
	2	80.87%
	3	84.81%
3	**1**	27.11%
	2	37.89%
	3	45.77%
4	**1**	5.71%
	2	14.30%
	3	17.41%
5	**1**	0.80%
	2	4.93%
	3	5.49%

proportion of conceptual mixtures for which a locus indicates a minimum NOC by the MAC method that is equivalent to the true NOC. In casework samples, the CAR metric would integrate the effects of allele masking, dropout, stutter and dropin. In conceptual mixtures with known genotypes, the effects of dropout, stutter and dropin are removed as variables, isolating allele masking as the driver for differences in locus informativeness.

Informativeness of loci for NOC designation by the MAC method declines with increasing number of contributors. However, at all mixture levels, category 2 and 3 alleles outperform category 1 alleles (Table 3). Individual loci that are relatively uninformative by length, tend also to be relatively uninformative by sequence (Table 4). However, significant exceptions exist where sequence diversity greatly improves the informativeness of a locus. The sequence diversity can be primarily in the STR locus (e.g. D8S1179) or in the amplicon regions flanking the STR locus (e.g. D5S818). Informativeness levels for all loci in all mixtures are included in Supplementary Table 5.

Table 4
Average accuracy of loci for number of contributors using the maximum allele count method across all conceptual mixture levels. Accuracy on a per-locus basis is expressed in terms of correct assignment rate representing the percentage of conceptual mixtures in which the locus indicated the correct number of contributors by the maximum allele count method. Loci are sorted in decreasing order of informativeness by category 3 allele type.

Marker	Average Correct Assignment Rate Across Mixture Levels By Allele Category		
	1	2	3
D12S391	40.19%	70.99%	70.99%
D2S1338	43.43%	65.86%	66.44%
D21S11	32.33%	54.94%	54.94%
D1S1656	43.36%	53.59%	53.59%
PentaE	47.33%	47.86%	47.86%
D5S818	16.20%	16.35%	45.64%
D3S1358	19.82%	43.66%	43.66%
D8S1179	22.99%	43.40%	43.55%
D18S51	38.88%	40.01%	40.01%
D7S820	24.50%	24.16%	39.96%
D13S317	21.41%	21.33%	38.96%
PentaD	35.98%	35.24%	37.25%
VWA	25.53%	36.37%	36.37%
FGA	35.60%	35.88%	35.88%
D16S539	20.77%	20.77%	34.81%
D2S441	21.24%	28.22%	29.16%
D19S433	25.61%	25.84%	25.84%
TH01	19.10%	19.10%	21.61%
D10S1248	20.98%	21.26%	21.26%
D22S1045	20.33%	20.33%	20.55%
CSF1PO	17.78%	18.34%	18.49%
TPOX	15.45%	15.45%	17.36%

Locus informativeness for NOC has been previously shown to be correlated with the degree of polymorphism [17]. Here, we consider heterozygosity of loci. Locus informativeness for NOC determination was found to be positively correlated with locus heterozygosity, but the correlation declines with increasing mixture level as increasing

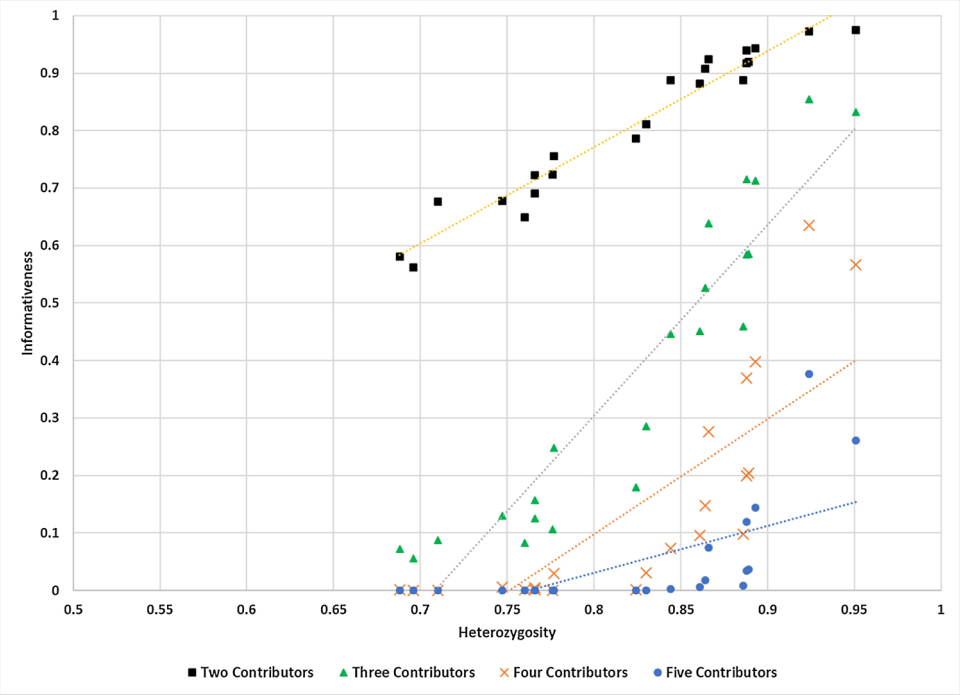

Fig. 3. Locus informativeness versus locus heterozygosity based on STR locus sequence and at 4 levels of conceptual mixture. Locus heterozygosities are from Gettings et al., 2016 [15] for the same data set.

numbers of loci exhibit "zero" informativeness due to allele masking (i.e. correct assignment rate < 1/10,000 or 0.01%) (Fig. 3).

4. Conclusions

In this study we have demonstrated that sequence-defined forensic STR alleles are more informative for NOC estimates by allele counting methods. This increased informativeness is due to higher heterozygosity of alleles based on sequence rather than length, which results in lower levels of allele masking. This study isolated the effect of allele masking and indicates that PCR-MPS methods will be more accurate for NOC by this measure. However, the level of allele dropout, dropin in PCR-MPS methods will also impact the utility of PCR-MPS for estimating NOC. These parameters have not yet been extensively studied. In this study accuracy in estimating the true NOC to a sample was measured by two different allele counting methods. The increase in heterozygosity due to sequence variation resulted in increased accuracy of NOC estimates by both methods. The effect is most pronounced in the MAC method, where accuracy declines rapidly in category 1 alleles in higher-order mixtures. By contrast, accuracy of estimates based on the total allele count method declines only slightly as mixture complexity increases; and all three categories of alleles decline by similar amounts. The increased heterozygosity of PCR-MPS should also benefit other methods for estimating NOC to include machine learning and maximum likelihood methods. Overall these results demonstrate that PCR-MPS can provide improved results when estimating NOC and suggest that the application of additional heterozygous loci can improve NOC estimates in the future.

Disclaimer

Points of view in this document are those of the authors (PMV and KGB) and do not necessarily represent the official position or policies of the U.S. Department of Commerce or the Department of Justice. Certain commercial equipment, instruments, and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by NIST, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose.

All work presented has been reviewed and approved by the NIST Human Subjects Protections Office.

Acknowledgements

This work was supported by internal funding at NicheVision Forensics (BAY and BM) and internal funding from the NIST Special Programs Office for Forensic DNA (PMV and KBG), an interagency agreement with the FBI Biometric Center of Excellence (PMV and KBG).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.fsigen.2018.09.007.

References

[1] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, N. Morling, M. Prinz, P.M. Schneider, B.S. Weir, DNA commission of the international society of forensic genetics: recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2006) 90–101, https://doi.org/10.1016/j.forsciint.2006.04.009.

[2] M.W. Perlin, A. Sinelnikov, An information gap in DNA evidence interpretation, PLoS One 4 (2009), https://doi.org/10.1371/journal.pone.0008327.

[3] J.A. Bright, R. Richards, M. Kruijver, H. Kelly, C. McGovern, A. Magee, A. McWhorter, A. Ciecko, B. Peck, C. Baumgartner, C. Buettner, S. McWilliams, C. McKenna, C. Gallacher, B. Mallinder, D. Wright, D. Johnson, D. Catella, E. Lien, C. O'Connor, G. Duncan, J. Bundy, J. Echard, J. Lowe, J. Stewart, K. Corrado, S. Gentile, M. Kaplan, M. Hassler, N. McDonald, P. Hulme, R.H. Oefelein, S. Montpetit, M. Strong, S. Noël, S. Malsom, S. Myers, S. Welti, T. Moretti, T. McMahon, T. Grill, T. Kalafut, M.M. Greer-Ritzheimer, V. Beamer, D.A. Taylor, J.S. Buckleton, Internal validation of STRmix™ – a multi laboratory response to PCAST, Forensic Sci. Int. Genet. 34 (2018) 11–24, https://doi.org/10.1016/j.fsigen.2018.01.003.

[4] C.H. Brenner, DNA-View, (n.d.). dna-view.com (Accessed 12 June 2018).

[5] S. Manabe, C. Morimoto, Y. Hamano, S. Fujimoto, K. Tamaki, Development and validation of open-source software for DNA mixture interpretation based on a quantitative continuous model, PLoS One 12 (2017) 1–18, https://doi.org/10.1371/journal.pone.0188183.

[6] H. Haned, L. Pène, J.R. Lobry, A.B. Dufour, D. Pontier, Estimating the number of

contributors to forensic DNA Mixtures: does maximum likelihood perform better than maximum allele count? J. Forensic Sci. 56 (2011) 23–28, https://doi.org/10.1111/j.1556-4029.2010.01550.x.
[7] 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, https://doi.org/10.1016/j.fsigen.2014.11.010.
[8] T. Tvedebrink, On the exact distribution of the numbers of alleles in DNA mixtures, Int. J. Legal Med. 128 (2014) 427–437, https://doi.org/10.1007/s00414-013-0951-3.
[9] 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, https://doi.org/10.1016/j.fsigen.2016.11.006.
[10] T.M. Clayton, J.P. Whitaker, R. Sparkes, P. Gill, Analysis and interpretation of mixed forensic stains using DNA STR profiling, Forensic Sci. Int. 91 (1998) 55–70, https://doi.org/10.1016/S0379-0738(97)00175-8.
[11] B. Budowle, A.J. Onorato, T.F. Callaghan, A. Della Manna, A.M. Gross, R.A. Guerrieri, J.C. Luttman, D.L. Mcclure, Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed dna profiles in forensic casework, J. Forensic Sci. 54 (2009) 810–821, https://doi.org/10.1111/j.1556-4029.2009.01046.x.
[12] 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) 1–15, https://doi.org/10.1186/s12863-016-0429-7.
[13] D.R. Paoletti, T.E. Doom, C.M. Krane, M.L. Raymer, D.E. Krane, Empirical analysis of the STR profiles resulting from conceptual mixtures, J. Forensic Sci. 50 (2005) 1361–1366, https://doi.org/10.1520/JFS2004475.
[14] J.S. Buckleton, J.M. Curran, P. Gill, Towards understanding the effect of uncertainty in the number of contributors to DNA stains, Forensic Sci. Int. Genet. 1 (2007) 20–28, https://doi.org/10.1016/j.fsigen.2006.09.002.
[15] 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 in high template and low template amounts, Croat. Med. J. 52 (2011) 314–326, https://doi.org/10.3325/cmj.2011.52.314.
[16] M.D. Coble, J.A. Bright, J.S. Buckleton, J.M. Curran, Uncertainty in the number of contributors in the proposed new CODIS set, Forensic Sci. Int. Genet. 19 (2015) 207–211, https://doi.org/10.1016/j.fsigen.2015.07.005.
[17] G.M. Dembinski, C. Sobieralski, C.J. Picard, Estimation of the number of contributors of theoretical mixture profiles based on allele counting: does increasing the number of loci increase success rate of estimates? Forensic Sci. Int. Genet. 33 (2018) 24–32, https://doi.org/10.1016/j.fsigen.2017.11.007.
[18] C. Van Neste, F. Van Nieuwerburgh, D. Van Hoofstat, D. Deforce, Forensic STR analysis using massive parallel sequencing, Forensic Sci. Int. Genet. 6 (2012) 810–818, https://doi.org/10.1016/j.fsigen.2012.03.004.
[19] C. Børsting, S.L. Fordyce, J. Olofsson, H.S. Mogensen, N. Morling, Evaluation of the Ion Torrent™ HID SNP 169-plex: a SNP typing assay developed for human identification by second generation sequencing, Forensic Sci. Int. Genet. 12 (2014) 144–154, https://doi.org/10.1016/j.fsigen.2014.06.004.
[20] X. Zeng, J. King, S. Hermanson, J. Patel, D.R. Storts, B. Budowle, An evaluation of the PowerSeq™ auto system: a multiplex short tandem repeat marker kit compatible with massively parallel sequencing, Forensic Sci. Int. Genet. 19 (2015) 172–179, https://doi.org/10.1016/j.fsigen.2015.07.015.
[21] F. Guo, Y. Zhou, F. Liu, J. Yu, H. Song, H. Shen, B. Zhao, F. Jia, G. Hou, X. Jiang, Evaluation of the early access STR kit v1 on the ion torrent PGM™ platform, Forensic Sci. Int. Genet. 23 (2016) 111–120, https://doi.org/10.1016/j.fsigen.2016.04.004.
[22] J.D. Churchill, M. Stoljarova, J.L. King, B. Budowle, Massively parallel sequencing-enabled mixture analysis of mitochondrial DNA samples, Int. J. Legal Med. (2018) 1263–1272, https://doi.org/10.1007/s00414-018-1799-3.
[23] M.R. Lindberg, S.E. Schmedes, F.C. Hewitt, J.L. Haas, K.L. Ternus, D.R. Kadavy, B. Budowle, A comparison and integration of MiSeq and MinION platforms for sequencing single source and mixed mitochondrial genomes, PLoS One 11 (2016) 1–16, https://doi.org/10.1371/journal.pone.0167600.
[24] K.B. Gettings, K.M. Kiesler, S.A. Faith, E. Montano, C.H. Baker, B.A. Young, R.A. Guerrieri, P.M. Vallone, Sequence variation of 22 autosomal STR loci detected by next generation sequencing, Forensic Sci. Int. Genet. 21 (2016) 15–21, https://doi.org/10.1016/j.fsigen.2015.11.005.
[25] N.M.M. Novroski, J.L. King, J.D. Churchill, L.H. Seah, B. Budowle, Characterization of genetic sequence variation of 58 STR loci in four major population groups, Forensic Sci. Int. Genet. 25 (2016) 214–226, https://doi.org/10.1016/j.fsigen.2016.09.007.
[26] C.R. Steffen, M.D. Coble, K.B. Gettings, P.M. Vallone, Corrigendum to 'U.S. Population Data for 29 Autosomal STR Loci' [Forensic Sci. Int. Genet. 7 (2013) e82–e83](S1872497312002712)(10.1016/j.fsigen.2012.12.004), Forensic Sci. Int. Genet. 31 (2017) e36–e40. doi:https://doi.org/10.1016/j.fsigen.2017.08.011.
[27] W. Bär, B. Brinkmann, B. Budowle, A. Carracedo, P. Gill, P. Lincoln, W.R. Mayr, B. Olaisen, DNA recommendations. Further report of the DNA commission of the ISFG regarding the use of short tandem repeat systems, Forensic Sci. Int. 87 (1997) 179–184, https://doi.org/10.1007/s004140050061.
[28] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, L. Gusmão, D.R. Hares, J.A. Irwin, J.L. King, P. De Knijff, N. Morling, M. Prinz, P.M. Schneider, C. Van Neste, S. Willuweit, C. Phillips, Massively parallel sequencing of forensic STRs: considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements, Forensic Sci. Int. Genet. 22 (2016) 54–63, https://doi.org/10.1016/j.fsigen.2016.01.009.
[29] C. Phillips, K.B. Gettings, J.L. King, D. Ballard, M. Bodner, L. Borsuk, W. Parson, "The devil's in the detail": release of an expanded, enhanced and dynamically revised forensic STR sequence guide, Forensic Sci. Int. Genet. 34 (2018) 162–169, https://doi.org/10.1016/j.fsigen.2018.02.017.