NicheVision Forensics, LLC

Technical Note
Allele Frequencies for
IDSeek® OmniSTR™ and IDSeek® CombiSTR Plus™ Kits

May 2026

Summary

Sequence-based frequencies can differ depending on factors not usually impactful for length-based frequencies. These include the forensic kit used, the read trimming protocol, and the design of the relevant allele frequency database. Unlike length-based allele frequencies, sequence-based allele frequencies cannot always be taken directly from an allele frequency database. This is true for forensic kits that amplify genomic regions different from those covered in the database, but it is also true if the bioinformatic trim protocol trims to different genomic regions than covered by the database. Derivation of sequence-based allele frequencies from the NIST 1036 allele frequency database (1, 2) is described here for the IDseek® OmniSTR™ kit and the autosomal loci in the IDSeek® CombiSTR Plus™ kit.

Sequence-Based Allele Frequencies

The objective of forensic DNA analysis via fragment analysis (PCR-CE) and sequence analysis (PCR-MPS) is to detect the PCR amplicons generated in a PCR step. Alleles in a profile will have different frequencies depending on whether the feature measured is length or sequence. **Figure *1*** shows the NIST database frequencies derived from a survey of 1036 individuals by length (3) and by sequence (1, 2). Nine and 19 distinct alleles were identified by their length and sequence features respectively. Sequence-based alleles always have frequencies that are equal to or lower than length-based alleles for the same allele number. For example, there is only one known sequence for allele number 8 alleles in the database. Thus, both length-based and sequence-based alleles of length-8 have equivalent frequencies (0.0212). By contrast allele number 9 alleles have only one length (by definition) but exhibit four different known sequences. Consequently, the frequencies of allele number 9 alleles are divided among the four different distinct sequences. If additional sequence-based alleles are discovered, the frequency will be further subdivided.

Length-Based Alleles		Sequence-Based Alleles				
		SEQ Shorthand			Frequencies	
Allele	Freq.	5' Flank	STR	3'Flank	SEQ	LEN
5.0	0.00048		[GATA]5	rs11642858	0.00048	**0.00048**
6.0						
6.3			[GATA]8		0.0212	**0.0212**
7.0						
8.0	0.0212		[GATA]9	rs11642858	0.1134	
8.1			[GATA]9		0.0483	
9.0	0.1626		GGTA [GATA]8		0.0005	
9.1		rs563997442	[GATA]9		0.0005	**0.1626**
9.3						
10.0	0.1081		[GATA]10	rs11642858	0.0772	
10.1			[GATA]10		0.0304	
10.2			[GATA]5 GACA [GATA]4		0.0005	**0.1081**
10.3						
11.0	0.2915		[GATA]11		0.2698	
11.2			[GATA]11	rs11642858	0.0174	
11.3			[GATA]11	rs114697632	0.0043	**0.2915**
12.0	0.2568					
12.2			[GATA]12		0.2539	
12.3			[GATA]12	rs11642858	0.0024	
13.0	0.1371		[GATA]12	rs114697632	0.0005	**0.2568**
13.2						
13.3			[GATA]13		0.1371	**0.1371**
13.4						
14.0	0.0217		[GATA]14		0.0212	
14.2			[GATA]14	rs11642858	0.0005	**0.0217**
14.3						
15.0	0.0005		[GATA]15		0.0005	**0.0005**
Sum	**1.0000**				**1.00000**	**1.00000**

Figure 1. Allele frequencies by length and sequence as reported in the NIST 1036 allele frequency database.

Sequence-Based Allele Frequencies Can Depend on the Forensic Kit

Different forensic kits have different primer placements. This is true for both PCR-CE and PCR-MPS kits. Differences in primer placement in PCR-CE kits rarely makes a difference in allele frequencies because there are few length-affecting (i.e., INDEL) variants that can be inside one kit primer pair and outside another kit primer pair. NIST, FBI and others validate their length-base allele frequency databases using all the major PCR-CE kits and rarely discover an inter kit difference (usually called a microvariant). By contrast, many sequence-affecting (i.e., SNP) variants exist that can be inside one kit primer pair but outside another (**Table *1***). Thus, two different PCR-MPS kits can generate amplicons from the same sample that differ in their sequence. Stated differently, the detected haplotypes are different, where haplotype is defined as the allelic state of the set of polymorphisms (STRs, SNPs, INDELs) contained in a DNA fragment. FIGURE illustrates this situation with the ForenSeq and OmniSTR amplicons covering the D5S818 locus. ForenSeq amplicons cover D5S818 and rs73801920 as does the NIST 1036 database. Frequencies of ForenSeq amplicons/alleles can be taken directly from the database. OmniSTR amplicons cover a larger genomic extent and include SNP loci not included in the database. SNPs rs541272009 and rs25768 were not covered in the database, meaning that the allele state of these two SNPs was not assessed. Thus, OmniSTR kit amplicons do not have a database frequency. However, the intersection of the kit amplicon with the database genomic extent does have a database-derived frequency (**Figure *2***). Notably, greater than 25% of humans exhibit the minor (C) allele at rs25768, making it likely that individuals surveyed for the database had a minor (non-reference) allele at that position.

Table 1. Count of length-affecting (INDEL) and sequence-affecting (SNP) polymorphisms known to exist at a population frequency >1% in the examined regions of the ForenSeq, PowerSeq and OmniSTR kits. Data are from the forensic sequence structure guide (https://strider.online/).

Polymorphism Type	Counts of Polymorphisms With Population Frequencies > 1%					
	Included in Kit, but not NIST 1036 Database			Included in NIST 1036 Database, but not Kit		
	ForenSeq	PowerSeq	OmniSTR	ForenSeq	PowerSeq	OmniSTR
INDEL	0	0	0	0	1	1
SNP	0	29	12	0	10	16

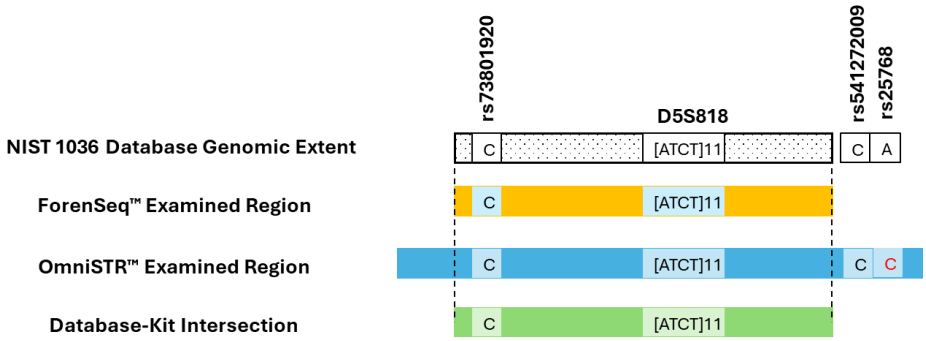

Figure 2. Illustration of amplicons covering D5S818 generated by the ForenSeq and OmniSTR kits.

Sequence-Based Allele Frequencies Depend on Bioinformatic Trimming

Bioinformatic trimming results in the same type of kit-to-database discordances inherent to primer placement. Trimming flanking regions around the STR locus can eliminate SNP positions that contribute to allele definitions. After trimming two sequences that were distinct due to a flank-located variant can become identical. This can be illustrated using amplicons covering the D5S818 locus. If the flanks are trimmed leaving only the traditional STR repeat region, SNP rs73801920 position will be eliminated. For example, this causes whole amplicon alleles 8_AN and 8_BL to coalesce into an identical sequence designated 8_WJ (**Table *2***).

Table 2. Sequence-based allele frequencies in the NIST 1036 database for the genomic extent covering D5S818. Trimming away the 5′ flank eliminates 7 of 19 alleles with a concomitant increase in the allele frequency of the coalesced alleles.

Locus	Allele	SID-2D		Frequencies		Counts	Database Sequence		
		STR	Amplicon	STR	Amplicon		5' Flank	Repeat Region	3' Flank
D5S818	7	CY	GO	0.0106	0.0106	22	TATTTATACCTCT	ATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	8	**WJ**	XQ	**0.0203**	0.0179	37	TATTTATACATCT	ATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	8	**WJ**	BL		0.0024	5	TATTTATACCTCT	ATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	9	**BF**	AN	**0.0463**	0.0434	90	TATTTATACATCT	ATCTATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	9	**BF**	WB		0.0029	6	TATTTATACCTCT	ATCTATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	10	**HZ**	KG	**0.0777**	0.0579	120	TATTTATACCTCT	ATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	10	**HZ**	TI		0.0198	41	TATTTATACATCT	ATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	11	**LZ**	VW	**0.3156**	0.2872	595	TATTTATACCTCT	ATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	11	**LZ**	BU		0.0285	59	TATTTATACATCT	ATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	12	**FI**	BU	**0.3533**	0.2539	526	TATTTATACCTCT	ATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	12	**FI**	QQ		0.0994	206	TATTTATACATCT	ATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	13	**NJ**	VC	**0.1540**	0.1153	239	TATTTATACCTCT	ATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	13	**NJ**	HB		0.0386	80	TATTTATACATCT	ATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	13	YM	RB	0.0092	0.0092	19	TATTTATACCTCT	ATCTATCTATCTATGTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	14	**QC**	ER	**0.0101**	0.0082	17	TATTTATACCTCT	ATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	14	**QC**	NA		0.0019	4	TATTTATACATCT	ATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	14	WN	MJ	0.0010	0.0010	2	TATTTATACCTCT	ATCTATCTATCTATGTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	15	QQ	KY	0.0014	0.0014	3	TATTTATACCTCT	ATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
D5S818	15	LX	UQ	0.0005	0.0005	1	TATTTATACCTCT	ATCTATCTATCTATGTATCTATCTATCTATCTATCTATCTATCTATCT	TCAAAAT
				1.0000	**1.0000**	**2072**			

Frequency Calculations

Allele frequencies are calculated in an intuitive way as the count of an allele observed in a sampled population divided by the total number of chromosomes in the sample. For autosomal STRs, the number of chromosomes in the sample is 2N where N is the number of people in the sample. Allele frequencies must be non-zero when used in genotype, profile, and forensic statistical strength of evidence calculations. Allele frequencies of zero can arise when an allele is observed in a sample that was not observed in the allele frequency database. They can also arise within a database when a particular allele is observed in one database sub population (e.g., Caucasian), but not in one or more of the other sub populations (African American, Asian, Hispanic, etc.). The frequency tables for these other sub populations have a count of zero for this allele.

The problem of zero counts was addressed by the National Research Council by replacing the zero count with a value of 5/2N. Typically, the value is applied to all allele counts ≤5 (**Equation *1***).

Equation 1

$$f_a = \begin{cases} x \leq 5: \dfrac{5}{2N} \\ x > 5: \dfrac{x}{2N} \end{cases}$$

where:

- f_a is the frequency assigned to an allele
- x is the allele count in the database
- N is the number of humans in the database sample

This approach has fallen out of favor, and has been replaced by methods more suited to probabilistic genotyping that adjust all frequencies in the database (4) using equations such as **Equation 2**.

Equation 2

$$f_{x_i} = \frac{x_i + \frac{1}{k+2}}{N+1}$$

Where:

x_i = count of allele ***i*** in the database (may be zero)

k = number of allele types at the locus

N = total number of alleles (i.e., chromosomes) in the database at the locus

One criticism of the NRC method is that it increases the frequencies of low-count alleles causing the total allele frequency at a locus to exceed 1.0, meaning that it is no longer a proper probability distribution. The Curran method does not have this side effect because when allele counts are adjusted for database size, some allele counts are increased while others are decreased (**Table *3***). It should be noted that sequence-based alleles experience lower and zero-frequency alleles than length-based alleles.

Table 3. Sequence-based allele frequencies at D1S1656 adjusted for database size by the NRC (5) and Curran and Buckleton (4) methods. The NRC method increases low-frequencies (green) whereas the Curran method both increases and decreases (red) frequencies relative to raw count frequencies. The NRC method generates a more conservative novel allele frequency (zero count alleles).

Locus	Allele	Count in Database	Raw Frequency	Adjusted from Raw Frequency			
				NRC	Delta	Curran	Delta
D1S1656	14_TE	1	0.00048	0.00241	0.00193	0.00050	0.00001
D1S1656	15_KH	1	0.00048	0.00241	0.00193	0.00050	0.00001
D1S1656	17_FJ	1	0.00048	0.00241	0.00193	0.00050	0.00001
D1S1656	17.3_QY	1	0.00048	0.00241	0.00193	0.00050	0.00001
D1S1656	17.3_AM	1	0.00048	0.00241	0.00193	0.00050	0.00001
D1S1656	10_US	2	0.00097	0.00241	0.00145	0.00098	0.00001
D1S1656	16.3_CZ	3	0.00145	0.00241	0.00097	0.00146	0.00001
D1S1656	14.3_VB	4	0.00193	0.00241	0.00048	0.00194	0.00001
D1S1656	14.3_SI	5	0.00241	0.00241	0.00000	0.00243	0.00001
D1S1656	17_TS	5	0.00241	0.00241	0.00000	0.00243	0.00001
D1S1656	11_JP	6	0.00290	0.00290	0.00000	0.00291	0.00001
D1S1656	13_IG	9	0.00434	0.00434	0.00000	0.00436	0.00001
D1S1656	16_SK	9	0.00434	0.00434	0.00000	0.00436	0.00001
D1S1656	16_TR	12	0.00579	0.00579	0.00000	0.00580	0.00001
D1S1656	18_UJ	12	0.00579	0.00579	0.00000	0.00580	0.00001
D1S1656	10_LV	13	0.00627	0.00627	0.00000	0.00628	0.00001
D1S1656	19.3_LQ	19	0.00917	0.00917	0.00000	0.00918	0.00001
D1S1656	15.3_KA	28	0.01351	0.01351	0.00000	0.01352	0.00001
D1S1656	15_IP	43	0.02075	0.02075	0.00000	0.02076	0.00000
D1S1656	13_JR	58	0.02799	0.02799	0.00000	0.02799	0.00000
D1S1656	15.3_IQ	58	0.02799	0.02799	0.00000	0.02799	0.00000
D1S1656	14_BN	61	0.02944	0.02944	0.00000	0.02944	0.00000
D1S1656	18.3_SV	74	0.03571	0.03571	0.00000	0.03571	0.00000
D1S1656	12_EY	81	0.03909	0.03909	0.00000	0.03909	-0.00001
D1S1656	17_EO	81	0.03909	0.03909	0.00000	0.03909	-0.00001
D1S1656	12_GH	98	0.04730	0.04730	0.00000	0.04729	-0.00001
D1S1656	11_MB	100	0.04826	0.04826	0.00000	0.04825	-0.00001
D1S1656	13_NR	130	0.06274	0.06274	0.00000	0.06272	-0.00002
D1S1656	16.3_FB	138	0.06660	0.06660	0.00000	0.06658	-0.00002
D1S1656	17.3_TS	215	0.10376	0.10376	0.00000	0.10373	-0.00004
D1S1656	14_SD	238	0.11486	0.11486	0.00000	0.11482	-0.00004
D1S1656	16_LG	274	0.13224	0.13224	0.00000	0.13219	-0.00005
D1S1656	15_MQ	291	0.14044	0.14044	0.00000	0.14039	-0.00005
	Sum	**2072**	**1.00000**	**1.01255**	**0.01255**	**0.99997**	**-0.00003**
	Novel Allele Frequency		**2.4×10^{-3}**			**1.38×10^{-5}**	

MixtureAce contains pre-calculated allele frequency tables for five different trim protocols for each forensic kit (**Table *4***). Users can select the trim protocol to use in the data entry popup window (**Figure *6***).

Table 4. Description of bioinformatic trimming protocols.

Protocol Shorthand	Alternative Name	Description
A	STR Only	Flanking sequences on both sides of the STR are trimmed.
B	Minus Primers	Primer sequences, when known, are trimmed.
C	Vendor Recommendation	Trim where kit vendor recommends, as identified in the forensic sequence structure guide (https://strider.online/).
D	Database Intersection	Trim at the intersection of the kit amplicon and the database genomic extent, thereby maximizing the number of database-derived frequencies (6).
E	ISFG Minimum Range	Trim at the minimum range as identified by ISFG in the forensic sequence structure guide (https://strider.online/).

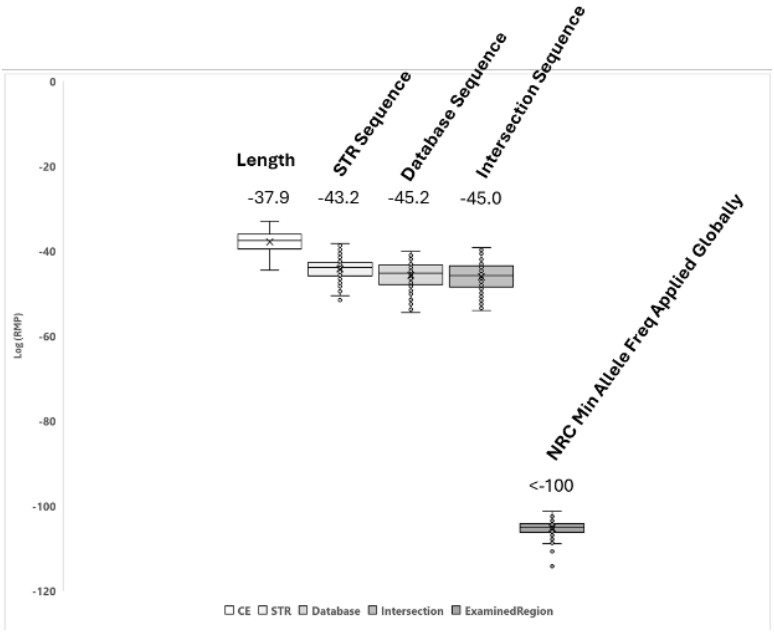

Figure 3. Negative logarithms of random match probabilities (RMP) for 1000 randomly generated profiles of 28 autosomal STR markers in the OmniSTR kit. Plots show distributions of RMPs for five different treatments of the same data.

How MixtureAce Manages Allele Frequencies

Allele frequences can be uploaded to MixtureAce from Excel spreadsheets via the Options menu (**Figure *4***). The file format must include columns for Race (Population), Locus, Allele, and Frequency. A source for the frequency data must also be provided.

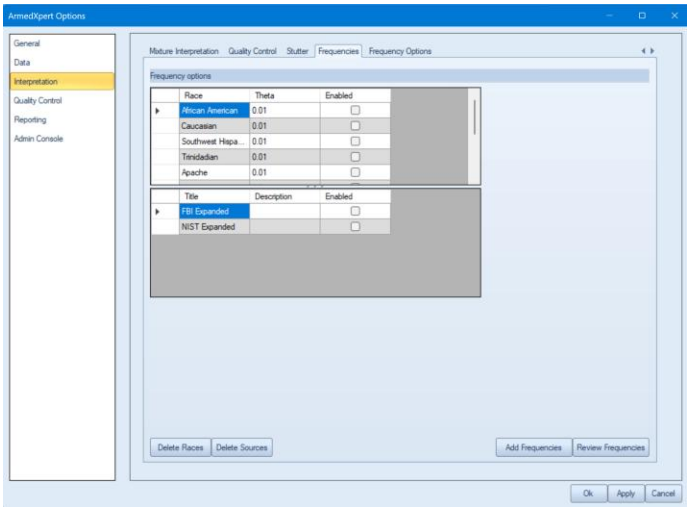

Figure 4. Allele frequency input window.

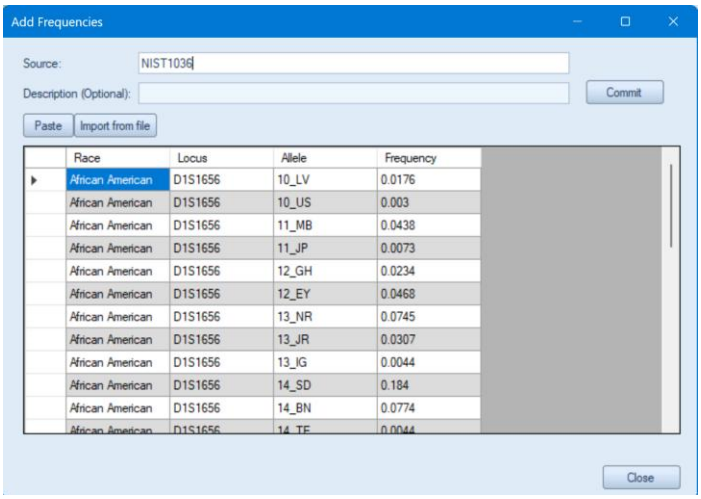

Figure 5. Adding frequencies from Excel spreadsheet by importing from a file or copy and paste.

Figure *3* illustrates the impact of trimming protocols on profile probabilities. Probabilities based on the STR only sequence (protocol A) can be several orders of magnitude lower than probabilities based on

allele number. Probabilities based on the intersection sequences (trim protocol D) are very similar to probabilities based on the database sequences. Using untrimmed sequences and applying the NRC minimum allele frequency to amplicons that do not match the database extent results in extremely low and unrealistic probabilities. This approach, which can be used for length-based alleles, is ***not recommended*** for sequence-based alleles due to the resulting extreme RMPs.

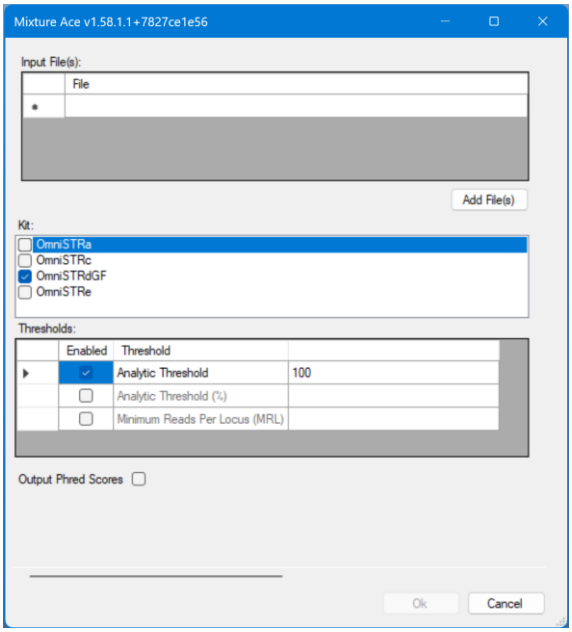

Figure 6. MixtureAce input screen for the OmniSTR kit specifying 4 of the 5 bioinformatic trimming options (trimming primers and at kit vendor recommended positions are the same).

Literature Cited

1. Borsuk LA, Gettings KB, Steffen CR, Kiesler KM, Vallone PM. Sequence-based US population data for the SE33 locus. Electrophoresis. 2018;39(21):2694-701.
2. Gettings KB, Borsuk LA, Steffen CR, Kiesler KM, Vallone PM. Sequence-based U.S. population data for 27 autosomal STR loci. Forensic Sci Int Genet. 2018;37:106-15.
3. Steffen CR, Coble MD, Gettings KB, Vallone PM. Corrigendum to 'U.S. Population Data for 29 Autosomal STR Loci' [Forensic Sci. Int. Genet. 7 (2013) e82-e83]. Forensic Sci Int Genet. 2017;31:e36-e40.
4. Curran JM, Buckleton JS. An investigation into the performance of methods for adjusting for sampling uncertainty in DNA likelihood ratio calculations. Forensic Sci Int Genet. 2011;5(5):512-6.
5. National Research Council (U.S.). Committee on DNA Forensic Science: an Update., National Research Council (U.S.). Commission on DNA Forensic Science: an Update. The evaluation of forensic DNA evidence. Washington, D.C.: National Academy Press; 1996. xv, 254 p. p.
6. Young B, Marciano M, Crenshaw K, Duncan G, Armogida L, McCord B. Match statistics for sequence-based alleles in profiles from forensic PCR-mps kits. Electrophoresis. 2021;42(6):756-65.