Sequence Data Generation

The goal of this experiment was to report complete profiles (no dropout as compared to CE genotypes with at least 30X coverage per allele in the Illumina ForenSeq Universal Analysis Software (UAS)) for all 27 autosomal STR loci reported in ForenSeq DNA Signature Prep. Optimization experiments were conducted and deviations from the manufacturer protocol (ForenSeq DNA Signature Prep Reference Guide, Document # 15049528 v01, September 2015) are outlined below.

For N = 658 population samples derived from blood, 5 µL input DNA of approximately 0.5 ng/µL (2.5 ng total input) yielded more complete profiles compared to the manufacturer recommended input of 1 ng; therefore, this was the input volume for these samples.

For population samples derived from buccal swabs (N = 378), previous testing with other assays showed possible decreased quality and quantity from the expected 0.5 ng/ µL; therefore, a subset (N = 64) were re-quantified with Quantifiler HP in preparation for sequencing. Average concentrations of 0.35 ng/µL and 0.12 ng/µL were obtained for the small and large targets, respectively. To increase the input DNA, these samples (N = 384) were processed by completely drying down 10 µL of DNA extract and adding 15 µL mastermix, replacing DNA input volume with water (5 µL primer, 4.7 µL reaction mix, 0.3 µL enzyme, 5 µL H_20). All samples were initially amplified using DNA Primer Mix B (DPMB), while a subset of reruns (N = 13) were processed with DNA Primer Mix A (DPMA).

Manufacturer's protocol states that purified libraries can be stored at 4 ˚C for up to one year and normalized libraries can be stored at 4 ˚C for up to 30 days. In our experiments, the most complete profiles were obtained when purified library storage time was minimized and normalized libraries were sequenced immediately.

During optimization experiments, library pool input was adjusted upward from the manufacturer recommended 7 µL. The goal of this change was to increase the cluster density while keeping the clusters passing filter within passing range. For samples derived from blood, the typical pool input was 10 µL and for samples derived from buccal swabs, the typical pool input was 12 µL, with a maximum input of 14 µL. The average cluster density from these 42 sequencing runs was 1 400 K/mm^2, and ranged from 702 K/mm^2 to 1853 K/mm^2 (manufacturer's recommended range is 400 to 1650 K/mm^2). All sequencing runs with pool input of 10 µL were within the manufacturer's recommended cluster density range, while nearly half the runs with pool input of 12 µL exceeded the recommended range (these results may vary by sample quantity/quality and instrument). The average clusters passing filter was 86.7 %, ranging from 74.2 % to 96.2 % (manufacturer's recommended level is ≥ 80 %). Full details on each of the 42 runs can be found in Supplementary Table 1. Figure A below compares the cluster density to the percentage of clusters passing filter, to explore how these metrics are related. Generally, cluster densities exceeding the recommended range result in lower percentages of clusters passing filter, but the effect appears gradual. As shown in Figure B, phasing and pre-phasing are all within manufacturer's recommended range; the average phasing and pre-phasing was 0.186 % and 0.096 %, respectively. Phasing and pre-phasing do not appear to be affected by cluster density, as shown by comparison of Figure A to Figure B.

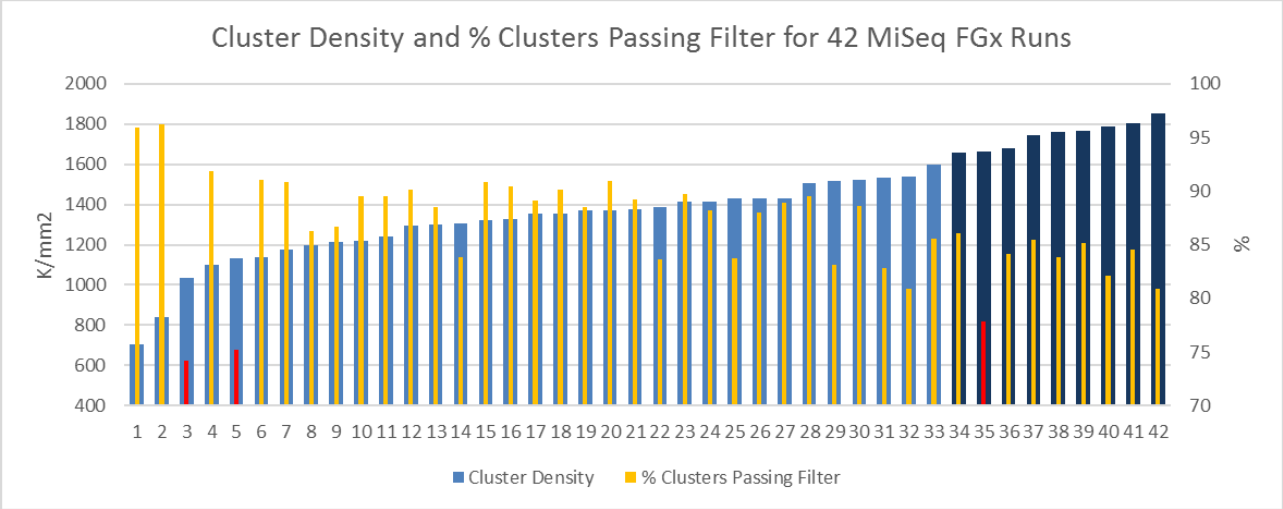

Figure A. Left side y-axis is cluster density in K/mm^2 and right side y-axis is percent clusters passing filter. Light blue bars represent cluster density within the manufacturer's recommended range of 400 to 1650 K/mm^2. Dark blue bars exceed the recommended range. Yellow bars represent percentage clusters passing filter above the recommended 80 %, while red bars are below 80 %. Runs are sorted left to right by increasing cluster density. Runs 3 and 5 were chronologically sequential, and subsequent troubleshooting diagnosed a camera focusing issue.

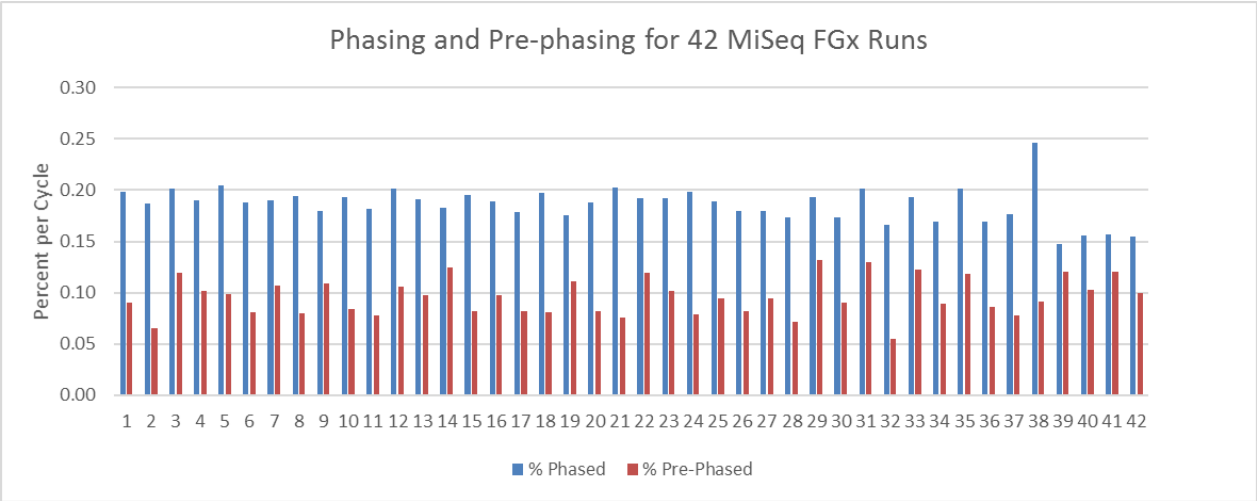

Figure B. Percent phasing and pre-phasing per cycle. Runs are sorted left to right by increasing cluster density, the same order as Figure A.

Thirteen samples requiring reruns because of low coverage or dropout (as determined by comparison to CE typing results) were processed and sequenced in one sequencing run of N=15 plus an amplification negative. Additional steps to improve coverage included the previously mentioned use of DPMA, and elution of the purified product in 22.5 µL instead of the recommended 52.5 µL for N=4 low coverage samples. This last change allowed nearly the entire purified product (20 µL) to be carried forward to normalization/pooling/sequencing. For one sample processed first in the typical manner and rerun with these changes, a comparison of coverage across the STR loci is presented in Figure C. It is important to note that these changes were undertaken for single source samples with full CE concordance data available; internal validation would be required to determine the effects of such changes prior to casework use.

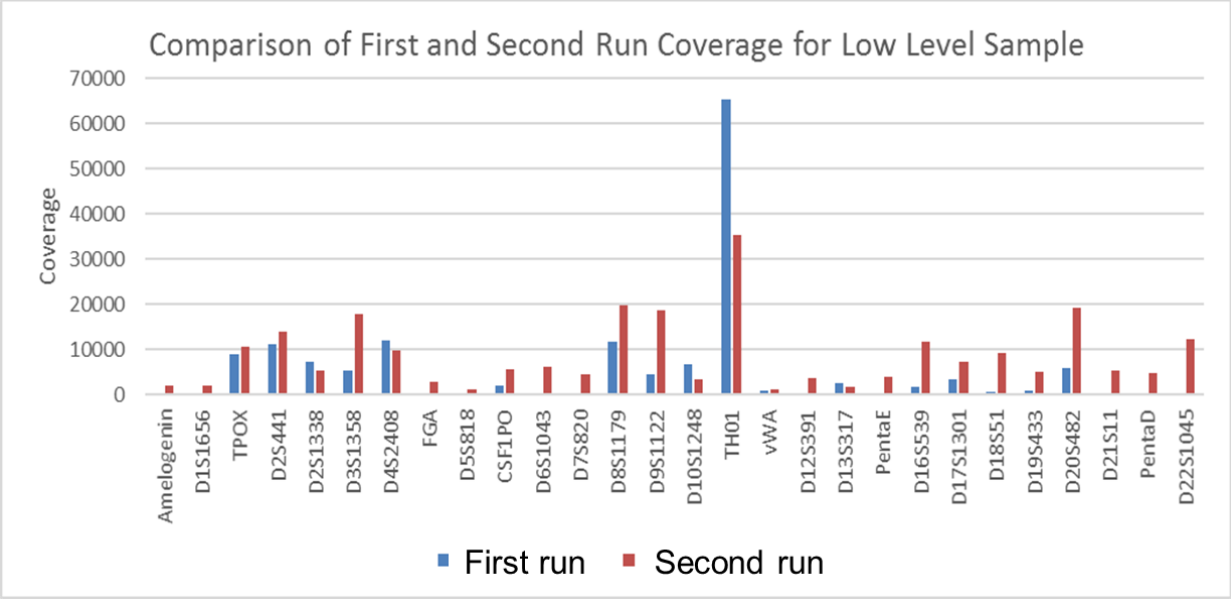

Figure C. Improved results were obtained for a low-level sample. The changes made for the second run are 1) Amplification with DPMA rather than DPMB, 2) Elution of purified product in a reduced volume (22.5 µL compared to 52.5 µL), 3) Decreased number of samples in sequencing run (15 samples compared to 24).

CE Data Generation

Length-based autosomal STR genotype results for the NIST 1036 samples across 24 CE multiplexes have been previously published [1, 2]. This data set served as the basis for CE concordance checks for 23 of the 27 autosomal STR loci reported by the UAS (D1S1656, D2S441, D2S1338, D3S1358, D5S818, D6S1043, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, CSF1PO, FGA, Penta D, Penta E, TH01, TPOX, and vWA). CE genotype data was also previously published for N = 665 of the NIST 1036 samples at the four remaining autosomal STR loci reported by the UAS ("NIST mini-STRs" D4S2408, D9S1122, D17S1301, and D20S482) [3]. The remaining 371 samples (1036 – 665) were CE genotyped at the four NIST mini-STRs for this project, using the parameters described in [3].

Data Analysis

Two bioinformatic analyses were performed: 1) the Illumina Universal Analysis Software (UAS)-processed .txt files were parsed off-platform with custom batch-enabled scripts and 2) the FASTQ files were analyzed with a custom pipeline based on STRait Razor 2.0 [4]. For both methods, a minimum allelic coverage threshold of 30X was targeted, and samples with alleles falling below this level were reprocessed. For UAS analysis, default stutter thresholds (see Illumina ForenSeq Universal Analysis Software Guide, Document # 15053876 v01) and 4.5 % allele calling threshold (when > 650 reads were present at a locus) were used. In parsing SR analysis results, an allelic balance threshold of 20 % was implemented to reduce artifacts requiring manual inspection. Both analyses' results were compared to the expected CE-based allele calls, and any differences investigated/ arbitrated.

Based on the results of the arbitration and a general evaluation of sequencing errors outside of the UAS reported region, a reportable range for SR was determined, as shown in **Supplementary Table 4**. Sequence string counts were converted to frequencies in Excel, and both are presented in **Supplementary Table 3**, along with condensed and full sequence string allele formats. Sequence strings were given unique designators and analyzed using STRAF [5] and Arlequin [6]. This analysis included calculation of allele frequencies (confirmed with Excel results), observed and expected heterozygosity, probability of identity, evaluation of conformity to Hardy Weinberg Equilibrium, and evaluation of linkage disequilibrium. Sequences mapping to the SE33 locus were present in SR results but were not accessible in the UAS. Analysis of the SE33 locus results is published elsewhere [7].

Sequence strings were condensed/formatted according to guidance from the International Society of Forensic Genetics [8, 9], including reporting sequences on the forward strand. The following loci are reported on the reverse strand in the UAS and were converted to forward strand for this publication: D1S1656, D2S1338, FGA, D5S818, CSF1PO, D6S1043, D7S820, vWA, Penta E, and D19S433.

Sequencing Results

Two lots of library preparation reagents were used in this study, with markedly different DPMB primer balances at some loci. Figure D below contains the average coverage across 592 samples processed with the first lot, compared with the average coverage across 329 samples processed with the second lot. Specifically, the autosomal STR D9S1122 routinely exhibited low coverage with instances of allelic dropout in the first lot, then contained comparatively moderate to high coverage in the second lot.

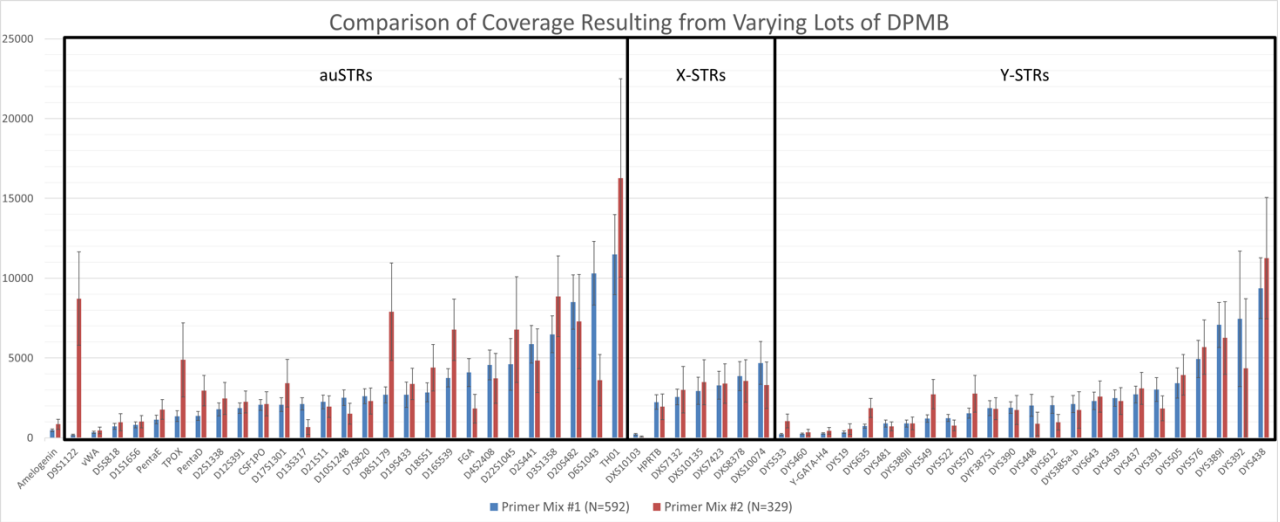

Figure D. Average coverage and standard deviation across two lots of primer mix. Primer Mix #1 data includes 19 sequencing runs encompassing 592 samples and Primer Mix #2 data includes 14 sequencing runs encompassing 329 samples. Nine sequencing runs from the population sample sequencing were not considered for this analysis of coverage: two were excluded due to suboptimal instrument performance and seven were excluded because of suboptimal run conditions (test runs or reruns of failed samples). Also excluded from this analysis were four female samples (all other samples are male) and 10 samples with very low coverage.

Although the goal was to obtain 30X coverage for each reported sequence, this was not possible for two samples, shown in Table A, below. One Penta E [TCTTT]18 allele which was expected based on CE data was not detected above 10X in the first run. After reamplification (at which point the sample was depleted), the sequence was present at 28X in UAS and 27X in SR. This is the standard motif at Penta E, and no flank is being reported for this locus; therefore, this sequence was reported/included in the frequency data. Additionally, one D22S1045 [ATT]16 ACT [ATT]2 allele was present at 29X in the SR analysis and 32X in the UAS analysis, and was also reported/included in frequency data.

Population, Sample Type	Locus	Genotype	UAS	SR
Caucasian, buccal	Penta E	10,18	18 = 28X	18 = 27X
African American, blood	D22S1045	14,19	19 = 32X	19 = 29X

Table A. Samples reported with coverage below 30X in N=1036.

D22S1045 generally exhibited allele specific imbalance, which decreased as the difference in allele size increased in a heterozygote pair. Table B, below, contains the most extreme cases which required manual data curation to achieve concordance with CE data. In all cases, the stutter product of the smaller allele exhibited greater coverage than the larger allele, resulting in a higher stutter ratio than the allele coverage ratio (ACR) of the true alleles. A similar trend was noted for DYS392 (not reported in this manuscript); however, as it is a single-copy YSTR locus, this did not result in discord. Rather, this trend was observed as a high standard deviation of locus coverage between samples. Both loci are trinucleotide repeats composed only of adenine and thymine bases.

n-4 Stutter Position of Allele 1	Coverage n-4 Stutter of Allele 1	Allele 1	Coverage 1	Allele 2	Coverage 2	n-4 Stutter Ratio of Allele 1	ACR	Allele Size Delta
							Coverage 2/ Coverage 1	
10	488	11	6621	19	100	7.40%	1.50%	8
10	235	11	4362	19	139	5.40%	3.20%	8
10	553	11	9719	18	31	5.70%	0.30%	7
10	545	11	10363	18	101	5.30%	1.00%	7
10	548	11	10254	17	462	5.30%	4.50%	6
10	399	11	6821	17	254	5.80%	3.70%	6
10	207	11	3679	17	156	5.60%	4.20%	6
10	419	11	8536	16	372	4.90%	4.40%	5
13	255	14	3226	19	32	7.90%	1.00%	5
13	185	14	2073	19	109	8.90%	5.30%	5
15	411	16	1952	17	289	21.10%	14.8%	1

Table B. D22S1045 UAS allele coverage for samples requiring manual data curation of the larger allele in N=1036. n-4 stutter ratio of the smaller allele is greater than the ACR of the true alleles.

Instances of allele dropout or severe imbalance other than the previously described D22S1045 issue are detailed in Table C, below. These results were attempted to be confirmed by reamplification/ resequencing, as shown. Sequence coverage data in this table is from the UAS, and the same genotype with similar coverage levels were observed in SR:

Population, Sample Type	Locus	CE	UAS-1st Run	UAS-2nd Run	Action
African American, Blood	D5S818	PPF6C 11-3038RFU 12-4665RFU	11-258X 12-15X	11-1790X 12-612X	11,12 reported
Hispanic, Blood	TH01	Varies by kit: IDPlex-6 PPF6C-6,7	6-3478X 7-28X 9.3-37X	6-13574X 7-123X 9.3-21X	Promega PowerSeq and Sanger sequence data reported for 7 allele
Asian, Buccal	D19S433	13-2767 RFU 14.2-2523 RFU	13-78X 14.2-1390X	13-198X 13.2-303X 14.2-3645X	13, 14.2 reported
African American, Buccal	D20S482	14-679 RFU 15-587 RFU	14-1301X 15-798X	Rerun failed, sample depleted	14,15 reported
African American, Blood	Amel	Varies by kit: IDPlex-X,Y PPF6C-Y	Y-254X	Y-1147X	Amel frequencies not reported

Table C. Allele dropout or severe imbalance observed in the sequence data at 26 auSTR loci (excluding D22S1045) and Amelogenin (Amel) in N=1036.

Instances of discordance between CE comparison to UAS and/or SR, are detailed in Table D, below. In each case, heterozygote balance and coverage were typical unless otherwise noted.

Population, Sample Type	Locus	CE	UAS	SR	Source
African American, Blood	Penta D	2.2,13.4	2.2,14	**2.2,13.4**	1 bp deletion rs536566765
African American, Buccal	D5S818	**7,12***	8,12	**8,12***	Assumed 4 bp deletion outside ForenSeq amplicon
Hispanic, Blood	D7S820	10.3,11	11,11	**10.3,11**	1 bp deletion, rs897512434
Caucasian, Blood	D9S1122	(12),14	12,14	**11.2,14**	2 bp deletion rs754976988, overlaps with CE primer binding site

Table D. Discordance between CE and sequence data observed in N=1036. Bolded genotypes used in allele frequency calculations. *8,12 used in 1036 sequence-based frequencies, 7,12 used in 1036 CE-based frequencies [2].

This low level of discordance is attributable to multiple factors. First, at some loci, a portion of the flanking region is included in the UAS analysis, which accounts for the most common flanking region indels (e.g. D13S317). Second, the CE dataset used for concordance checking was generated via multiple CE assays with overlapping loci. Previously identified issues of discord between CE assays have been evaluated, sometimes including Sanger Sequencing, to produce a CE-based frequency dataset representative of the individuals' true genotypes. It is possible that a laboratory performing a similar study and using a single CE assay to develop concordance data for each locus will observe a higher level of discordance.

Three triallelic patterns were observed in this study, summarized in Table E, below. A triallelic pattern of 9,10,11 was observed for one African American sample at the TPOX locus; each allele had comparable coverage and identical sequence motif. This pattern at TPOX has been observed in the literature and hypothesized as a duplication of a portion of Chromosome 5 onto the X Chromosome, with the original duplicated TPOX allele being 10 repeats in length [10, 11]. Exploring the flanking regions of these three alleles revealed no differences. A triallelic pattern of 11,(14),15 was observed for one Hispanic sample at the Penta D locus; each allele had identical sequence motif, but imbalance was observed with coverages of 1292X, 458X, and 858X, respectively (and similarly imbalanced CE genotype). A triallelic pattern of 11,11,13 was observed for one Hispanic sample at the D9S1122 locus; the 11 alleles were of two different sequence motifs, resulting in a sequence-based triallele, with each allele having comparable coverage. These three samples were excluded from frequency calculations at their respectively affected loci, resulting in a decrease in N by one for TPOX, Penta D, and D9S1122 (noted and shown in sample counts in **Supplementary Table 3**).

Population, Sample Type	Locus	CE Result	NGS Result	Action
African American, Blood	TPOX	9,10,11	9,10,11	Removed sample from TPOX frequency calculation
Hispanic, Buccal	Penta D	11,(14),15	11,(14),15	Removed sample from Penta D frequency calculation
Hispanic, Buccal	D9S1122	11,13	11,11,13	Removed sample from D9S1122 frequency calculation

Table E. Triallelic genotypes observed in N=1036.

Additional evaluation was performed in an effort to detect erroneous sequences which could not be determined by CE concordance evaluation, namely sequence-based heterozygotes or isoalleles. All isoalleles (947 length-based homozygotes, 1894 sequence-based alleles) were manually reviewed for observation of both alleles in other samples in this N=1036 dataset. Twenty-three sequence-based alleles were not observed in other samples, and were investigated further. Three of these sequence-based alleles at Penta D were determined not true alleles. All three had the same sequencing error in the 3' flank, which was recognized due to imbalance (16%, 20%, and 13% coverage ratios) and all three erroneous alleles being detected on the same sequencing run, which had unusually high phasing of 0.24% (within manufacturer's recommended range but outlying compared to the other 41 sequencing runs, see run #38 in Figure B). One sequence-based allele at D21S11 was determined not a true allele. A sequencing error was recognized in the 3' flank due to imbalance (23% coverage ratio). Also, this sample was sequenced three times and the erroneous allele was only observed in one run. The additional 19 sequence-based alleles demonstrated typical coverage ratios from 0.77 to 0.99 and typical sequence coverage from 261X to 6712X, but were evaluated further as described below.

Based on the detection of the above four erroneous sequences, the 1036 dataset was searched at the 27 autosomal STR loci for any sequences with only one observation. A total of 215 sequences were identified in this search, including the aforementioned 19 which were present in isoalleles. These 215 sequences were manually reviewed for observation of alleles of the same motif in this 1036 dataset. Sixty-one sequences were of a motif not observed in other alleles in this dataset, and were investigated further. The first level of confirmation was to determine if the sequences had been observed in samples outside of the 1036 dataset, either in-house or published. A search of [12] for matching sequences or sequences of the same motif resulted in the confirmation of eight and three sequences, respectively. One sequence was confirmed by matching a Sanger sequence present in SRM 2391c Component E. Seven additional sequences were confirmed by their presence in the father sample of the associated son included in 1036. Forty-two remaining sequences were confirmed by retyping either in Promega PowerSeq (29 samples) or ForenSeq (14 samples). No additional erroneous sequences were detected in this process.

It should be noted that observation of the same sequence or motif in another sample is not conclusive verification (evidenced by the three erroneous alleles at Penta D), nor is the single observation of an allele inherently a cause for concern. The above evaluation was performed to increase confidence in rarely occurring sequences in this dataset. Additionally, the erroneous bases detected at Penta D and D21S11 were in flanking sequences outside of manufacturer software range.

REFERENCES

[1] C.R. Hill, D.L. Duewer, M.C. Kline, M.D. Coble, J.M. Butler, U.S. population data for 29 autosomal STR loci, Forensic Sci Int Genet 7(3) (2013) e82-3.
[2] 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], Forensic Sci Int Genet 31 (2017) e36-e40.
[3] C.R. Hill, M.C. Kline, M.D. Coble, J.M. Butler, Characterization of 26 MiniSTR Loci for Improved Analysis of Degraded DNA Samples, J Forensic Sci 53(1) (2008) 73-80.
[4] D.H. Warshauer, J.L. King, B. Budowle, STRait Razor v2.0: the improved STR Allele Identification Tool--Razor, Forensic Sci Int Genet 14 (2015) 182-6.
[5] A. Gouy, M. Zieger, STRAF-A convenient online tool for STR data evaluation in forensic genetics, Forensic Sci Int Genet 30 (2017) 148-151.
[6] L. Excoffier, H.E.L. Lischer, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, Molecular Ecology Resources 10(3) (2010) 564-567.
[7] L. Borsuk, K.B. Gettings, C.R. Steffen, K.M. kiesler, P.M. Vallone, Sequence-based U.S. population data for the SE33 locus Electrophoresis [accepted manuscript] (2018).
[8] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, L. Gusmao, D.R. Hares, J.A. Irwin, J.L. King, P. Knijff, N. Morling, M. Prinz, P.M. Schneider, C.V. 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.
[9] 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.
[10] A.B. Lane, The nature of tri-allelic TPOX genotypes in African populations, Forensic Science International: Genetics 2 (2008) 134-137.
[11] J.B. Picanco, P.E. Raimann, G.A. Paskulin, L. Alvarez, A. Amorim, S.E. Batista Dos Santos, C.S. Alho, Tri-allelic pattern at the TPOX locus: a familial study, Gene 535(2) (2014) 353-8.
[12] N.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.