Contents lists available at ScienceDirect

Forensic Science International: Reports

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

Genetic Population Data

Sequence-based U.S. population data for 7 X-STR loci

Lisa A. Borsuk *, Carolyn R. Steffen, Kevin M. Kiesler, Peter M. Vallone, Katherine B. Gettings

U.S. National Institute of Standards and Technology, Biomolecular Measurement Division, 100 Bureau Drive, Gaithersburg, MD, 20899, USA

ARTICLE INFO

Keywords:
X chromosome
Short tandem repeat (STR)
Sequence
Population data

ABSTRACT

The National Institute of Standards and Technology (NIST) U.S. population sample set of unrelated individuals was used to determine allele and haplotype frequencies for seven X-chromosome short tandem repeat (STR) loci in four linkage groups. DXS7132, DXS7423, DXS8378, DXS10074, DXS10103, DXS10135, and HPRTB were sequenced using the ForenSeq DNA Signature Prep Kit on a MiSeq FGx instrument from Verogen. Capillary electrophoresis data produced using the Qiagen Investigator Argus X-12 was compared to ForenSeq length-based alleles and found to be 99 % concordant. For three loci (DXS10103, DXS10074, and HPRTB) the length-based allele call is affected by the extent of flanking region included in the reported sequence. Six of the seven loci gained alleles by sequencing compared to length-based determinations. The increase in alleles are found in both the repeat and flanking region sequences. All sequences for which frequencies are reported in this dataset were cataloged as GenBank records in the STRSeq NCBI BioProject (https://www.ncbi.nlm.nih.gov/bioproject/380127). Frequency information for both the loci and linkage groups is reported, along with results of statistical tests including gene diversity, polymorphism information content, power of discrimination, and linkage disequilibrium. All supplemental files are available at the NIST Public Data Repository – U.S. population data for Human Identification Markers (https://doi.org/10.18434/t4/1500024).

1. Introduction

X-chromosome short tandem repeat (X-STR) markers are recognized as useful tools to supplement kinship testing in the forensic setting. Studies of allele and haplotype frequencies based on traditional capillary electrophoresis (CE) length-based analyses of these loci have been reported in the literature for various population groups, for examples see [1–4]. More recently, new technologies capable of providing sequenced-based information with a higher level of marker multiplexing have been investigated for characterization of forensic loci, including X-STRs [5–10].

The National Institute of Standards and Technology (NIST) U.S. Population Sample Set, consisting of 1036 unrelated individuals were sequenced using the MiSeq FGx Forensic Genomics System (Verogen, San Diego, CA), including the ForenSeq DNA Signature Prep Kit (Verogen), which targets STR markers commonly used for human identification and relationship testing [11]. Seven X-STR loci are reported in this assay: DXS10135, DXS10074, DXS7132, DXS10103, DXS7423, DXS8378, and HPRTB [12], with at least one marker representing each of the four linkage groups established on the X-chromosome [13]. The core repeat region was assessed with the ForenSeq Universal Analysis Software v1 (UAS) whereas the entire reported region variation was assessed with a customized bioinformatic approach.

Sequence-based allele and haplotype frequencies along with other relevant population genetic parameters for each population group were determined. The information provided in this study will serve to facilitate the application of sequence-based methods to X-STR profiling in the forensic setting. The sequence data are available as NCBI GenBank records within the STRSeq X-Chromosomal STR Loci BioProject, accession PRJNA380348 [14]. All supplemental files for this paper are available at the NIST Public Data Repository – U.S. population data for Human Identification Markers (https://doi.org/10.18434/t4/1500024).

2. Materials and methods

2.1. Population and samples

A total of 1036 anonymous samples with self-reported ancestries were sequenced in what has been previously reported as the “NIST 1036” [11], divided among four U.S. populations: African American (AfAm, N = 342), Asian (N = 97), Caucasian (Cauc, N = 361), and Hispanic (Hisp, N = 236).

Abbreviations: X-STR, X-chromosome short tandem repeat; CE, capillary electrophoresis; UAS, universal analysis software; SLR, sample level report; FRR, flanking region report; DoC, depth of coverage; ACR, allele coverage ratio.

* Corresponding author.
E-mail addresses: lisa.borsuk@nist.gov (L.A. Borsuk), becky.steffen@nist.gov (C.R. Steffen), kevin.kiesler@nist.gov (K.M. Kiesler), peter.vallone@nist.gov (P.M. Vallone), katherine.gettings@nist.gov (K.B. Gettings).

http://doi.org/10.1016/j.fsir.2020.100160
Received 5 October 2020; Received in revised form 17 November 2020; Accepted 18 November 2020
Available online 23 November 2020

Five samples were removed from NIST 1036 for this analysis of X-STR sequence data for a total of 1031 samples in this study. Four female samples were removed: one African American, one Asian, and two Caucasian. Of the remaining 1032 male samples, one African American male sample was removed due to the appearance of a high duplication rate on the X chromosome (five of the seven ForenSeq X-STR loci appeared to be duplicated in both ForenSeq and Investigator Argus X-12).

All work presented in this paper has been reviewed and approved by the NIST Research Protections Office.

2.2. Genotyping

2.2.1. Investigator Argus X-12

CE length-based genotypes were previously generated for a subset of these samples (n = 663) using the Investigator Argus X-12 kit (Qiagen, Hilden, Germany) [4], and the remaining samples were analyzed with Investigator Argus X-12 in the course of this study. The samples were run following the manufacturer's protocol for Investigator Argus X-12. The samples were analyzed on the 16-capillary Applied Biosystems Prism 3130xl Genetic Analyzer and the data was analyzed using GeneMapper*ID* v3.2 (Thermo Fisher).

2.2.2. MiSeq FGx forensic genomics system

The samples were sequenced on the MiSeq FGx Forensic Genomics System using the ForenSeq DNA Signature Prep Kit as previously described by Gettings et al. [11] and associated supplemental material (https://doi.org/10.18434/t4/1500024).

2.3. Sanger sequencing and additional CE analysis

Sanger sequencing of the DXS10135 locus was performed to further complement the dataset using an in-house designed assay. The primers used for amplification and sequencing the DXS10135 locus were: Forward: 5′-ACTCCCGACCTCAGGTGAT-3′ and Reverse: 5′-TATGGAGC-TAAGGGGTGACA-3′ (Eurofins Genomics, Louisville, KY). A PCR annealing temperature of 63 °C was used and was the only change made to the previously published Sanger sequencing protocol [15].

For one sample, an in-house CE analysis was performed for the DXS10103 locus to evaluate a possible duplication. The primer sequences used to amplify the DXS10103 locus were dye-labeled Forward: 5′-6FAM-TACTGCAGGCCTTCCTGAAT-3′ (Thermo Fisher) and unlabeled Reverse: 5′-ATTTTTGACAGGGTGCCAAG-3′ (Eurofins Genomics), while PCR conditions were consistent with a previously published protocol [16].

2.4. Data analysis

2.4.1. ForenSeq UAS analysis

After each sequencing run was completed the UAS automatically analyzed the run, generating a text results file for each sample. The default settings were used for the analysis [17]. The CE length-equivalent information was extracted from the text results files and used to evaluate concordance with CE measurements.

2.4.2. STRait Razor analysis

The raw FASTQ files were analyzed using STRait Razor v3.0 [18] with a modified configuration file (Supplemental File 1, ForenSeq_X-STR.config). The resulting sample files were processed to identify the length-based and sequenced-based allele calls using an in-house Perl script. A set of allele calling rules were established: 1) A depth of coverage (DoC) greater than or equal to 10 was required of a sequence to be considered further, 2) The sequence within a locus with the highest DoC was reported as an allele, 3) Additional sequences were reported as potential duplicated alleles if the allele coverage ratio (ACR is defined as the coverage of sequence under consideration divided by coverage of highest DoC sequence) was above 20 %, 4) Samples with multiple potential alleles at a locus were evaluated individually, as males are expected to have only one X-STR allele for these seven loci. These additional potential alleles were evaluated to determine if they may be high stutter. Additionally, sequences with the same length as the highest DoC sequence but with a single base difference (as sequenced by ForenSeq) were considered noise.

2.4.3. Concordance data sets

Two X-STR sequence data sets were created for CE concordance check purposes: 1) UAS data equivalent to the UAS Sample Level Report (SLR) and 2) STRait Razor analyzed data equivalent to the UAS Flanking Region Report (FRR). See Table 1 for the differences between the two UAS report ranges. The sequence information is provided in Supplemental File 2, "UAS Reported Sequences" and a diagram of the data processing is in worksheet "Data Processing Workflow".

2.4.4. Statistical analysis

Frequencies for both length-based and sequence-based (FRR range – see Table 1) alleles were calculated in Excel and StatsX v2.0, which was also used to calculate linkage group frequencies, gene diversity (GD), polymorphism information content (PIC), power of discrimination for male (PD_M) and female (PD_F), and mean exclusion change (MEC) calculations. StatsX calculates several versions of MEC: Kruger, Kishida, Desmarais, and Desmarais duo [19]. Additionally, StatsX generated an Arlequin formatted file, and Arlequin v3.5.2.2 was used for linkage disequilibrium analysis [20]. The linkage disequilibrium analysis was run with one million steps in Markov Chain instead of the default 10,000 in order to return a p-value for all of the pairs compared. All other parameters were left as default for Arlequin analysis.

3. Results

3.1. Quality

3.1.1. Sequence

Length-based and sequence-based alleles were reported for all 1031 samples for the seven X-STR loci. One null allele was reported for one sample at the DXS8378 locus where both CE and sequencing failed to return a result. For one sample, CE typing was unsuccessful at several loci, likely due to low sample quality, but length-based data was successfully determined from the sequence data. Two loci presented challenges when generating both data sets, DXS10135 and DXS10103, described below.

Table 1
The UAS report ranges for the seven ForenSeq X-STR loci including the SLR and FRR. Base pairs (bp) additional to the repeat region are indicated for the 5′ and 3′ flanks.

	UAS Report Ranges					
	Sample Level Report (SLR)			Flanking Region Report (FRR)		
Locus	5′ bp		3′ bp	5′ bp		3′ bp
DXS10135[a]	0	Repeat	49	0	Repeat	120
DXS8378	0	Region	0	52	Region	72
DXS7132	0		0	24		76
DXS10074	0		0	45		64
DXS10103	0		0	52[b]		2
HPRTB	0		0	80		28
DXS7423	0		0	60		17

[a] The length of 3′ sequence for which frequency data is reported for this locus in this manuscript is the UAS SLR range, which includes 49 bp of 3′ flanking sequence. The additional 71 bp of 3′ flanking sequence included in the UAS FRR was not of sufficient quality across all 1031 samples to be reported.

[b] The UAS FRR has been observed to exclude the first two bases (GC) of this 5′ flanking sequence.

3.1.1.1. Additional review of ForenSeq UAS data. A preliminary comparison of length-based allele calls obtained from CE experiments with those from ForenSeq UAS SLRs was performed. For six of the seven loci, the UAS SLR contains the repeat region only. However, in the case of DXS10135 an additional 49 bases of 3′ flanking region are included in the UAS SLR; these additional bases capture a known three base pair deletion (rs201630737) and any other insertions or deletions in this area, see Supplemental File 2, “UAS Reported Sequences”.

High n-1 stutter was removed from analysis for 10 samples for the DXS10135 locus. The proportion of this high stutter was 22% to 42% of the called allele. One high n-2 stutter (11 %) was removed from a sample at DXS10135 for which a high n-1 stutter (38 %) was also removed. One sample reported an n + 1 allele for DXS7132, which was removed as forward stutter with an ACR less than eight percent. For DXS10103, 19 samples were below the default UAS reporting threshold of 30 × . These sequences and associated length-based alleles were recovered from the UAS raw data with DoC ranging from 14× to 29 × . These length-based alleles were concordant with CE data.

3.1.1.2. Additional review of STRait Razor data. The length-based alleles used for a second CE concordance check were obtained from the STRait Razor analysis and are equivalent to the UAS FRR sequence range in six of the seven loci, see Table 1.

The exception is DXS10135, for which the reported sequence contains 71 bp less 3′ flanking region than the UAS FRR, due to quality issues in the unreported region in this data set. The reported sequence range for DXS10135 is consistent with the sequence range in the UAS SLR. Using the UAS FRR range resulted in 37 allele dropouts (DoC below 10×) and 98 stutter/noise sequences reported above the 20 % ACR cutoff; however, using the UAS SLR range, this was reduced to three allele dropouts and 23 stutter/noise sequences reported above 20 % ACR which required manual review. The sequences reported for three remaining DXS10135 allele dropouts (DoC below 10× in ForenSeq) were determined by Sanger sequencing.

Additional samples evaluated for stutter/noise were resolved to one allele. The highest n-1 stutter artifact removed was 54 % ACR. The highest noise sequence removed was 22 % ACR. DXS10103 exhibited high stutter/noise when using the UAS FRR sequence. No alleles were missing but 369 stutter/noise sequences were present above the 20 % ACR cutoff. Ninety-nine percent (364) of these were dinucleotide stutter from a dinucleotide repeat in the 5′ flanking region. True allelic variation of an additional dinucleotide was also observed in this region in three samples, resulting in x.2 microvariant alleles by length. Based on the size of the Argus X-12 amplicons (111bp to 135bp ladder size range), it is assumed this dinucleotide repeat is not contained within the reported range of the Argus X-12 CE kit. Therefore, alleles with true variation in this region would appear discordant between CE and the UAS FRR; however, would appear concordant between CE and the UAS SLR. In addition, for DXS10103, using the UAS flanking region range resulted in DoC between 10× and 30× for 40 samples. None of these sequences were unique and all were concordant with the CE data; therefore, these sequences were reported for the UAS flanking region range.

3.1.2. CE

Length-based data was obtained for 1030 samples using the Investigator Argus X-12 kit as described in Materials and Methods. One sample repeatedly resulted in poor quality electropherograms. This sample was not included in the CE set for concordance evaluation.

3.2. Concordance

As described above, two concordance checks were performed: 1) ForenSeq UAS length-based alleles from the SLR compared to Investigator Argus X-12 (referred to as UAS SLR vs CE) and 2) STRait Razor analysis length-based alleles determined from sequence consistent with the range of the UAS FRR compared to Investigator Argus X-12 (referred to as UAS FRR vs CE). See Supplemental File 2, “UAS Reported Sequences” for range of sequences compared. All referenced SNPs, insertions and deletions are reported in this worksheet. A table of discordances is reported in worksheet “Discordance Table”. These discordances will be discussed below.

Five discordant allele calls were found between both types of sequencing analysis and CE. Three samples exhibited allele drop-out in sequencing results of larger alleles at DXS10135, likely to be a combination of sample quality and allele length: 32.1, 37.2, and 39.2. Sanger sequencing subsequently confirmed these were concordant with Investigator Argus X-12 allele calls, and there was no evidence of sequence variants in the flanking regions. For one sample at the DXS10103 locus, sequence data reported two alleles (16,19; DoC 112, 373; ACR 30 %) whereas Investigator Argus X-12 reported only the 19 allele (3,828 RFUs). CE analysis was repeated using in-house primers, and again, the 19 allele (25,355 RFUs) was present whereas the 16 allele was not observed. This sample/locus is reported as a 19 allele in this dataset. Lastly, one sample at the DXS10074 locus contained a 7 allele whereas no allelic signal was observed on the CE electropherogram. The allele calls for the remainder of the profile were present and of good quality; therefore, this was considered a CE null allele. In the next two sections, discordances found in only one of the two concordance checks are described.

3.2.1. UAS SLR vs CE

The concordance between the UAS SLR and Investigator Argus X-12 is over 99 %. Two samples were discordant at HPRTB due to a four-base deletion (rs768895696) in the 5′ flank that is not included in the UAS SLR but is included in the CE data. This resulted in a difference of an additional repeat in the sequence data. Likewise, two samples were discordant at DXS10074 due to a two-base insertion (rs1346168465) in a dinucleotide repeat in the 5′ flank of the sequence that is not included in the UAS SLR but is included in the reporting of the CE data.

For DXS7132, three alleles present in CE data were missing from the UAS data. These three DXS7132 alleles have a SNP in common (rs778986795), which has been previously reported to result in a bioinformatic null allele in the UAS [21]. This SNP is adjacent to the repeat region in the 3′ flank. Depth of coverage in the STRait Razor analysis for these three sequences is consistent with average locus coverage.

DXS8378 and DXS10074 each have an additional allele called that is not in the n-1 stutter position (DXS8378: 10, 12; DoC 145, 1958; ACR 7% and DXS10074: 15, 16; DoC 4829, 454; ACR 9%). In both cases the n-1 sequence is present with DoC less than the second called allele. These are not reported in the STRait Razor results because the ACR are less than 20 % for both. In addition, these additional alleles are not visible in the CE data.

3.2.2. UAS FRR vs CE

The concordance between the UAS FRR (the equivalent range which was analyzed in STRait Razor) and Investigator Argus X-12 is over 99 %. Four of the seven loci (DXS7132, DXS7423, DXS8378, and HPRTB) are 100 % concordant between the UAS FRR and CE. In addition to the issues identified above common to both SLR and FRR vs CE at DXS10135 and DXS10074, three samples were discordant at DXS10103. These samples reported a 19.2 microvariant allele by sequence whereas the CE method reported a 19 allele. As previously mentioned, there is a dinucleotide repeat in the 5′ flank of the DXS10103 sequence that is not included within the primers for the CE data, nor is it included in the UAS SLR. A two base insertion (rs112400988) is observed in the repeat of these samples.

3.3. Statistics

3.3.1. Locus and linkage groups

All of the population linkage groups 1 through 3 show an increase in number of haplotypes by sequence compared to length, except for linkage group 3 in the Asian population group, where no increase is observed. Linkage group 4 is represented by a single locus, DXS7423, which in the populations studied, shows no increase in information between length and sequence alleles. Frequencies were calculated for each locus (Supplemental File 2, “Locus - Frequency”) and for each linkage group (worksheet “Linkage Group - Frequency”). The frequency calculations for loci and linkage groups did not include the DXS8378 null allele or the DXS10135 duplication. However, these samples were included in the other loci and linkage groups for frequency calculations. These two samples were not included forensic parameters calculations.

Statistical information for all the linkage groups and loci are included in the Supplemental File 2, “Locus Forensic Parameters” and “Haplotype Forensic Parameters”. Table 2 shows increases in alleles and haplotypes due to sequencing information by locus and linkage group, respectively. By both length and sequence, DXS10135 is the most informative locus. PIC values for the four populations ranged between 0.9114−0.9511 for length and 0.9221−0.9738 for sequence. The least informative locus is DXS7423 for African American and Asian population groups for both length and sequence. There is no difference in information content between length and sequence for this locus. DXS8378 is the least informative locus for Caucasian and Hispanic population groups by both length and sequence.

3.3.2. Linkage disequilibrium

Linkage disequilibrium calculations were performed using Arlequin for both length-based and sequence-based data for each group. The length and sequence data resulted in about the same number of significant (p-value $<$ 0.05) pairs of loci. After Bonferroni correction (p-value $<$ 0.0006) there are still significant pairs present in three of the four populations (Supplemental File 2, “Linkage Disequilibrium”). The following pairings show statistically significant evidence of nonrandom association of alleles after correction for multiple testing: in the African American, Caucasian, and Hispanic population groups, DXS10103 and HPRTB (by sequence); in the Hispanic population, DXS10135 and HPRTB as well as DXS10103 and HPRTB (by length). The DXS10103 and HPRTB are associated with linkage group 3. In the African American and Caucasian population groups DXS10103 and HPRTB (by length) were significant before p-value correction as was the Hispanic population DXS10135 and HPRTB (by sequence).

Evidence of statistical linkage disequilibrium can result for a variety of reasons including distance between markers, founder effect, and mutations, among others [1]. Due to the fact that the U.S. is a “melting pot”, drawing people from all over the globe, the population groups comprising the NIST 1036 sample set are generally representative of continents (Africa, Asia, Europe, and Latin America), with admixture expected in the African American and Hispanic population groups. It is not possible to distinguish subpopulations within these population samples. It is unclear if the samples and sample sizes in each population group are sufficient to properly resolve the linkage disequilibrium; therefore, this information should be considered within the context of other studies.

4. Conclusion and discussion

This NIST X-STR population data is intended to support forensic casework and kinship analyses by providing high-confidence sequence-based allelic and linkage group frequencies. This data was over 99 % consistent with the reported CE data for these samples, with differences largely attributable to varying primer placement between assays. Such differences highlight possible issues which may arise when using both CE and sequence-based data in one analysis, which are the same issues laboratories encounter when comparing data across CE assays with varying locus primer placements. In addition to addressing how to report such differences, it may be useful for laboratories to describe the more common sources of discordance, such as those presented here, in their interpretation protocols. The sequences are publicly available as GenBank records within the STRSeq BioProject (https://www.ncbi.nlm.nih.gov/bioproject/380127) and subproject X-Chromosomal STR Loci (https://www.ncbi.nlm.nih.gov/bioproject/380348) [14].

In order to facilitate implementation of this frequency data, the bioinformatic analysis attempted to mimic the range of sequence contained in the UAS FRR. The lone exception is the DXS10135 locus, where the reported sequence range in this manuscript contains 71 bp less $3'$ sequence than the UAS FRR, due to data quality (in fact, the reported range matches the UAS SLR, which is easily accessible by ForenSeq users). It is possible that the UAS or another bioinformatic analysis may improve these results and extend the reportable range for this locus.

Further work could include the addition of DXS3877 and DXS10148, which are present in the ForenSeq FASTQ files but are not reported by the UAS. These two loci have data quality issues which appear similar to SE33, also present in the ForenSeq FASTQ files but not reported by the UAS [22]. The additional data from DXS10148 and DXS3877 would expand linkage group 1 to include three loci and add a second locus to linkage group 4, respectively. Both of the loci appear to be highly variable by sequence (data not shown).

Supplemental content

All supplemental files are available at the NIST Public Data Repository –U.S. population data for Human Identification Markers (https://doi.org/10.18434/t4/1500024). This includes a STRait Razor v3.0 formatted config file for the 7 X STR loci included in the paper and an excel file which contains worksheets as follows: UAS Reported Sequences, Data Processing Workflow, Discordance Table, Locus – Frequency, Linkage Group – Frequency, Locus Forensic Parameters, Haplotype Forensic Parameters, and Linkage Disequilibrium.

Funding and disclaimer

This work was funded in part by the Federal Bureau of Investigation (FBI) of the United States of America interagency agreements DJF-15-1200-0000174 and DJF-16-1200- 0000246: “Forensic DNA Typing as a Biometric Tool”. Points of view in this document are those of the authors 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

Table 2
Gains of alleles and haplotypes by sequence over length in loci and linkage groups (LG).

	DXS10135	DXS8378	LG1	DXS7132	DXS10074	LG2	DXS10103	HPRTB	LG3	DXS7423	LG4
All	75	6	133	5	26	58	11	5	36	0	0
AfAm	47	3	67	3	19	34	8	3	18	0	0
Asian	9	0	6	2	0	2	0	0	0	0	0
Cauc	41	2	61	0	12	15	9	0	17	0	0
Hisp	24	1	33	0	12	13	9	2	14	0	0

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.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgments

The authors wish to thank Dr. Toni M. Diegoli for assistance with statistical analysis and Dr. Tunde I. Huszar for her review of the manuscript.

References

[1] K. Bentayebi, et al., Genetic diversity of 12 X-chromosomal short tandem repeats in the Moroccan population, Forensic Sci. Int. Genet. 6 (1) (2012) e48–9.

[2] G. Horvath, et al., A genetic study of 12 X-STR loci in the Hungarian population, Forensic Sci. Int. Genet. 6 (1) (2012) e46–7.

[3] X.P. Zeng, et al., Genetic polymorphisms of twelve X-chromosomal STR loci in Chinese Han population from Guangdong Province, Forensic Sci. Int. Genet. 5 (4) (2011) e114–6.

[4] T.M. Diegoli, et al., Allele frequency distribution of twelve X-chromosomal short tandem repeat markers in four U.S. Population groups, Forensic Sci. Int. Genet. Suppl. Ser. 3 (1) (2011) e481–e483.

[5] J.M. Salvador, et al., Filipino DNA variation at 12 X-chromosome short tandem repeat markers, Forensic Sci. Int. Genet. 36 (2018) e8–e12.

[6] F.R. Wendt, et al., Flanking region variation of ForenSeq DNA signature prep kit STR and SNP loci in Yavapai Native americans, Forensic Sci. Int. Genet. 28 (2017) 146–154.

[7] N.M.M. Novroski, et al., Characterization of genetic sequence variation of 58 STR loci in four major population groups, Forensic Sci. Int. Genet. 25 (2016) 214–226.

[8] C. Hussing, et al., The Danish STR sequence database: duplicate typing of 363 danes with the ForenSeq DNA signature prep kit, Int. J. Legal Med. 133 (2) (2019) 325–334.

[9] A. Delest, et al., Sequenced-based French population data from 169 unrelated individuals with Verogen's ForenSeq DNA signature prep kit, Forensic Sci. Int. Genet. 47 (2020) 102304.

[10] C. Phillips, et al., Global patterns of STR sequence variation: sequencing the CEPH human genome diversity panel for 58 forensic STRs using the Illumina ForenSeq DNA signature prep kit, Electrophoresis 39 (21) (2018) 2708–2724.

[11] K.B. Gettings, et al., Sequence-based U.S. Population data for 27 autosomal STR loci, Forensic Sci. Int. Genet. 37 (2018) 106–115.

[12] Verogen, ForenSeq DNA Signature Prep Reference Guide Sept. 2015 cited 2020; Document# 15049528 v01: Available from:, (2015) https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/forenseq/forenseq-dna-signature-prep-guide-15049528-01.pdf.

[13] R. Szibor, et al., Use of X-linked markers for forensic purposes, Int. J. Legal Med. 117 (2) (2003) 67–74.

[14] K.B. Gettings, et al., STRSeq: a catalog of sequence diversity at human identification Short Tandem Repeat loci, Forensic Sci. Int. Genet. 31 (2017) 111–117.

[15] M.C. Kline, et al., STR sequence analysis for characterizing normal, variant, and null alleles, Forensic Sci. Int. Genet. 5 (4) (2011) 329–332.

[16] C.R. Hill, J.M. Butler, P.M. Vallone, A 26plex autosomal STR assay to aid human identity testing*, J. Forensic Sci. 54 (5) (2009) 1008–1015.

[17] Verogen, ForenSeq Universal Analysis Software Guide cited 2020; Document # 15053876 v01: Available from:, (2016) https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/forenseq-universal-analysis-software/forenseq-universal-analysis-software-guide-15053876-01.pdf.

[18] A.E. Woerner, J.L. King, B. Budowle, Fast STR allele identification with STRait razor 3.0, Forensic Sci. Int. Genet. 30 (2017) 18–23.

[19] Y. Lang, F. Guo, Q. Niu, StatsX v2.0: the interactive graphical software for population statistics on X-STR, Int. J. Legal Med. 133 (1) (2019) 39–44.

[20] L. Excoffier, H.E. Lischer, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, Mol. Ecol. Resour. 10 (3) (2010) 564–567.

[21] R. Li, et al., Improved pairwise kinship analysis using massively parallel sequencing, Forensic Sci. Int. Genet. 38 (2019) 77–85.

[22] L.A. Borsuk, et al., Sequence-based US population data for the SE33 locus, Electrophoresis 39 (21) (2018) 2694–2701.