Population Analysis and Forensic Utility of X-Chromosomal Short Tandem Repeat (X-STR) loci

Lisa A. Borsuk*[1], MS, Toni M. Diegoli[2], PhD, Katherine B. Gettings[1], PhD, Carolyn R. Steffen[1], MS, Kevin M. Kiesler[1], MS, and Peter M. Vallone[1], PhD

[1]National Institute of Standards and Technology, Applied Genetics Group, Gaithersburg, MD 20899-8314
[2]Forensic Science Laboratory – Washington, Bureau of Alcohol, Tobacco, Firearms, and Explosives, Beltsville, MD 20705
*Corresponding author lisa.borsuk@nist.gov

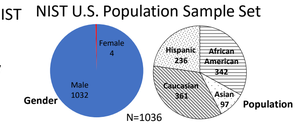
Abstract: X-STR markers are recognized as useful tools to supplement kinship testing in the forensic setting. Numerous studies of allele and haplotype frequencies based on traditional length-based analyses of these loci have been reported in the literature for various population groups. More recently, 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. Here, the details of sequencing and analysis of seven X-STRs in U.S. populations will be presented.

The National Institute of Standards and Technology (NIST) U.S. Population Sample Set consists of 1036 unrelated individuals (1032 male, 4 female) with four population groups represented: African American (n = 342), Asian (n = 97), Caucasian (n = 361), and Hispanic (n = 236). These samples have been sequenced using the MiSeq FGx Forensic Genomics System, including the ForenSeq DNA Signature Prep Kit, which targets important STR markers commonly used for human identification and relationship testing [1]. Seven X-STR loci are included in this assay: DXS10135, DXS10074, DXS7132, DXS10103, DXS7423, DXS8378, HPRTB [2], with at least one marker representing each of the four linkage groups found on the X-chromosome. The core repeat region as well as flanking region variation was assessed with a customized bioinformatic approach. This approach also detected two additional X-STR loci (DXS10148 and DXS8377) which are sequenced with the assay but not reported in the associated Universal Analysis Software (UAS). These two 'extra' loci are being evaluated for potential inclusion in the population set.

Sequence-based allele and haplotype frequencies along with other relevant population genetic parameters for each population group were determined. Results from this study were compared to allele calls and frequencies derived from previous analyses using length-based methods [3]. 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 will be made publicly available at NCBI STRSeq X Chromosomal STR Loci BioProject accession PRJNA380348 [4].

Materials and Methods:

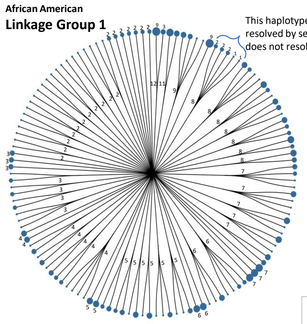
The NIST U.S Population sample set has been evaluated using a number of capillary electrophoresis (CE) and sequencing kits for human identification.



Four African American male samples were removed from this analysis due to poor quality of X-STR sequencing or CE results. One African American male sample appears to be XXY. The female sample populations are 1 African American, 1 Asian, and 2 Caucasian. Length-based genotypes were previously generated for this sample set using the Qiagen Investigator Argus X-12 kit[3].

Using Illumina's FGx MiSeq and ForenSeq kit 1036 samples were sequenced. The FASTQ files were trimmed using BBDuk [5] and analyzed using a modified version of STRait Razor v2.0 [6] with a modified configuration file. The resulting files were processed to identify the length and sequenced-based allele calls. A set of allele calling rules were established. The additional loci required separate hands on evaluation after going through the analysis process.

African American
Linkage Group 1

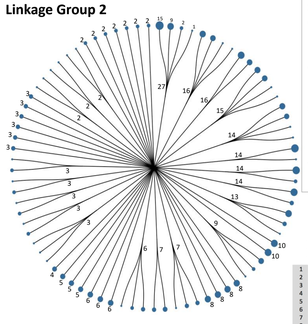


	African American Linkage Group 1 Length	Sequence
1	34	96
2	18	25
3	7	16
4	7	12
5	8	6
6	4	3
7	8	4
8	7	
9	2	2
10		
11	1	
12	1	
N	336	336
T#H	97	164

Distribution of Haplotypes by Length and Sequence – the numbers on the left side of the tables are number of observations made for each haplotype group. The Length and Sequence columns are the number of different haplotypes that contain the same number of observations. The tables are broken up by population and linkage group. Located on the bottom left is T#H, which is total number of haplotypes.

Examples – In the table of African American Linkage Group 1 haplotypes above, there are 18 different length-based haplotypes that are observed twice. There are 25 different sequenced-based haplotypes that are observed twice. There is 1 length-based haplotype that is observed 12 times (gray arrow). This length-based haplotype is separated by sequencing into 2 haplotypes, one that is observed 9 times (green arrow) and one that is observed 3 times (purple arrow). The colored arrows indicate where these sets are represented in the above table.

African American
Linkage Group 2

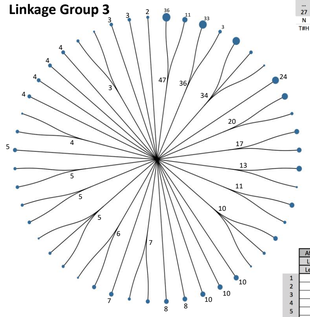


	African American Linkage Group 2 Length	Sequence
1	22	47
2	10	15
3	11	12
4	1	2
5	3	4
6	4	7
7	2	2
8	5	5
9	1	2
10	2	3
11		1
12	1	1
13	1	
14	4	
15	1	1
16	2	
27	1	
N	336	336
T#H	69	103

	Asian Linkage Group 1 Length	Sequence	Linkage Group 2 Length	Sequence	Linkage Group 3 Length	Sequence	
1	24	33	8	10	9	9	1
2	9	7	6	6	2	2	2
3		1	5	6	1	1	3
4	7	5	3	2	1	1	4
5						2	5
6	2	2	1	1	2	2	6
7	2	2			1	1	7
8							8
9					1	1	9
10			1	2	1	1	10
11			3	2	1	1	11
12					1	1	12
13					1	1	13
14							14
N	96	96	96	96	96	96	N
T#H	44	50	27	29	21	21	T#H

	Caucasian Linkage Group 1 Length	Sequence	Linkage Group 2 Length	Sequence	Linkage Group 3 Length	Sequence	
1	25	73	16	29	9	20	1
2	12	24	5	9	6	9	2
3	8	10	5	4	3	4	3
4	7	12	2	1		2	4
5	6	9	1	1			5
6	2	6	2	3	1	3	6
7	4	4	2	2	4	3	7
8	2	1	3	3	1	3	8
9	4	1	1		1		9
10	3	1	1	2	1		10
11	2	1	1			1	11
12	1	1	1	2			12
13	2		1			3	13
14	1		2	2	3		14
15	1			1	2		15
16	1		2	2	1	2	16
17			1		1	1	17
18				1	1		18
19					1	1	19
20							20
21			1	1			21
22			1				22
23				1		1	23
24			1	1			24
25			1		1		25
...							...
30				1			30
31			1				31
...							...
42						1	42
...							...
46					1		46
...							...
55					1	1	55
N	358	358	359	359	359	359	N
T#H	81	143	51	66	38	55	T#H

African American
Linkage Group 3



	African American Linkage Group 3 Length	Sequence
1	7	14
2	1	10
3	3	7
4	5	4
5	4	3
6	1	
7	2	2
8	2	3
9		2
10	4	3
11	1	1
12		1
13	1	
14		
15		
16		
17	1	
18		1
19		
20	1	
24	1	1
...		
33		1
34	1	
35		1
36	1	1
...		
47	1	
N	336	336
T#H	37	55

	Hispanic Linkage Group 1 Length	Sequence	Linkage Group 2 Length	Sequence	Linkage Group 3 Length	Sequence	
1	30	58	22	30	6	14	1
2	8	19	7	10	2	6	2
3	10	9	5	9	3	4	3
4	4	2	4	2	1	3	4
5	3	6	1	3	2	6	5
6	4	5	4	3	2	2	6
7	3		2	1	1		7
8	3	1	1	1	2		8
9	2		1	2	1	2	9
10		1	2	1	2		10
11			2	1	2	2	11
12	1				2	1	12
13		1	1	1			13
14		1					14
15	2		1	1			15
16							16
17				1			17
18							18
19			1	1			19
20			1				20
...							...
22						1	22
23					1		23
...							...
26					1	1	26
...							...
43					1	1	43
N	236	236	236	236	235	235	N
T#H	70	103	55	67	29	43	T#H

Improvements in Discrimination with Sequence Information for Non-Unique African American Haplotypes– The spokes of the diagram represent individual haplotypes as determined by length-based data. If the spoke splits, it is due to additional haplotypes being distinguished with the addition of sequence data. The numbers in the plot at the branching and the nodes are the number of times that length- or sequenced-based haplotype was observed respectively. The nodes are scaled within each plot.

An example in African American Linkage Group 3 (above) is a set of 47 samples that have the same haplotype by length (located at the 12 o'clock position). This length-based haplotype is split into two haplotypes by sequence. One sequenced-based haplotypes was observed 36 times and the other 11 times.

Materials and Methods continued.

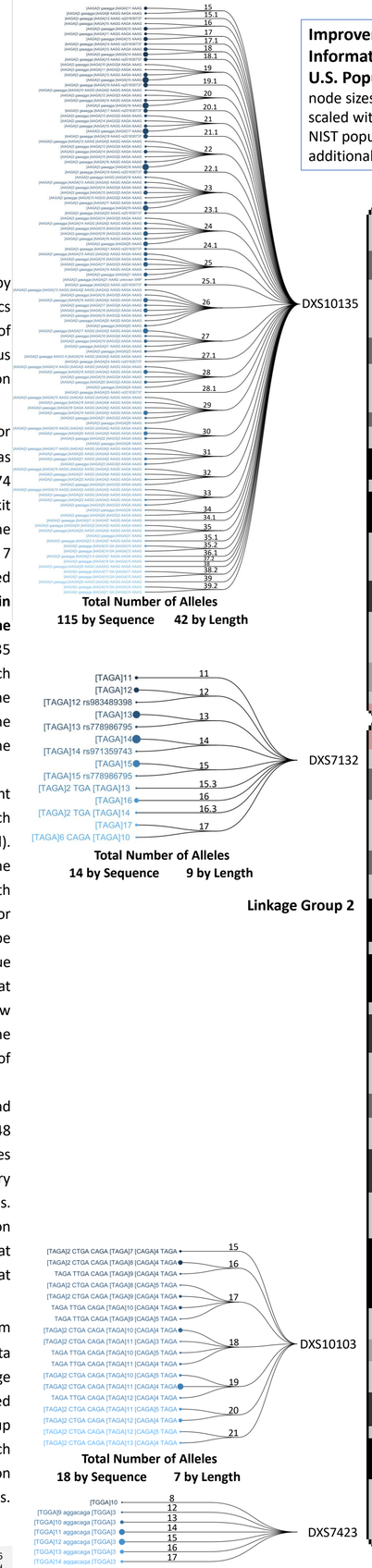
Allele and haplotype frequencies for each population were calculated by hand using a spreadsheet program. Forensic efficiency statistics (genetic/haplotype diversity, polymorphism information content, power of discrimination in both males and females, mean exclusion chance for various scenarios) were calculated using the Forensic ChrX Research website version 2.0 [7] and/or StatsX v2.1 [8].

Results: Length- and sequence-based data was completely generated for 1032 of the 1036 samples for the 7 X-STR loci in the ForenSeq Kit. There was one discordance between length and sequence data found. At the DXS10074 locus, no results were generated with the Qiagen Investigator Argus x-12 kit whereas sequencing data was obtained. Otherwise, concordance of the length-based allele calls between the two methods was found for these 7 loci. For 6 of the 7 loci, sequencing increased the number of alleles observed for the combined U.S. population. This can be seen in the **Improvements in Discrimination with Sequence Information for Alleles at 7 X-STR Loci in the Combined U.S. Population figure on the right**. In this data set, DXS10135 had the greatest increase in the number of alleles by sequence, which included a combination of repeat variation and SNPs identified in the additional flanking sequence. This is an almost three fold increase in the number of alleles from 42 by length to 115 by sequence. Sequencing of the DXS7423 locus not provide any additional information for this data set.

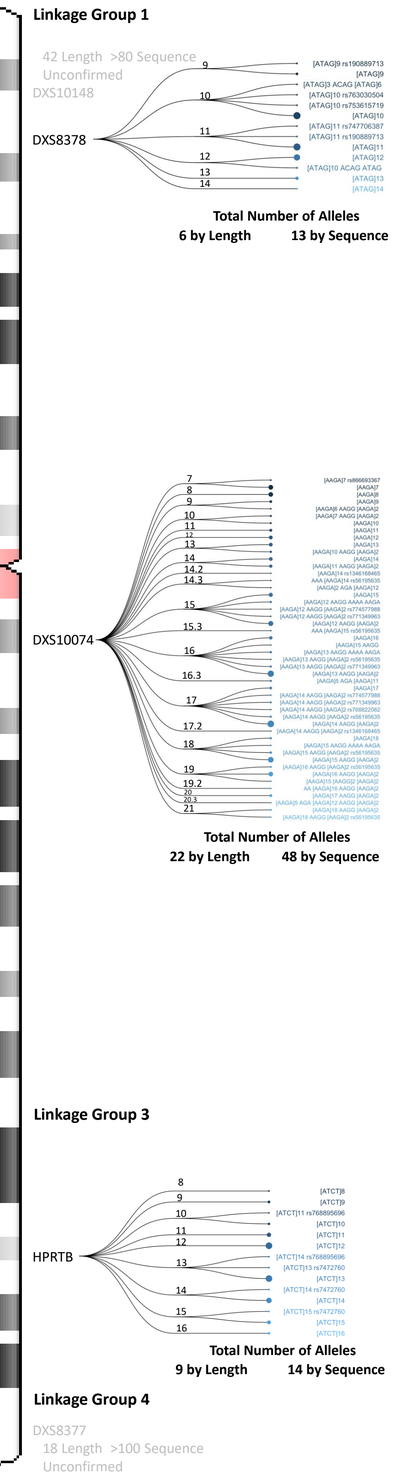
Forensic efficiency statistics remained the same or showed improvement with the sequencing data as compared to the length-based for each population considering both alleles and haplotypes (data not presented). There was an increase in the number of haplotypes overall as well as the number of unique haplotypes observed with sequencing data than with length-based methods for all linkage groups and populations, except for linkage group 3 for the Asian population, which did not change. This can be seen in the **linkage group tables to the left**. The number of unique haplotypes increased three-fold for the African American population at Linkage Group 1. The **linkage group plots to the left** demonstrate how haplotypes are further resolved by sequencing. In linkage group 3 for the African American population, there are an additional 18 haplotypes with 7 of them being unique.

The additional loci DXS10148 and DXS8377 were evaluated separately and were not included in the linkage group analysis. Analysis of the DXS10148 locus was complicated by sequencing noise, resulting in multiple sequences of the same length being called with significant coverage. DXS8377 had very low coverage and high stutter. In many cases there was no call for this locus. Additionally, no capillary electrophoresis data was available for comparison for DXS8377. Both DXS10148 and DXS8377 have complicated repeat structures and demonstrate, with sequences that were obtained, that sequencing would add information for these loci.

Conclusion: For the 7-X STR loci present in the ForenSeq Kit, results from both sequence- and length-based methods were consistent. Sequencing data added additional information that improved discrimination within linkage groups. All populations experienced gains in the number of observed haplotypes through sequencing except the Asian population at linkage group 3. However, a number common haplotypes observed multiple times in each population persisted despite these gains. The African American population samples had the largest overall increase in the number of haplotypes. Additional evaluation of all loci, including the 'extra' loci, is continuing.



Improvements in Discrimination with Sequence Information for Alleles at 7 X-STR Loci in the Combined U.S. Population – STRs are placed in approximate positions. End node sizes are not consistent between dendrograms. They are scaled within each plot to give an idea of allele distribution in the NIST population for each locus. The light gray loci are the additional X-STRs not reported in the UAS.



References:

1. Gettings, K.B.; Borsuk, L.A.; Steffen, C.R.; Kiesler, K.M.; Vallone, P.M.; U.S. Population Sequence Data for 27 Autosomal STR Loci. Forensic Science International: Genetics. 2018; https://doi.org/10.1016/j.fsigen.2018.07.013
2. Verogen, Inc. ForenSeq DNA Signature Prep Reference Guide. Sept. 2015. Document# 15049528 v01
3. Diegoli, T.M.; Linacre, A.; Vallone, P.M.; Butler, J.M.; Coble, M.D.; Allele frequency distribution of twelve X-chromosomal short tandem repeat markers in four U.S. population groups, Forensic Science International Genetics Supplement Series 3 (2011) e481-e483.
4. Gettings, K.B.; Borsuk, L.A.; Ballard, D.; Bodner, M.; Budowle, B.; Devesse, L.; King, J.; Parson, W.; Phillips, C.; Vallone, P.M.; STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci. Forensic Science International: Genetics. 2017 Nov.;31:111-117; https://doi.org/10.1016/j.fsigen.2017.08.017
5. BBDuk - http://seqanswers.com/forums/showthread.php?t=42776
6. STRait Razor v2.0: the improved STR Allele Identification Tool--Razor. Warshauer et al., Forensic Sci Int Genet. 2015 (14):182-6
7. R. Szibor, S. Hering, J. Edelmann, A new Web site compiling forensic chromosome X research is now online, Int. J. Legal Med. 120 (2006) 252–254.
8. Y. Lang, F. Guo, Q. Niu, StatsX v2.0: the interactive graphical software for population statistics on X-STR, Int. J. Legal Med. 133 (2019) 39-44.

Poster # B2 at:
71st AAFS Meeting,
Baltimore, MD
February 18-23, 2019