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

Lisa A. Borsuk1, MS, Toni M. Diegoli2, PhD, Katherine B. Gettings1, PhD, Carolyn R. Steffen1, MS, Kevin M. Kiesler3, MS, and Peter M. Vallerone1, PhD

1National Institute of Standards and Technology, Applied Genetics Group, Gaithersburg, MD 20899-8314
2Forensic Science Laboratory – Washington, Bureau of Alcohol, Tobacco, Firearms, and Explosives, Beltsville, MD 20705
3Corresponding 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 sequencing-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 1032 unrelated individuals (510 male, 4 female) with four population groups represented: African American (n = 434), Asian (n = 97), Caucasian (n = 361), and Hispanic (n = 236). These samples have been sequenced using the MiSeq DNA Sequencing 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 study: DXS1013, DXS9207, DXS7183, DXS7103, DXS7143, DXS8373, and HMBDR [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 DXS3177) 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 on NCBI STRbase X Chromosomal STR (X) locus accession PRJNA301044 [4].

Materials and Methods: The NIST U.S. Population Sample Set 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 XX. 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 Qigen Investigator Argus X-12 kit.

Using Illumina’s MiSeq and ForenSeq kit 1036 samples were sequenced. The FASTQ files were trimmed using BBduk [5] and analyzed using a modified version of StathRazer v2.0.5 [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 was established. The additional loci required separate hands-on evaluation after going through the analysis process.

Features:

- 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 to both males and females, mean exclusion chance for various scenarios) were calculated using the Forensic ChRk Research website version 2.0 [7] and/or Statsoft v2.1 [8].

Results: Length-based 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 DXS10174 locus; no results were generated with the Qigen 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 all of the 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 DXS423 locus did 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 is an additional 18 haplotypes with 7 of them being unique. The additional loci DXS10148 and DXS3177 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. DXS3177 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 DXS3177. DXS10148 and DXS3177 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. Sequence 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, 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 – STs are placed in approximate positions. End node cases 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.