B192 United States Population Sequence Data for 27 Autosomal Short Tandem Repeat (STR) Loci, 24 Y-Chromosomal Short Tandem Repeat (Y-STR) Loci, and 7 X-Chromosomal Short Tandem Repeat (X-STR) Loci

Katherine B. Gettings, PhD, NIST, 100 Bureau Drive, MS 8314, Gaithersburg, MD 20899; Kevin Kiesler, MS, 100 Bureau Drive, MS 8314, Gaithersburg, MD 20899; Becky Steffen, MS, NIST, 100 Bureau Drive, MS 8314, Gaithersburg, MD 20899; Lisa Borsuk, MS, NIST, 100 Bureau Drive, Gaithersburg, MD 20899; and Peter M. Vallone, PhD, 100 Bureau Drive, Gaithersburg, MD 20899-8311*

After attending this presentation, attendees will better understand sequence-based STR allele frequencies, how this new data type can be implemented into statistical analysis, and which STR loci exhibit significant allelic gains by sequence in a large population dataset representative of four common United States populations.

This presentation and the availability of this data will impact the forensic science community by facilitating sequence-based statistical calculations when STR sequencing applications are utilized. Recent publications have been limited to smaller sample sizes, less loci, or populations less representative of the United States.[1-3]

STR loci allows for determination of repeat motif variations within the STR (or entire PCR amplicon) that cannot be ascertained by size-based Polymerase Chain reaction (PCR) fragment analysis. Sanger sequencing has been used in research laboratories to further characterize STR loci, but it is currently impractical for routine forensic use due to the laborious nature of the procedure in general and additional steps required to separate heterozygous alleles. Recent advances and cost reductions in Next Generation Sequencing (NGS), also known as massively parallel sequencing, have opened the door to routinely obtaining STR sequence data in the forensic laboratory, which may be particularly valuable in challenging cases of mixture interpretation or complex kinship. Thus, several commercial manufacturers have developed assays for sequencing various combinations of STR loci (autosomal, X, and Y), with some also containing Single Nucleotide Polymorphisms (SNPs) and mitochondrial DNA (mtDNA).

This sequence-based population study includes >1,000 DNA samples, divided among Caucasian, African American, Hispanic, and East Asian individuals, at 27 autosomal STR loci, 24 Y-STR loci, and 7 X-STR loci using an assay designed for NGS. Data analysis for genotyping was performed using the manufacturer's software and an alternate in-house pipeline based on freeware.[4,5] Allele calls for Capillary Electrophoresis (CE) and sequencing methods were compared for concordance at every locus. In the case of a discordant allele call, the cause was determined.

The resulting population data reveals general gains in discriminatory power per locus, as well as population-specific gains for some loci. Categorizing the 58 STR loci as simple or compound/complex based on the motif present in GRCh38 Human Genome Reference Sequence, the 22 compound/complex loci exhibit, on average, more than double the number of alleles by sequence compared to alleles by length, with some loci containing more than four or five times the alleles. The 36 simple repeat loci demonstrate a more modest increase in alleles by sequence compared to alleles by length, as expected. For the autosomal STR loci, the resulting sequence-based frequency data is applied to generate Random Match Probabilities (RMP) for each individual per locus and across the profile, and comparison to length-based RMP values reveals expected gains in statistics. Y-STR and X-STR loci are evaluated for gains in haplotype diversity.

This data is fundamental to implementation and support of sequencing technology, and the demonstrated gains in alleles and RMP values will assist laboratories in weighing costs/benefits.

Reference(s):

1. Gelardi C., Rockenbauer E., Dalsgaard S., Borsting C., Morling N. Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles. *Forensic Sci Int Genet.* 12 (2014) 38-41.
2. van der Gaag K.J., de Leeuw R.H., Hoogenboom J., Patel J., Storts D.R., Laros J.F., et al. Massively parallel sequencing of short tandem repeats-Population data and mixture analysis results for the PowerSeq system. *Forensic Sci Int Genet.* 24 (2016) 86-96.
3. Wendt F.R., Churchill J.D., Novroski N.M.M., King J.L., Ng J., Oldt R.F., et al. Genetic analysis of the Yavapai Native Americans from West-Central Arizona using the Illumina MiSeq FGx™ Forensic Genomics System. *Forensic Science International: Genetics.* (2016).
4. Illumina, ForenSeq Universal Analysis Software Guide, Part #4470483 Rev. A, (2011).
5. Warshauer D.H., King J.L., Budowle B. STRait Razor v2.0: The improved STR Allele Identification Tool – Razor. *Forensic Science International: Genetics.* 14 (2015) 182-186.

STR Sequence, United States Population, Allele Frequency