Sequencing of the highly polymorphic STR locus SE33

National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce

Email:
Lisa.Borsuk@nist.gov

L.A. Borsuk[1], K.B. Gettings[1], C.R. Steffen[1], K.M. Kiesler[1], and P.M. Vallone[1]

[1]U.S. National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA

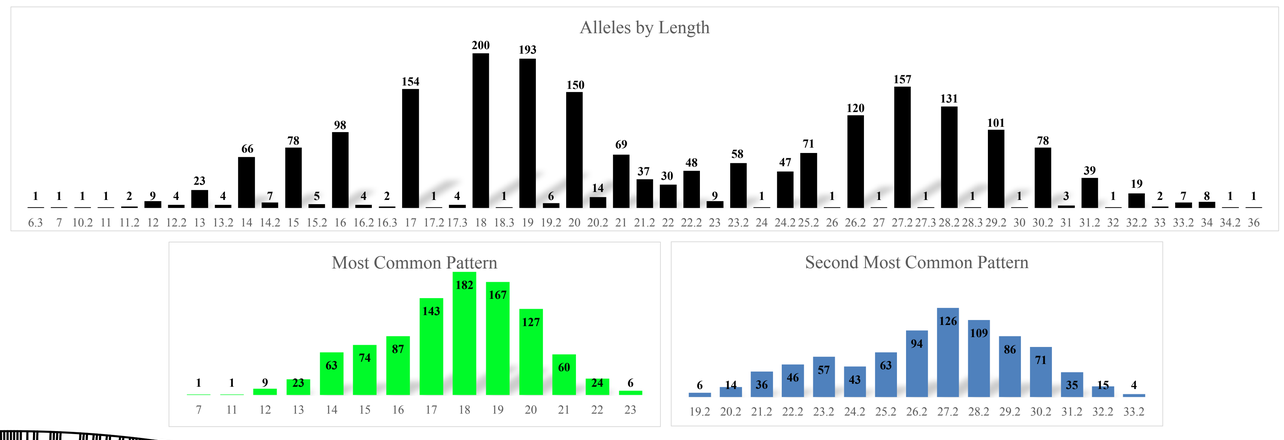
P01-53
170253

ABSTRACT – The NIST U.S. Population Sample Set consists of 1036 unrelated individuals. There are four population groups represented: African American (n = 342), Asian (n = 97), Caucasian (n = 361), and Hispanic (n = 236). These samples have been analyzed using next generation sequencing technology targeting important STR sequences commonly used for human identification. The analysis of SE33 included in this data set required a customized bioinformatic approach to identify and process the allelic information. The locus SE33 is one of the most polymorphic markers used by the forensic community [1]. SE33 is a highly variable locus by length and sequencing has resulted in a four-fold increase in the number of observed alleles. The NIST Population Sample Set has an observed range of 6.3 to 36 tetranucleotide repeats [2]. It has 52 unique alleles by length and 264 unique alleles by sequence. Analysis of this data set shows 100% concordance with length based methods when flanking sequence is considered. The different categories (classes) of repeat motifs revealed will be illustrated and further stratified by population group.

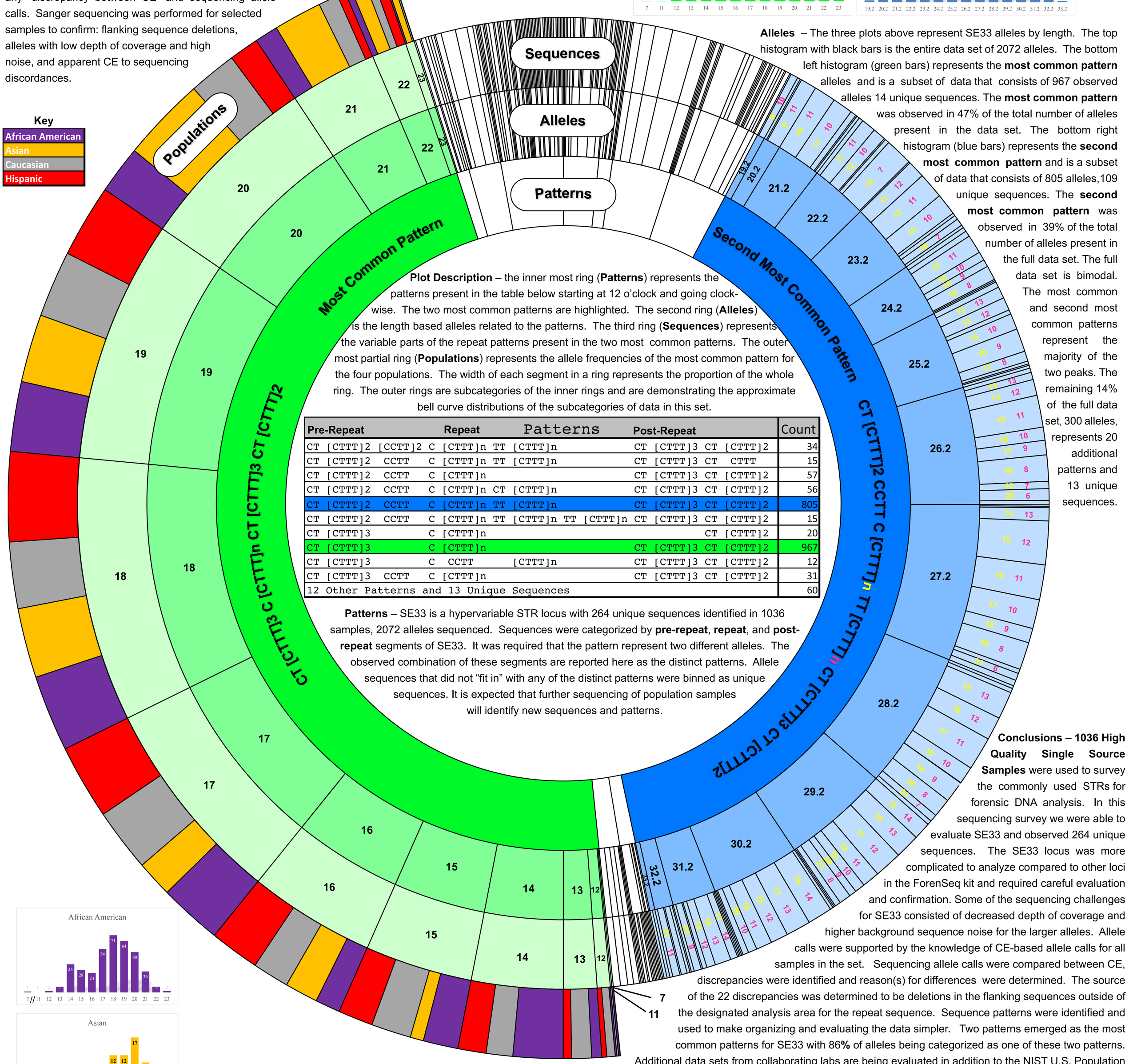
Specific Aim: Determine sequences of the highly polymorphic STR locus SE33 using next generation sequencing of the NIST population samples.

Method - Using Illumina's FGx MiSeq and ForenSeq kit 1036 high quality single source samples were sequenced. The FASTQ files were trimmed using BBDuk [3] and analyzed using a modified version of STRait Razor v2.0 [4] 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 including minimum depth of coverage at 30x and SE33 specific allele coverage ratios.

Interpretation of SE33 necessitated relaxing typical allele coverage ratio constraints (e.g. below 0.3). It also required additional curation and allele recovery based on knowledge of CE allele lengths. Flanking sequences were analyzed separately using a modified version of STRait Razor v2.0, which evaluated an additional 75 bases beyond the bioinformatic anchor site located at the end of the 3' repeat. Additional evaluation of the flanking sequences beyond the 75 bases was conducted to explain any discrepancy between CE and sequencing allele calls. Sanger sequencing was performed for selected samples to confirm: flanking sequence deletions, alleles with low depth of coverage and high noise, and apparent CE to sequencing discordances.



Alleles – The three plots above represent SE33 alleles by length. The top histogram with black bars is the entire data set of 2072 alleles. The bottom left histogram (green bars) represents the **most common pattern** alleles and is a subset of data that consists of 967 observed alleles 14 unique sequences. The **most common pattern** was observed in 47% of the total number of alleles present in the data set. The bottom right histogram (blue bars) represents the **second most common pattern** and is a subset of data that consists of 805 alleles,109 unique sequences. The **second most common pattern** was observed in 39% of the total number of alleles present in the full data set. The full data set is bimodal. The most common and second most common patterns represent the majority of the two peaks. The remaining 14% of the full data set, 300 alleles, represents 20 additional patterns and 13 unique sequences.



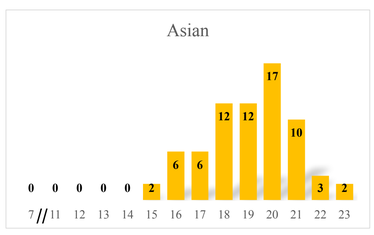
Plot Description – the inner most ring (**Patterns**) represents the patterns present in the table below starting at 12 o'clock and going clock-wise. The two most common patterns are highlighted. The second ring (**Alleles**) is the length based alleles related to the patterns. The third ring (**Sequences**) represents the variable parts of the repeat patterns present in the two most common patterns. The outer most partial ring (**Populations**) represents the allele frequencies of the most common pattern for the four populations. The width of each segment in a ring represents the proportion of the whole ring. The outer rings are subcategories of the inner rings and are demonstrating the approximate bell curve distributions of the subcategories of data in this set.

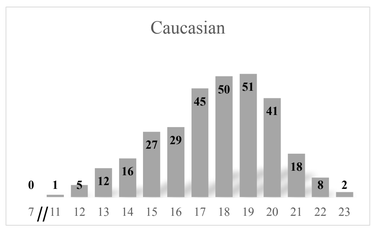
Pre-Repeat	Repeat	Patterns	Post-Repeat	Count
CT [CTTT]2 [CCTT]2	C [CTTT]n TT [CTTT]n		CT [CTTT]3 CT [CTTT]2	34
CT [CTTT]2 CCTT	C [CTTT]n TT [CTTT]n		CT [CTTT]3 CT CTTT	15
CT [CTTT]2 CCTT	C [CTTT]n		CT [CTTT]3 CT [CTTT]2	57
CT [CTTT]2 CCTT	C [CTTT]n CT [CTTT]n		CT [CTTT]3 CT [CTTT]2	56
CT [CTTT]2 CCTT	C [CTTT]n TT [CTTT]n		CT [CTTT]3 CT [CTTT]2	805
CT [CTTT]2 CCTT	C [CTTT]n TT [CTTT]n TT [CTTT]n		CT [CTTT]3 CT [CTTT]2	15
CT [CTTT]3	C [CTTT]n		CT [CTTT]2	20
CT [CTTT]3	C [CTTT]n		CT [CTTT]3 CT [CTTT]2	967
CT [CTTT]3	C CCTT	[CTTT]n	CT [CTTT]3 CT [CTTT]2	12
CT [CTTT]3 CCTT	C [CTTT]n		CT [CTTT]3 CT [CTTT]2	31
12 Other Patterns and 13 Unique Sequences				60

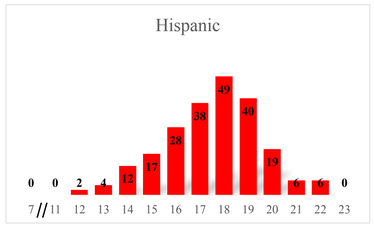
Patterns – SE33 is a hypervariable STR locus with 264 unique sequences identified in 1036 samples, 2072 alleles sequenced. Sequences were categorized by **pre-repeat**, **repeat**, and **post-repeat** segments of SE33. It was required that the pattern represent two different alleles. The observed combination of these segments are reported here as the distinct patterns. Allele sequences that did not "fit in" with any of the distinct patterns were binned as unique sequences. It is expected that further sequencing of population samples will identify new sequences and patterns.

Conclusions – 1036 High Quality Single Source Samples were used to survey the commonly used STRs for forensic DNA analysis. In this sequencing survey we were able to evaluate SE33 and observed 264 unique sequences. The SE33 locus was more complicated to analyze compared to other loci in the ForenSeq kit and required careful evaluation and confirmation. Some of the sequencing challenges for SE33 consisted of decreased depth of coverage and higher background sequence noise for the larger alleles. Allele calls were supported by the knowledge of CE-based allele calls for all samples in the set. Sequencing allele calls were compared between CE, discrepancies were identified and reason(s) for differences were determined. The source of the 22 discrepancies was determined to be deletions in the flanking sequences outside of the designated analysis area for the repeat sequence. Sequence patterns were identified and used to make organizing and evaluating the data simpler. Two patterns emerged as the most common patterns for SE33 with 86% of alleles being categorized as one of these two patterns. Additional data sets from collaborating labs are being evaluated in addition to the NIST U.S. Population Sample Set. Unique SE33 alleles will be uploaded to STRseq (See presentation O04 [5] and poster P01-49), SE33 BioProject PRJNA380562 at NCBI.

A copy of this poster will be available at: http://strbase.nist.gov/NISTpub.htm#Presentations







Populations – This is the most common SE33 sequence pattern by population. The histograms to the left represent the counts of the ***most common pattern*** observed alleles separated by population. Note: the histograms Y axes are not on the same scale. The **most common pattern** represents 54% of all the African American alleles, but only 36% of all of the Asian alleles observed in our data set. For the Caucasian and Hispanic populations the **most common pattern** represents 42% and 47% of their alleles, respectively. The patterns and how they relate to the population data continues to be explored.

Funding
FBI Biometrics Center of Excellence (BCOE): *Forensic DNA Typing as a Biometric tool.*

NIST Special Programs Office: *Forensic DNA.*

References
[1] Butler JM, Hill CR, Kline MC, Duewer DL, Sprecher CJ, McLaren RS, Rabbach DR, Krenke BE, and Storts DR: The single most polymorphic STR Locus: SE33 performance in U.S. population. Forensic Science International: Genetics Supplement Series. 2009; 2:23-24
[2] Hill CR, Duewer DL, Kline MC, Coble MD, and Butler JM: U.S. population data for 29 autosomal STR loci. Forensic Science International: Genetics. 2013; 7:e82-e83
[3] BBDuk - http://seqanswers.com/forums/showthread.php?t=42776
[4] STRait Razor v2.0: the improved STR Allele Identification Tool--Razor. Warshauer et al., Forensic Sci Int Genet. 2015 (14):182-6
[5] "STRseq: a resource for sequence-based STR analysis" is available at: http://strbase.nist.gov/NISTpub.htm#Presentations