STR Sequence Diversity in Population Samples and Nomenclature Guidance for the "Next Generation"

Katherine B. Gettings[1], Seth A. Faith[2], Brian Young[2], Esley Heizer Jr.[2], Kevin M. Kiesler[1], Elizabeth Montano[2], Christine Baker[2], Angela Minard-Smith[2], Richard Guerrieri[2] and Peter M. Vallone[1]

[1]National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899

[2]Battelle, 505 King Avenue, Columbus, OH 43201

ABSTRACT

As STR loci were being identified in the 1990s, various nomenclature systems were developed for different loci, with the primary variation being whether or not to "count" non-repeat bases interspersed in the repeat motif. In 1997, the ISFG issued guidelines on STR nomenclature, in an attempt to provide a common currency for information exchange. Historical precedent already existed for some loci, and this was maintained to avoid confusion, resulting in several commonly used forensic loci having complicated and contradictory nomenclature systems. This has not been an issue within the forensic community, as the capillary electrophoresis (CE)-length analyses are kit-based, with corresponding computer programs that automatically count repeats in a standardized manner. Now, as the costs associated with next-generation sequencing (NGS) methods decline, forensic research laboratories are beginning to explore the increase in information sequencing STR loci may provide. As a new generation of scientists begins interrogating these loci on a deeper level, an understanding of historical nomenclature is needed to achieve bioinformatic concordance with existing CE data. In the work presented here, NGS results from population samples exemplify the sequence variation that exists in forensic STR loci (SNPs and InDels within and outside of STR allele regions and repeat motif changes) as well as the complexity and inconsistency of the current nomenclature. This experimental sequence data gives an indication of the level of diversity expected in the larger population and provides examples of how sub-alleles can improve discrimination and mixture deconvolution in forensic casework. The different purposes of nomenclature—manual comparisons, forensic reports, database searching, court explanations—are discussed and examples of possible NGS-compatible nomenclature systems that may meet the needs of the forensic community are shown.

MATERIALS and METHODS

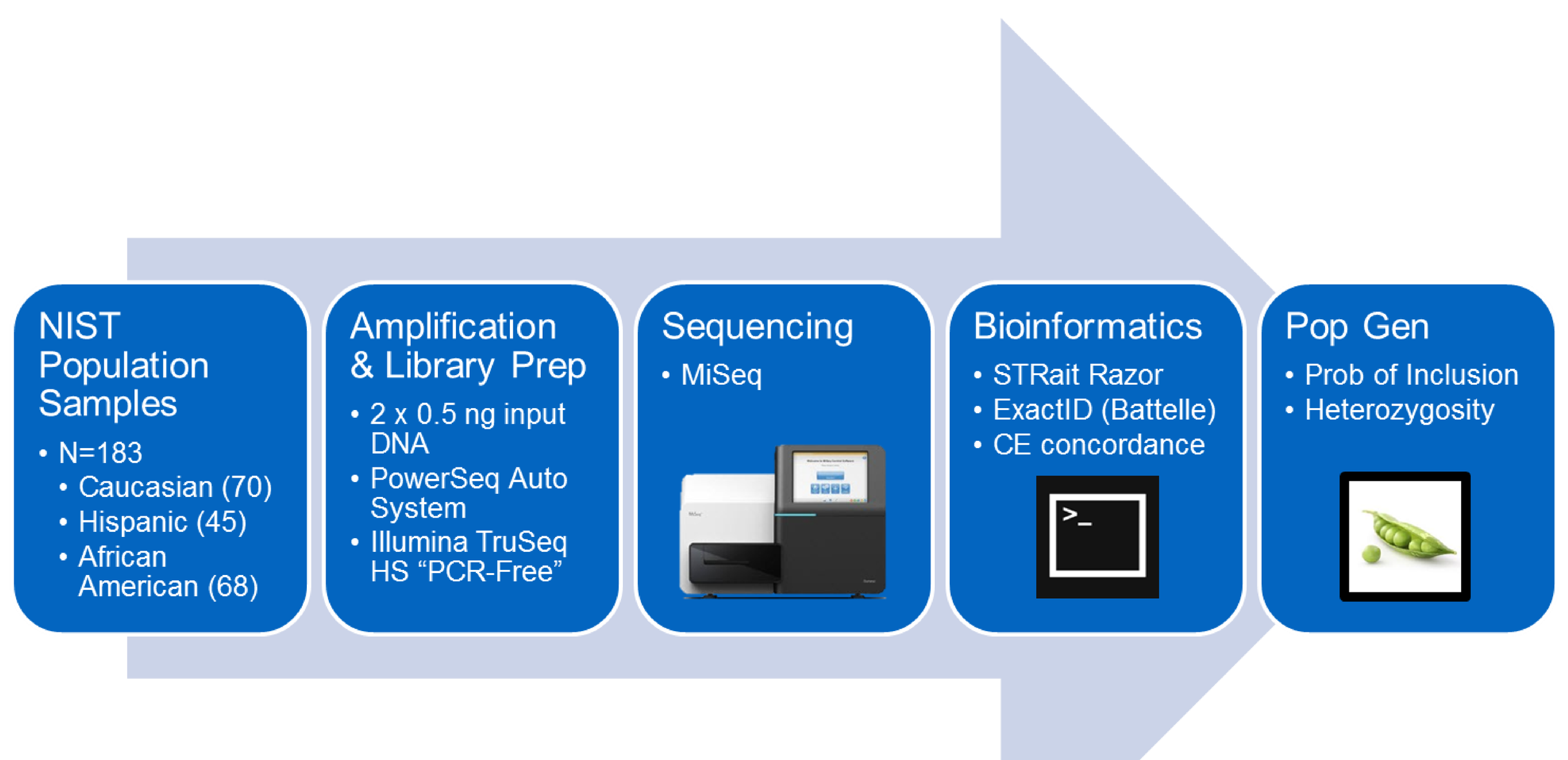


Figure 1. Overview of experimental design.

NIST population samples (N=183 consisting of 70 Caucasian, 68 African American, and 45 Hispanic individuals) were amplified twice in 96-well plates, with 0.5 ng input DNA per sample in 25 µL reaction volumes. Duplicate amplicons were combined during the clean-up step, prior to library generation. Sequencing template libraries were prepared in 96-well format with the TruSeq DNA PCR-Free Sample Preparation Kit HS (Illumina, San Diego CA, USA). Sequencing was performed in two runs (96 samples/run) on the MiSeq system (Illumina) using the 600 cycle MiSeq Reagent Kit v3 (Illumina).

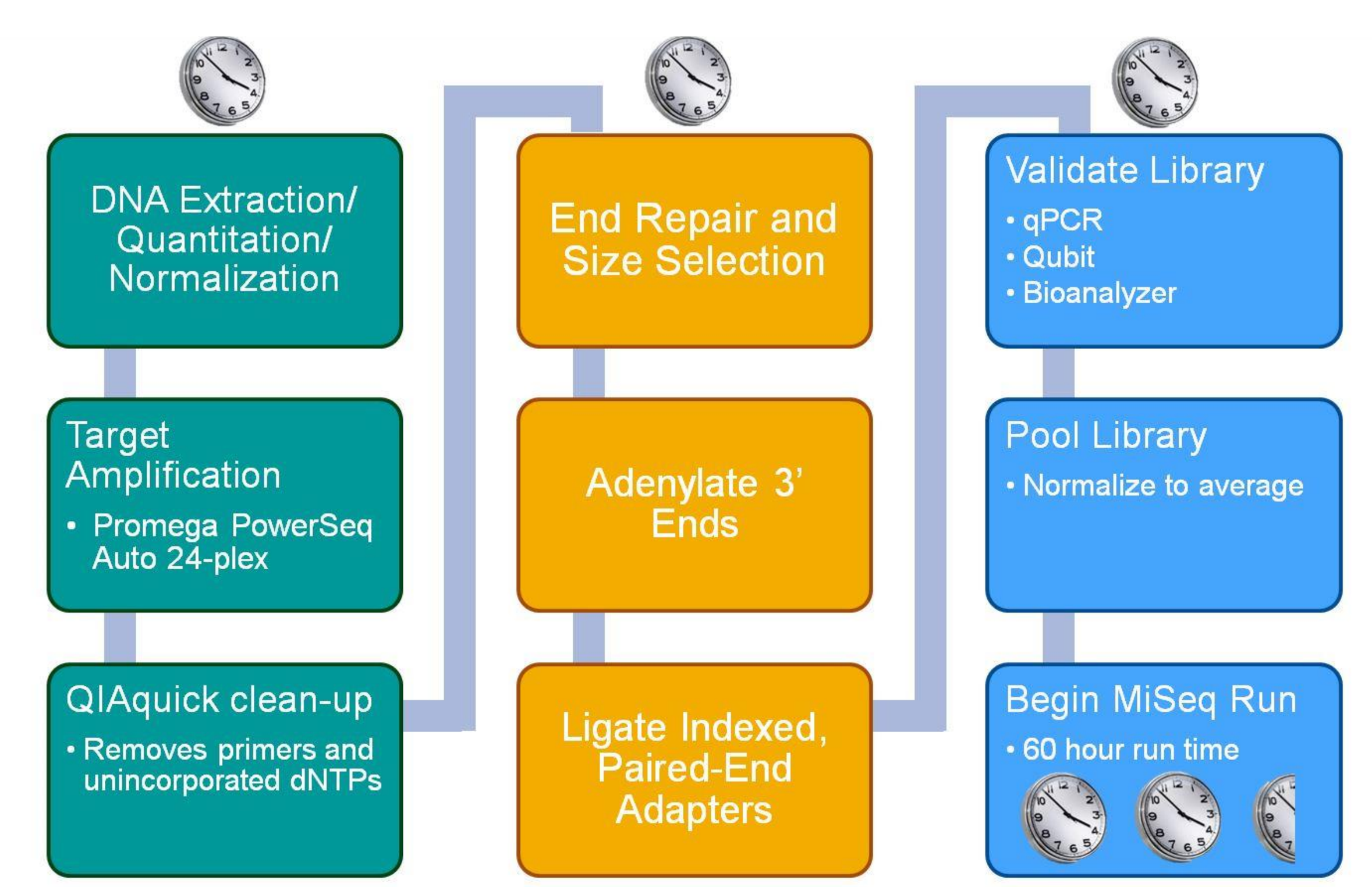


Figure 2. Overview of sample preparation workflow. Each clock represents a minimum of one day. All steps other than the actual MiSeq run were performed manually in a 96-well format.

Analysis of .fastq files to produce STR allele calls was performed with two different bioinformatic pipelines: ExactID (Battelle Memorial Institute, Columbus OH, USA, see ISHI 2014 poster #69 for more information), and STRait Razor [1]. Allelic balance based on coverage was evaluated to determine zygosity. Only majority sequences (two for heterozygotes or one for homozygotes) were considered as evidence supporting allele calls, and only the repeat regions of the majority sequences were analyzed further (e.g. sequences that were consistent with stutter, and sequences that did not match the majority sequence within the repeat region were excluded from further analysis). Genotypes from both ExactID and STRaitRazor were independently analyzed for concordance to CE based genotypes (generated previously with PowerPlex Fusion (Promega)). Discordances were evaluated further to determine the true genotype/sequence.

RESULTS

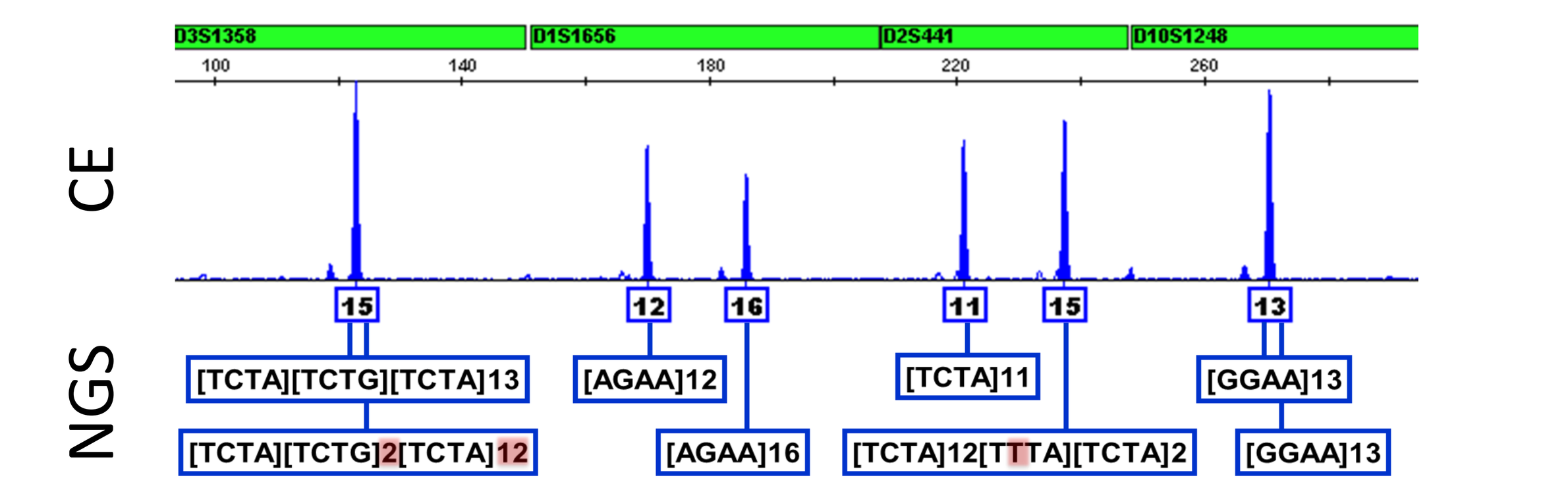


Figure 3. CE (PowerPlex Fusion, four loci shown) electropherogram for one sample of the 183 tested. Below the CE-based allele designations are the sequences obtained. D3S1358 is homozygous by length but heterozygous by sequence, D1S1656 is a simple repeat heterozygote, D2S441 has a simple repeat 11 allele and a different motif caused by a C→T SNP at the 15 allele, and D10S1248 is a simple repeat homozygote.

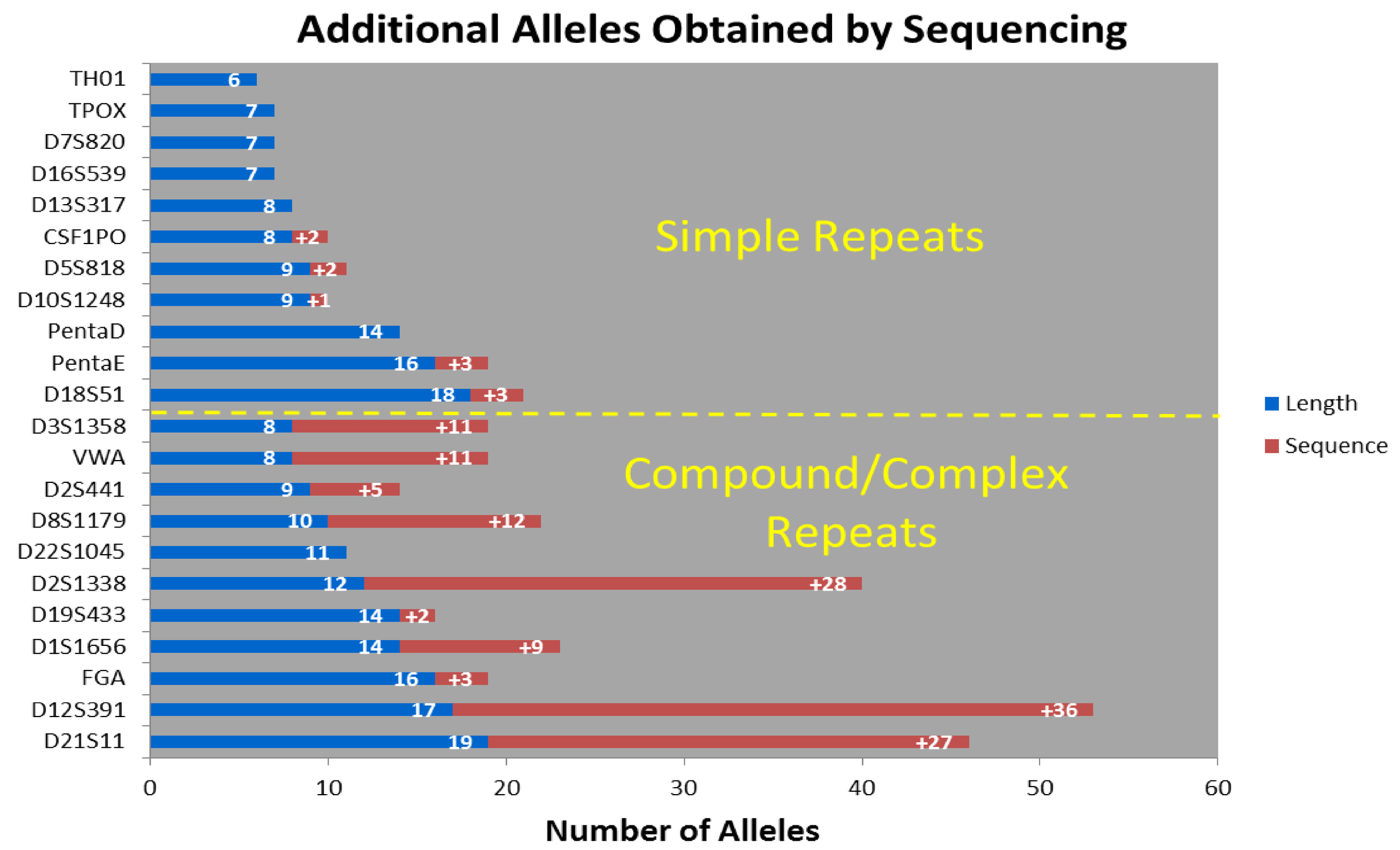


Figure 4. In blue are the number of different length-based alleles observed in this dataset (N=183), and in red are the number of additional sequenced-based alleles observed. Loci are grouped by repeat motif type (simple vs compound/complex) and sorted within each group by number of length-based alleles, smallest to largest.

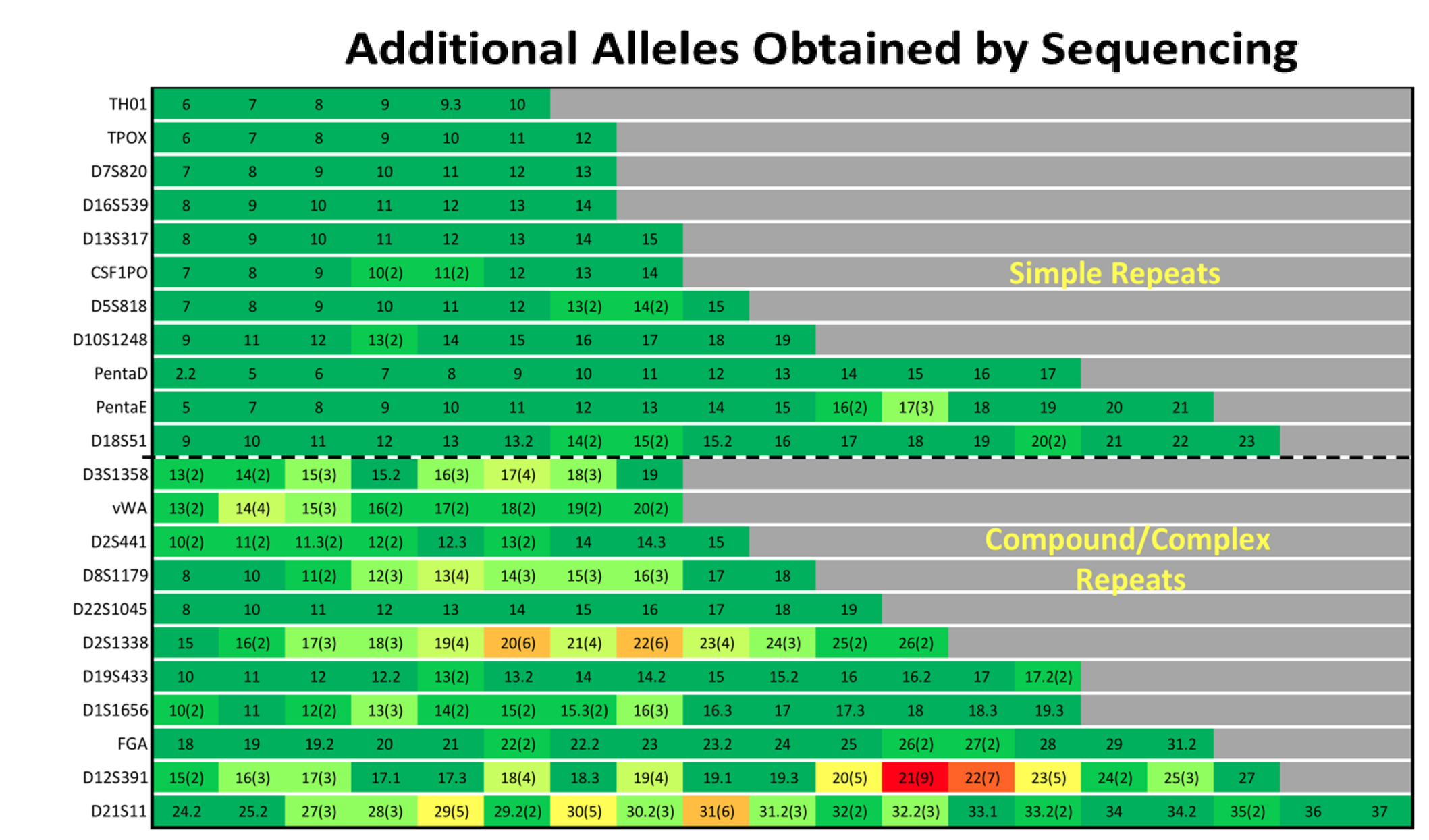
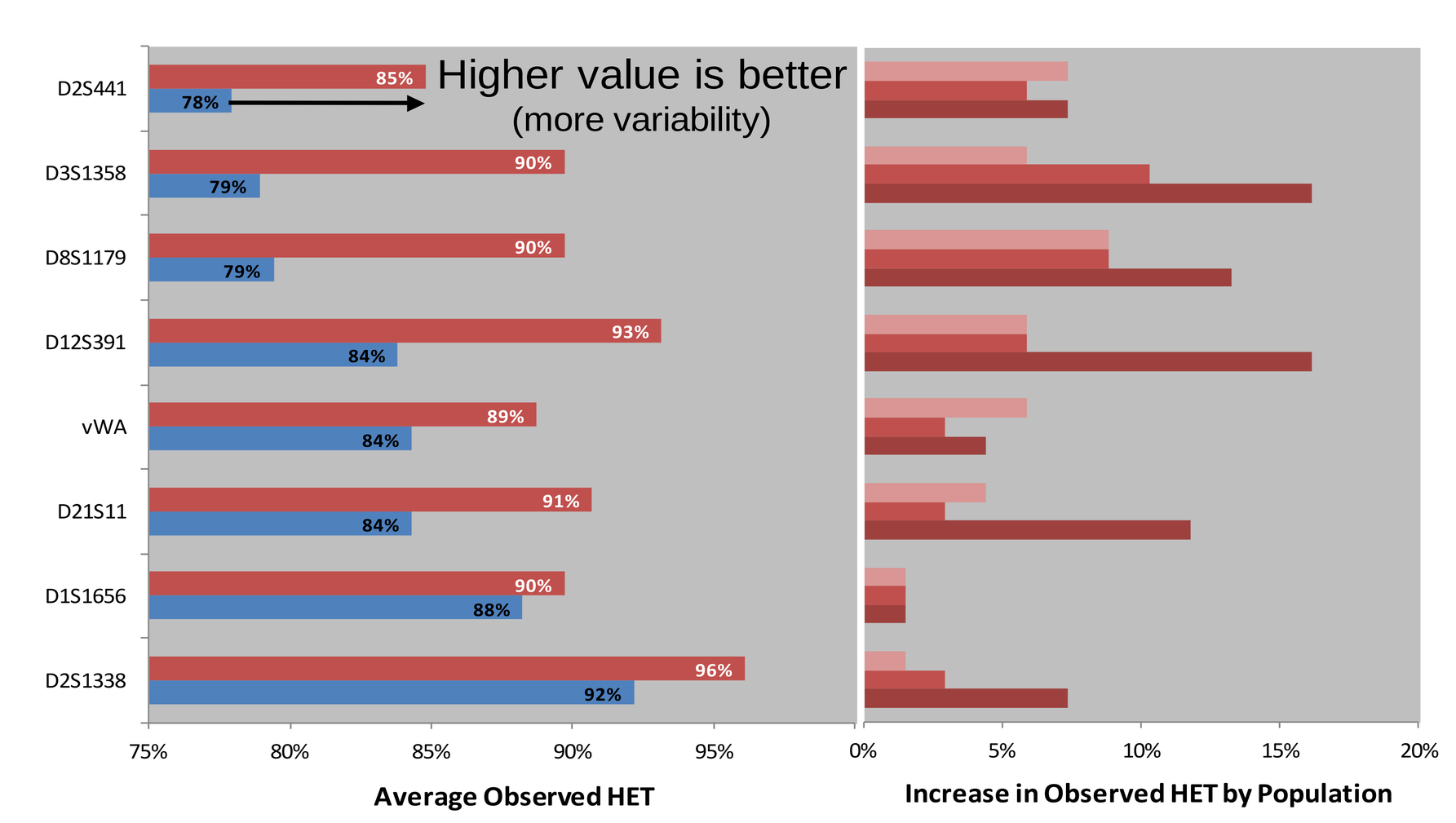
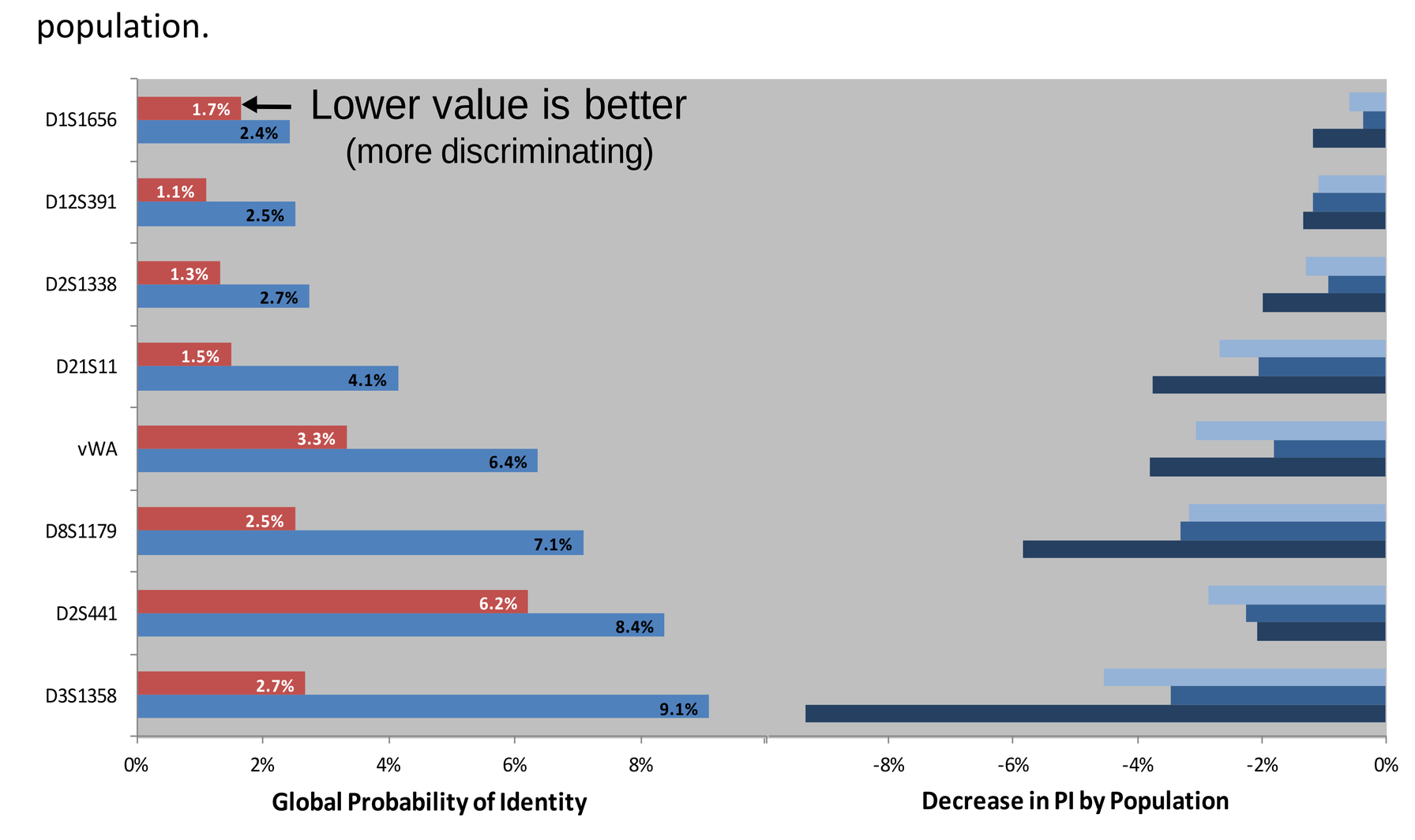


Figure 5. Heatmap showing the variant count for each allele (counts in parentheses, alleles with no parenthetical notation show no sequence variants in this dataset). Alleles are color coded with the darkest green shading representing no sequence variation; shading changes to yellow-orange-red with increasing sequence variation.



Definition: Heterozygosity
The number of heterozygous loci divided by total number of loci; a measure of allelic variability at a locus.

Figure 6. (left) Percent heterozygosity by length (blue) and by sequence (red), averaged from three populations (N=183) rank ordered by HET length, top eight loci shown. (right) The increase in heterozygosity by sequence, broken out by population.



Definition: Probability of Identity
Chance of randomly choosing two unrelated individuals with the same genotype at a locus.

$$\sum_{i=1}^{n} x_i^2$$

Figure 7. (left) Percent probability of identity by length (blue) and by sequence (red), based on average allele frequency from these three populations (N=183) rank ordered by PI length, top eight loci shown. (right) The decrease in probability of identity by sequence, broken out by population.

EXAMPLE LOCUS

The D21S11 locus provides an example of the sequence complexity that exists and the need for a system of nomenclature that is meaningful and expandable as more unique alleles are sequenced. For the set of population samples sequenced in this project (N=183), at the D21S11 locus, 16 sequences were found which had not previously been reported (equaling 4.4% of chromosome 21 sequences), see Table 1 rows in red.

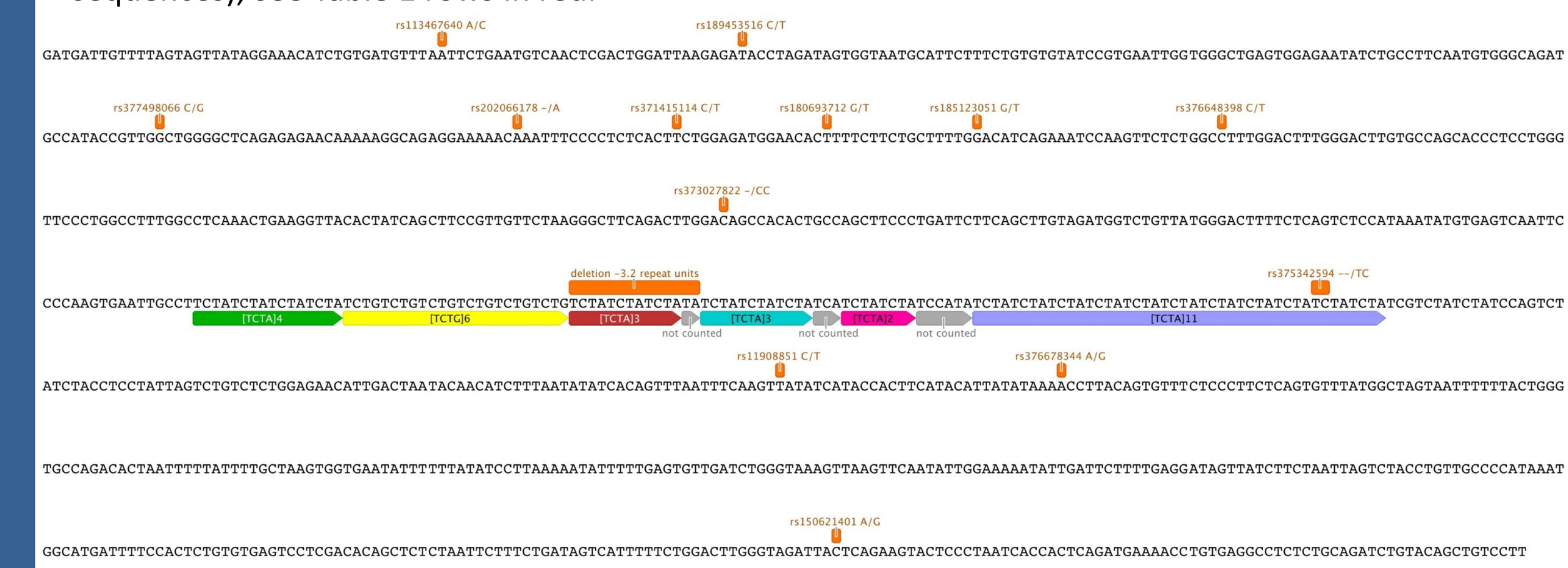


Figure 8. GRCh38 sequence (GenBank) at the D21S11 locus (repeat region and 500 bases upstream and downstream), annotated with information relevant to forensic NGS. The subunits of the repeat motif are shown in green, yellow, red, aqua, pink and purple (these regions are "counted", resulting in a 29 allele), while the interspersed non-repeat sequences are shown in gray (these regions are "not counted"). Two different regions where deletions result in .2 alleles, as well as the locations of numerous SNPs (from dbSNP) that have been observed in the flanking regions are shown in orange. Annotation created with Geneious v7.1.7.

D21S11 [TCTA]n [TCTG]n [TCTA]n TA [TCTA]n TCA [TCTA]n TCCATA [TCTA]n

allele	[TCTA]4-13	[TCTG]3-11	[TCTA]3 TA [TCTA]2-3 TCA [TCTA]2 TCCATA	[TCTA]6-15			Platform	Reference
24	[TCTA]4	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]6			Sanger	Griffiths *et al.* (1998)
25	[TCTA]4	[TCTG]3	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]10			Sanger	Schwartz *et al.* (1996)
26	[TCTA]4	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]8			Sanger	Möller *et al.* (1994)
26	[TCTA]6	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]7			Sanger	Wang *et al.* (2014)
27	[TCTA]4	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]9			Sanger	Möller *et al.* (1994)
27	[TCTA]5	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]9			Sanger	Griffiths *et al.* (1998)
27	[TCTA]6	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]8			Sanger	Schwartz *et al.* (1996)
28	[TCTA]4	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11			454	Gelardi et al. (2014)
28	[TCTA]4	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]10			Sanger	Möller *et al.* (1994)
28	[TCTA]5	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]10			Sanger	Wang et al. (2014)
28	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]2 TCA [TCTA]2 TCCATA	[TCTA]10			MiSeq	NIST 183
28	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]9			Sanger	Zhou *et al.* (1997)
28	[TCTA]6	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]9			454	Gelardi et al. (2014)
29	[TCTA]4	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11			Sanger	Griffiths *et al.* (1998)
29	[TCTA]4	[TCTG]7	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]10			Sanger	Planz *et al.* (2012)
29	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]2 TCA [TCTA]2 TCCATA	[TCTA]11			MiSeq	NIST 183
29	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]10			454	Gelardi *et al.* (2014)
29	[TCTA]6	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]10			Sanger	Zhou *et al.* (1997)
30	[TCTA]4	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12			Sanger	Schwartz *et al.* (1996)
30	[TCTA]4	[TCTG]7	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11			MiSeq	NIST 183
30	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11			Sanger	Zhou *et al.* (1997)
30	[TCTA]6	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11			Sanger	Griffiths (1998)
30	[TCTA]6	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]10			Sanger	Brinkmann *et al.* (1996a)
30	[TCTA]7	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]10			454	Gelardi *et al.* (2014)
31	[TCTA]4	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]13			454	Gelardi *et al.* (2014)
31	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12			Sanger	Griffiths *et al.* (1998)
31	[TCTA]6	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12			Sanger	Möller *et al.* (1994)
31	[TCTA]6	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11			Sanger	Zhou *et al.* (1997)
31	[TCTA]7	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11			Sanger	Schwartz *et al.* (1996)
31	[TCTA]7	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]10			MiSeq	NIST 183
31	[TCTA]4	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]13			Sanger	Wang et al. (2014)
32	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]13			Sanger	Zhou *et al.* (1997)
32	[TCTA]6	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]13			Sanger	Griffiths *et al.* (1998)
32	[TCTA]6	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12			MiSeq	NIST 183
32	[TCTA]8	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11			454	Gelardi *et al.* (2014)
33	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]14			Sanger	Zhou *et al.* (1997)
33	[TCTA]6	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]14			454	Gelardi *et al.* (2014)
33	[TCTA]9	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11			Sanger	Wang *et al.* (2014)
34	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]15			Sanger	Zhou *et al.* (1997)
34	[TCTA]8	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]13			Sanger	Wang *et al.* (2014)
34	[TCTA]10	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11			Sanger	Brinkmann *et al.* (1996a)
35	[TCTA]9	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11			Sanger	Wang *et al.* (2014)
35	[TCTA]10	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12			Sanger	Griffiths *et al.* (1998)
35	[TCTA]10	[TCTG]7	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]10			MiSeq	NIST 183
35	[TCTA]11	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11			Sanger	Brinkmann *et al.* (1996a)
36	[TCTA]9	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]14			Sanger	Wang *et al.* (2014)
36	[TCTA]10	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]13			Sanger	Brinkmann *et al.* (1996a)
36	[TCTA]10	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12			Sanger	Brinkmann *et al.* (1996a)
36	[TCTA]10	[TCTG]7	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11			MiSeq	NIST 183
36	[TCTA]11	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12			Sanger	Griffiths *et al.* (1998)
37	[TCTA]9	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]15			Sanger	Wang *et al.* (2014)
37	[TCTA]11	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]13			Sanger	Griffiths *et al.* (1998)
37	[TCTA]11	[TCTG]8	[TCTA]2 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11			MiSeq	NIST 183
38	[TCTA]13	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12			Sanger	Griffiths *et al.* (1998)

allele	[TCTA]4-6	[TCTG]5-6	[TCTA]2-3 TA [TCTA]2-3 TCA [TCTA]2 TCCATA	[TCTA]8-16	TA	[TCTA]	Platform	Reference
28.2	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]8	TA	TCTA	Sanger	Zhou *et al.* (1997)
29.2	[TCTA]5	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]10	TA	TCTA	Sanger	Zhou *et al.* (1997)
29.2	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]9	TA	TCTA	MiSeq	NIST 183
30.2	[TCTA]5	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11	TA	TCTA	Sanger	Schwartz *et al.* (1996)
30.2	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]2 TCA [TCTA]2 TCCATA	[TCTA]11	TA	TCTA	MiSeq	NIST 183
30.2	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]10	TA	TCTA	Sanger	Griffiths *et al.* (1998)
31.2	[TCTA]5	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12	TA	TCTA	MiSeq	NIST 183
31.2	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]2 TCA [TCTA]2 TCCATA	[TCTA]12	TA	TCTA	MiSeq	NIST 183
31.2	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11	TA	TCTA	Sanger	Griffiths *et al.* (1998)
32.2	[TCTA]4	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]13	TA	TCTA	Sanger	Brinkmann *et al.* (1996a)
32.2	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12	TA	TCTA	Sanger	Griffiths *et al.* (1998)
32.2	[TCTA]5	[TCTG]6	[TCTA]2 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]13	TA	TCTA	Sanger	Brinkmann *et al.* (1996a)
32.2	[TCTA]6	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12	TA	TCTA	MiSeq	NIST 183
32.2	[TCTA]6	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11	TA	TCTA	454	Gelardi *et al.* (2014)
33.2	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]13	TA	TCTA	Sanger	Griffiths *et al.* (1998)
33.2	[TCTA]6	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]13	TA	TCTA	Sanger	Brinkmann *et al.* (1996a)
33.2	[TCTA]6	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12	TA	TCTA	Sanger	Brinkmann *et al.* (1996a)
34.2	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]14	TA	TCTA	Sanger	Griffiths *et al.* (1998)
35.2	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]15	TA	TCTA	Sanger	Zhou *et al.* (1997)
36.2	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]16	TA	TCTA	Sanger	Zhou *et al.* (1997)

allele	[TCTA]5-11	[TCTG]6-14	[----] -- [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]9-13			Platform	Reference
24.2	[TCTA]5	[TCTG]6	[----] -- [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]9			Sanger	Griffiths *et al.* (1998)
25.2	[TCTA]5	[TCTG]6	[----] -- [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]10			Sanger	Griffiths *et al.* (1998)
25.2	[TCTA]6	[TCTG]6	[----] -- [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]9			MiSeq	NIST 183
37.2	[TCTA]7	[TCTG]14	[----] -- [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12			Sanger	Walsh *et al.* (2003)
37.2	[TCTA]9	[TCTG]12	[----] -- [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12			Sanger	Walsh *et al.* (2003)
38.2	[TCTA]9	[TCTG]12	[----] -- [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]13			Sanger	Walsh *et al.* (2003)
37.2	[TCTA]9	[TCTG]13	[----] -- [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11			Sanger	Walsh *et al.* (2003)
38.2	[TCTA]9	[TCTG]13	[----] -- [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12			Sanger	Walsh *et al.* (2003)
37.2	[TCTA]10	[TCTG]11	[----] -- [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12			Sanger	Walsh *et al.* (2003)
38.2	[TCTA]10	[TCTG]11	[----] -- [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]13			Sanger	Walsh *et al.* (2003)
39.2	[TCTA]10	[TCTG]13	[----] -- [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12			Sanger	Bagdonavicius *et al.* (2002)
41.2	[TCTA]10	[TCTG]14	[----] -- [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12			Sanger	Bagdonavicius *et al.* (2002)
37.2	[TCTA]11	[TCTG]11	[----] -- [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]11			Sanger	Walsh *et al.* (2003)
39.2	[TCTA]11	[TCTG]12	[----] -- [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]12			Sanger	Bagdonavicius *et al.* (2002)

allele	[TCTA]5-6	[TCTG]5-6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]5-6	TCA	[TCTA]6	Platform	Reference
30.3	[TCTA]6	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]5	TCA	[TCTA]6	Sanger	Tsuji *et al.* (2006)
31.3	[TCTA]7	[TCTG]5	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]5	TCA	[TCTA]6	Sanger	Wang *et al.* (2014)
33.1	[TCTA]5	[TCTG]6	[TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCATA	[TCTA]6	TCA	[TCTA]6 TA TCTA	MiSeq	NIST 183

Table 1. Non-exhaustive list of unique allele sequences at D21S11, obtained from either Appendix 1 of *Advanced Topics in Forensic DNA Typing: Methodology* [2] (rows in white, see book for complete references), from more recent literature [3-5](rows in blue), or from this project (rows in red), grouped by repeat motif (motifs and subunit size ranges shown in dark gray). This table contains 91 unique sequences for 33 length-based genotypes. Then number of unique sequences will continue to grow as more samples are sequenced. Top row: consensus motif at D21S11, color coded to match Figure 8.

NOMENCLATURE

STR genotypes based on sequence data should maintain back-compatibility with length-based genotyping. The original guidance document for length-based genotypes [6] is paraphrased here:

Guidelines for STR sequence and repeat designations

1. In protein coding genes, pseudogenes, and introns, the coding strand should be reported
2. For loci that are not known to be coding, the sequence originally described in the literature is the standard reference and strand for nomenclature
3. If forensic nomenclature is already established but is not in accordance with the aforementioned guidelines, the existing nomenclature should be maintained
4. The first 5' nucleotides that form a motif are used to define the repeat sequence motif
5. Allele designations should observe these structural principles:

Simple repeats are straightforward, the number of repeat units are counted	CSF1PO	$[AGAT]_{10}$ = 10 allele
In compound repeats, alleles are designated by counting the total number of full repeats	vWA	$[TCTA][TCTG]_{4}[TCTA]_{13}$ = 18 allele
Microvariant alleles are designated by counting the number of full repeats, adding a decimal point, and then counting the number of basepairs in the incomplete repeat	TH01	$[AATG]_{6}$ A-TG$[AATG]_{3}$ = 9.3 allele
Complex repeat systems should have a mathematical relationship to the bp length of a consensus allele	D21S11 consensus	$[TCTA]_4[TCTG]_6[TCTA]_3$TA$[TCTA]_3$TCA$[TCTA]_2$TCCATA$[TCTA]_9$ $\frac{119\text{ bp} - 11\text{ bp}}{4}$ = 27 allele*
For more highly variable systems, alleles should be identified according to their size in bp, compared to a sequenced ladder	D21S11 deletion in "constant" region	$[TCTA]_4[TCTG]_6$[- - - -]$_3$ - - $[TCTA]_3$TCA$[TCTA]_2$TCCATA$[TCTA]_9$ $\frac{105\text{ bp} - 11\text{ bp}}{4}$ = 23.2 allele*

*Note: Designation of D21S11 alleles has changed since the guidelines [6] were published, the above example reflects current nomenclature, where bases shown in gray are "not counted" toward the allele designation.

NGS Nomenclature Considerations and Possibilities

Obtaining full sequence data at forensic STR loci will be within reach of forensic casework and databasing laboratories in the near future. How this information will be used for comparisons and database searching, reported to investigators, and stored, are questions that will need to be addressed prior to implementation. **Three options for representing the sequence data, and their possible applications, are outlined below.**

(1) Complete Sequence String The entire string of generated sequence, including all flanking regions:

GACAGAATTGCACCAAATATTGGTAATTAAATGTTTACTATAGACTATTTAGTGAGATAAAAAAAAACTATC
AATCTGTCTATCTATCTATCTATCTATCTATCTATCTATCGTTAGTTCGTTCTAAACTATGACAAGTGTTCT
ATCATACCCTTTATATATATTAACCTTAAAATAACTCCATAGTCA

(2) Bracketed sequence This format is described in detail in the recent article by Gelardi, *et al.* [4]. It primarily consists of the repeat region, with repetitive elements enclosed in brackets and a numeric representation of the repeat length (as seen in Table 1 and in [2-6]). Additionally, polymorphisms (SNPs or InDels) in the flanking regions should be identified by their "rs" number (the dbSNP ID number, e.g. rs206437, see Figure 8). These rs numbers correspond to specific locations in the GenBank human genome assembly (current version GRCh38), and therefore eliminate the ambiguity that could result from lab-originated designations such as "upstream 13 bp C→T". If an rs number is not present in dbSNP for a particular flank polymorphism, the data can be submitted to NCBI for rs number assignment; however, some lab-specific designation may be needed in the interim.

(3) Unique Identifier A designation for each allele, either numeric (representing the repeat length) with an additional sequence-specific designator and flank polymorphism designator (e.g. "13d rs206437C" where 13 is the repeat length, d is the sequence version, and the rs number is a flank polymorphism), or a computer-generated code that is applied to each unique sequence string within a defined region (e.g. "@j*5").

Reporting/Manual Comparisons Any or all of the above options could be used to report STR region sequences. If the unique identifier can be readily "decoded" by a human, this may be helpful for quick comparisons. The bracketed sequence is intuitive and may help in explaining results to the investigator. The complete sequence could be appended to the report.

Database Searching The ideal nomenclature for database searching is unambigous and computationally inexpensive (i.e. fast). The two most likely possibilities are the computer-generated unique identifier and the complete sequence string.

Data Storage The string of nucleotides from the complete region sequenced, as well as the corresponding quality scores (as reported in the .fastq files that are automatically generated in an NGS run) will need to be maintained for re-analysis/possible future analyses. In this project, sequencing 24 loci at approximately 1000x coverage per locus on the MiSeq generates .fastq files of approximately 50 MB per sample. Based on these metrics, .fastq files from 20,000 samples could be stored on a 1 TB drive (current cost < $100). The .fastq files are generated by the instrument itself, using signal detection algorithms that have been optimized by the manufacturer. If it is possible for the NGS user to make changes to these algorithms for a particular NGS platform, then all files generated from the sequencer should be maintained. Within this project, maintaining all files would equal approximately 200 MB per sample, meaning approximately 5,000 samples could be stored on a 1 TB drive. It should be noted that larger storage systems are less expensive per unit of data.

Summary/Conclusions Sequencing more population samples at forensic STR loci will help guide nomenclature decisions. As was the case with mtDNA, phylogenetic approaches may be useful. In addition, the informatic approaches available/created for genotyping and databasing of sequences is an important factor. These issues should be addressed by the global forensic community, and we encourage an open dialogue among forensic experts in forums such as an ISFG subcommittee, the SWGDAM NGS Working Group, and a NIST-OSAC subcommittee.

Acknowledgements

The authors thank Doug Storts (Promega) and Jay Patel (Promega) for amplification primers and advice, Becky Hill (NIST) for CE genotype data, Mike Coble (NIST) for population genetic analysis guidance, and John Butler (NIST) for comments/feedback.

References

[1] Warshauer, D. H., D. Lin, K. Hari, R. Jain, C. Davis *et al.*, 2013 STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. Forensic Sci Int Genet 7**:** 409-417.

[2] Butler, J. M., 2012 *Advanced Topics in Forensic DNA Typing: Methodology*. Elsevier, Inc., Waltham, MA, USA.

[3] Wang, L., X. C. Zhao, J. Ye, J. J. Liu, T. Chen *et al.*, 2014 Construction of a library of cloned short tandem repeat (STR) alleles as universal templates for allelic ladder preparation. Forensic Sci Int Genet 12**:** 136-143.

[4] Gelardi, C., E. Rockenbauer, S. Dalsgaard, C. Borsting and N. Morling, 2014 Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles. Forensic Sci Int Genet 12C**:** 38-41.

[5] Planz, J. V., K. A. Sannes-Lowery, D. D. Duncan, S. Manalili, B. Budowle *et al.*, 2012 Automated analysis of sequence polymorphism in STR alleles by PCR and direct electrospray ionization mass spectrometry. Forensic Sci Int Genet 6**:** 594-606.

[6] Bär, W., B. Brinkmann, B. Budowle, A. Carracedo, *et al.*, 1997 DNA Recommendations – further report of the DNA Commission of the ISFH regarding the use of short tandem repeat systems. Forensic Sci Int 87:179-184.