STRSEQ: A RESOURCE FOR SEQUENCE-BASED STR ANALYSIS

Katherine B. Gettings, Lisa A. Borsuk, and Peter M. Vallone

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

ABSTRACT

The STR Sequencing Project (STRSeq) was initiated to facilitate the description of sequence-based alleles at the Short Tandem Repeat (STR) loci targeted in human identification assays. STRSeq data are maintained as GenBank records at the U.S. National Center for Biotechnology Information (NCBI). Each GenBank record contains: observed sequence of an STR region, annotation of the repeat region ("bracketing" consistent with the guidance of the International Society for Forensic Genetics) and flanking region polymorphisms, information regarding the sequencing assay and data quality, and backward compatible length-based allelic designation. STRSeq GenBank records are organized within a BioProject at NCBI (www.ncbi.nlm.nih.gov/bioproject/380127), which is sub-divided by Commonly used autosomal STR Loci, Alternate autosomal STR Loci, Y-chromosomal STR loci, and X-chromosomal STR loci. Each of these categories is further divided into locus-specific projects. The BioProject will initially contain aggregate alleles across 4,612 samples submitted by four laboratories: National Institute of Standards and Technology (NIST, the project organizer), University of North Texas Health Sciences Center, Kings College London, and University of Santiago de Compostela. In addition to providing a framework for communication among laboratories, the ability to search the BioProject can be leveraged as QC for rare sequences encountered in forensic casework. Future plans for this NIJ-funded effort include a pathway for researchers to submit additional alleles and customized interface tools.

TECHNOLOGY TRANSITION

Length variations among individuals in short tandem repeat (STR) loci have been used in forensic applications since the 1990s, due to the ease with which these loci can be multiplexed combined with a high degree of heterogeneity. At some STR loci commonly used in human identification, sequence variation can provide an additional level of discrimination, as shown in figures 1 and 2.

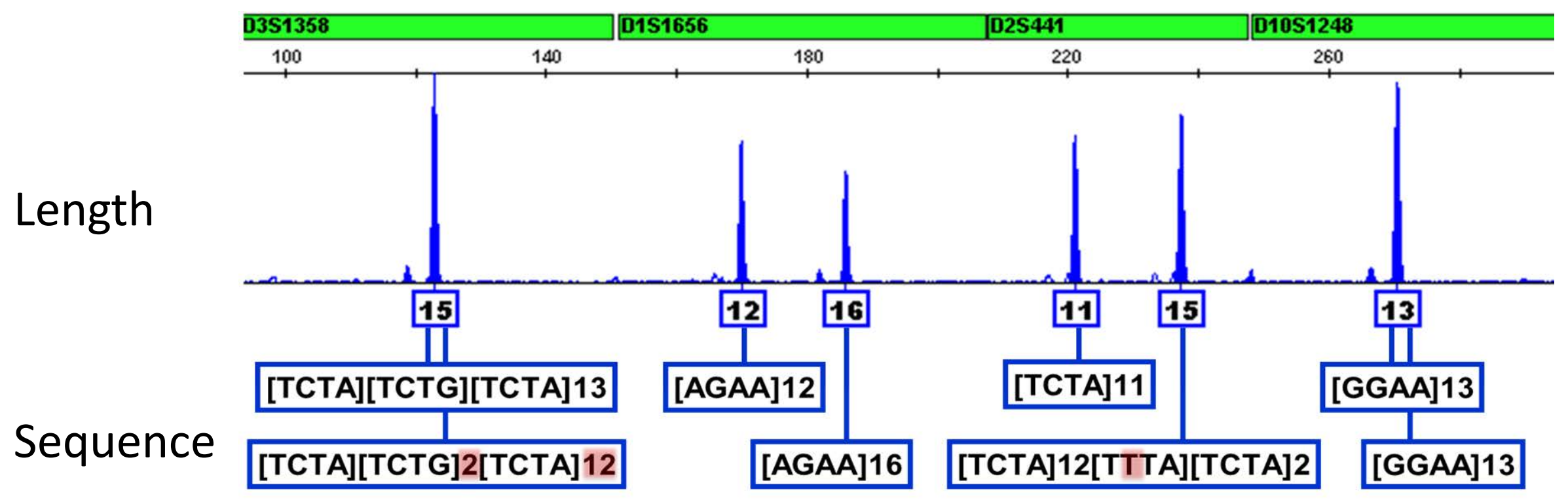


Figure 1. An example DNA profile at four STR loci used in human identification. The length-based profile is derived from peaks in an electropherogram. Peaks are given numerical assignments based on detection time compared to an allelic ladder. The numerical alleles correspond to counts of repetitive nucleotides (most often tetranucleotide repeats). DNA sequencing can be used to derive the same length-based profile, and sometimes provides additional discrimination of length-based alleles, as shown for the D3S1358 locus in this example.

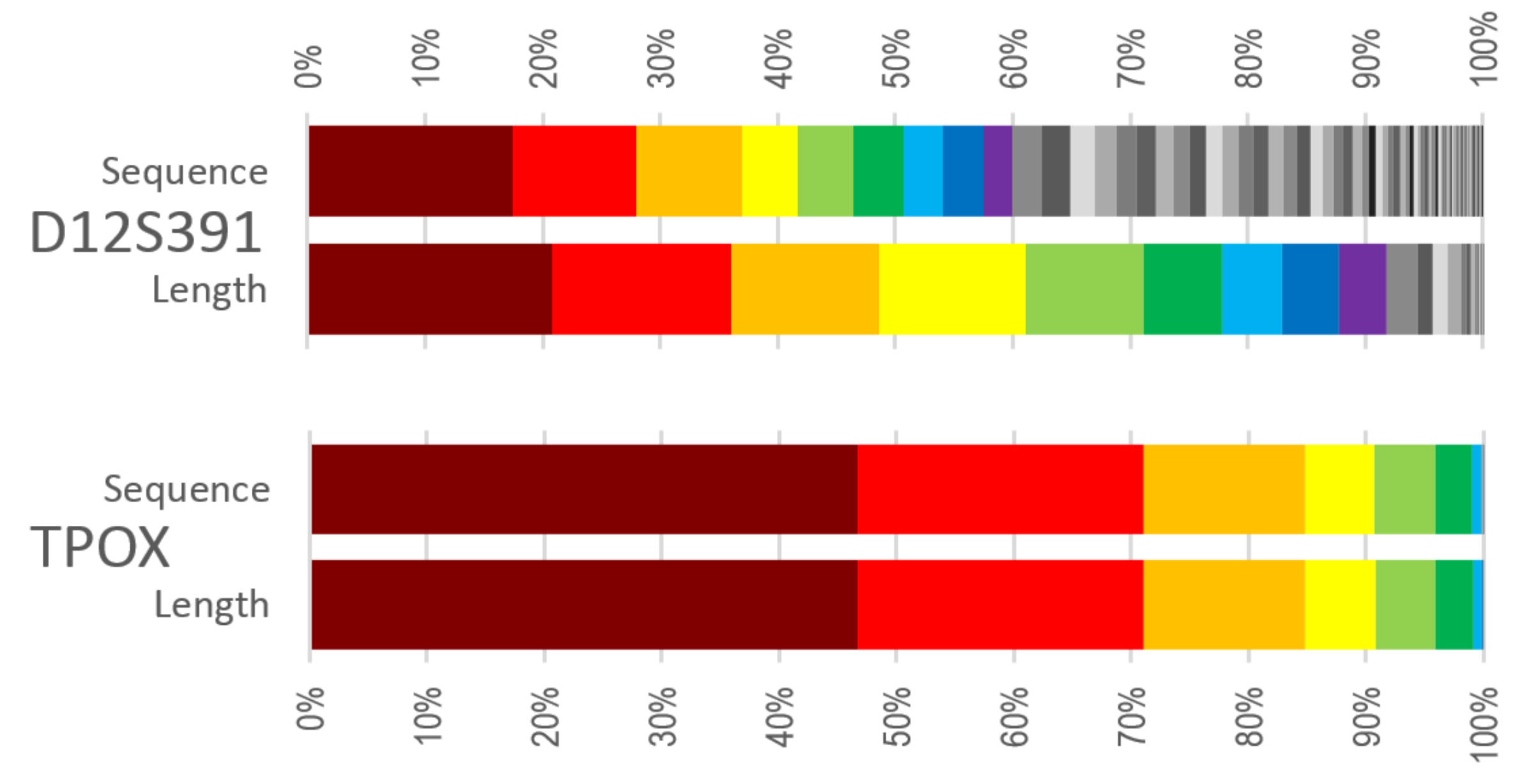


Figure 2. Comparison of allele frequency distribution of two loci by length and by sequence. The D12S391 locus is one of the more polymorphic loci by length used in human identification, and becomes further discriminating by sequence. The most common D12S391 allele by length is present at a frequency of 20.8% in the NIST population set of 1036 individuals, and this frequency drops to 17.5% by sequence, in addition to a proliferation of rare alleles shown in grayscale. The TPOX locus is far less polymorphic by length and does not improve by sequence.

As the forensic DNA community evaluates the potential of sequencing applications for Short Tandem Repeat (STR) loci, it is imperative to define the allelic diversity in these regions of the human genome, in order to standardize nomenclature for sequence-based STR alleles and ensure backward compatibility to traditional length-based alleles. The STRSeq BioProject [1] has been initiated to facilitate the description of sequence-based alleles at the STRs targeted in human identification assays. This resource consists of a curated catalog of sequence diversity at forensic STR loci, along with the key elements of nomenclature conforming to current guidelines [2], and will serve as the data backbone during this time of transition, as well as a stable resource for the future.

SAMPLES

The STRSeq Bioproject will initially contain aggregate alleles across 4,612 samples sequenced among the partner laboratories, as shown in figure 3. Importantly, sequence data for NIST population samples is compared to length-based (CE) data derived across multiple manufacturer assays and instruments, and a subset have now been sequenced with multiple assays. This cross validation lends confidence to the sequence data and evaluates back compatibility. Figures 4 and 5 demonstrate the alleles obtained across laboratories for one highly polymorphic locus, and the submission strategy.

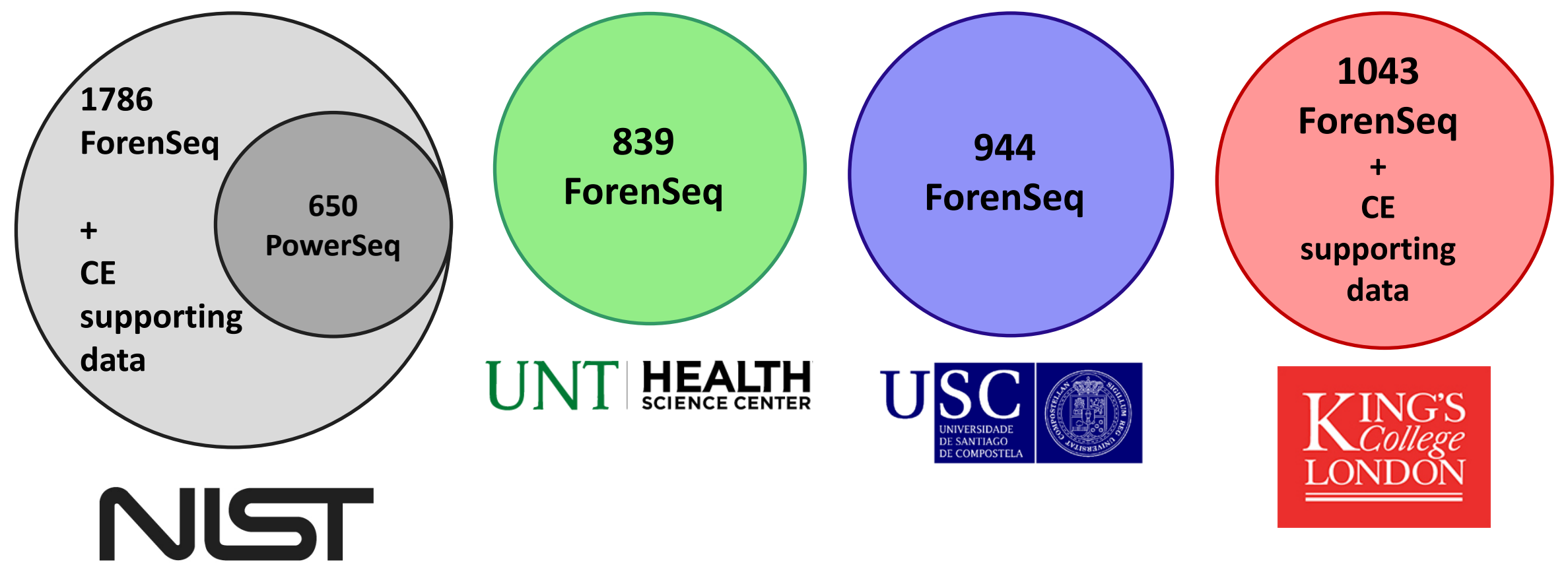


Figure 3. Sample submission by laboratory for the initial formation of the BioProject. NIST population samples have all been sequenced with the Illumina ForenSeq assay (MiSeq FGx), while a subset have been sequenced with the Promega PowerSeq Auto-Y assay (MiSeq) and are expected to be sequenced with ThermoFisher GlobalFiler NGS (S5) in 2018. The other three laboratories are submitting population data sequenced with Illumina ForenSeq. Varying levels of length-based (CE) supporting data exist across the laboratories.

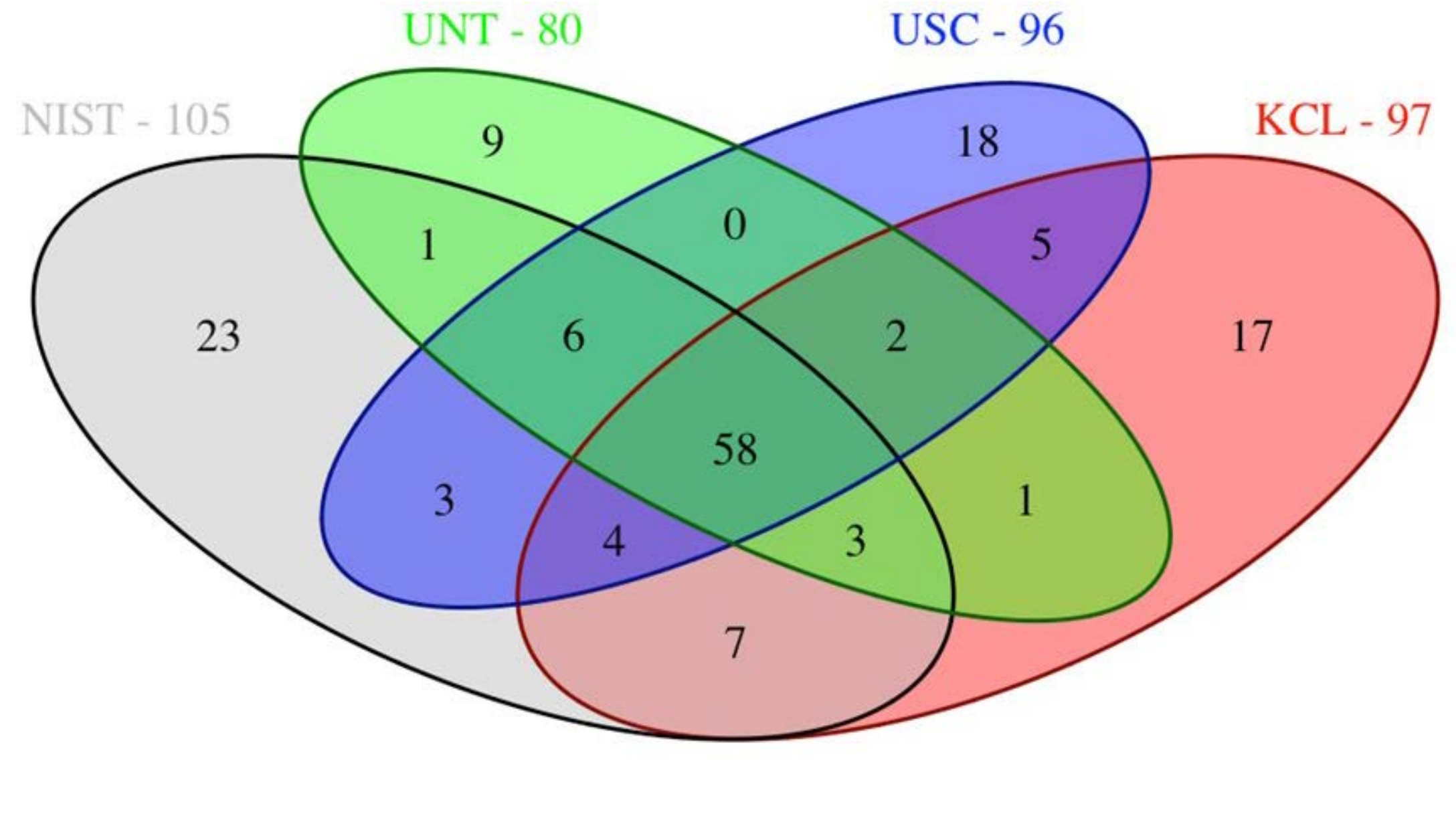
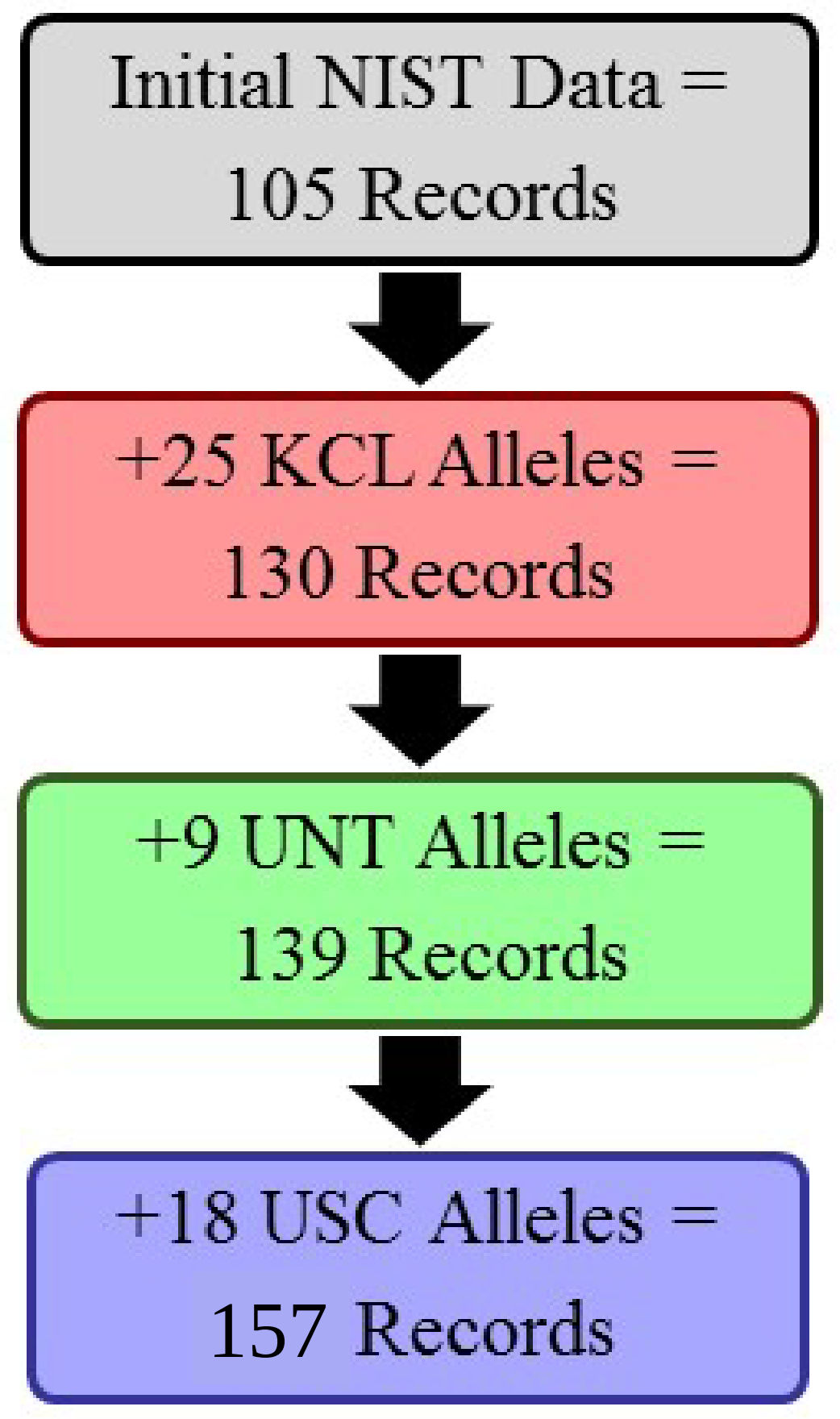


Figure 4. Venn diagram demonstrating the overlap of D12S391 sequence-based alleles observed among the four laboratories, and the total number of unique sequence-based alleles observed within each laboratory. STRSeq GenBank records will be created for each of the 157 unique sequence-based alleles at D12S391.



Initial NIST Data = 105 Records

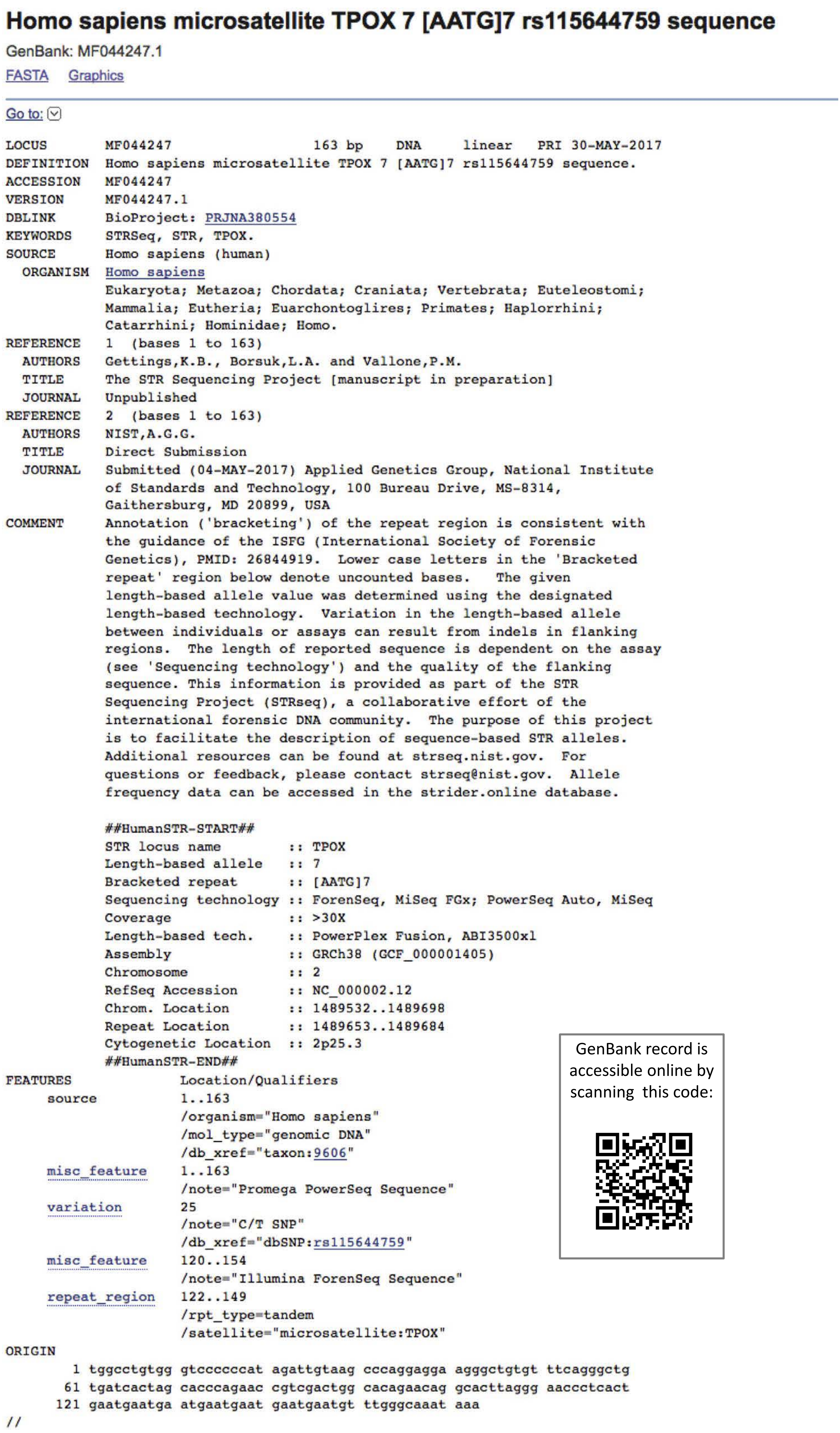
+25 KCL Alleles = 130 Records

+9 UNT Alleles = 139 Records

+18 USC Alleles = 157 Records

Figure 5. Submission strategy for 157 unique sequence-based alleles observed at the D12S391 locus. The 105 unique alleles generated at NIST form the basis of STRSeq records. Subsequent submissions from KCL, UNT, and USC will add records for sequences generated at each laboratory for which records do not already exist (25, 9, and 18 records, respectively).

EXAMPLE RECORD



Homo sapiens microsatellite TPOX 7 [AATG]7 rs115644759 sequence

GenBank: MF044247.1

FASTA Graphics

Go to:

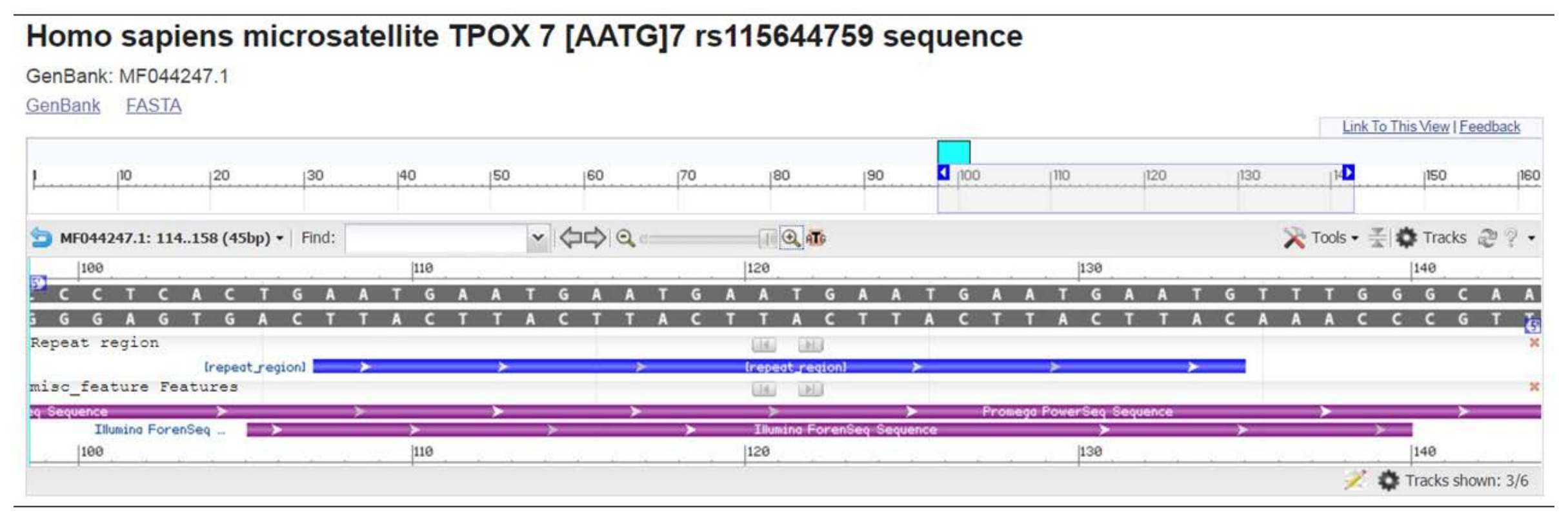
```
LOCUS       MF044247                 163 bp    DNA     linear   PRI 30-MAY-2017
DEFINITION  Homo sapiens microsatellite TPOX 7 [AATG]7 rs115644759 sequence.
ACCESSION   MF044247
VERSION     MF044247.1
DBLINK      BioProject: PRJNA380554
KEYWORDS    STRSeq, STR, TPOX.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 163)
  AUTHORS   Gettings,K.B., Borsuk,L.A. and Vallone,P.M.
  TITLE     The STR Sequencing Project [manuscript in preparation]
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 163)
  AUTHORS   NIST,A.G.G.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-MAY-2017) Applied Genetics Group, National Institute
            of Standards and Technology, 100 Bureau Drive, MS-8314,
            Gaithersburg, MD 20899, USA
COMMENT     Annotation ('bracketing') of the repeat region is consistent with
            the guidance of the ISFG (International Society of Forensic
            Genetics), PMID: 26844919.  Lower case letters in the 'Bracketed
            repeat' region below denote uncounted bases.   The given
            length-based allele value was determined using the designated
            length-based technology.  Variation in the length-based allele
            between individuals or assays can result from indels in flanking
            regions.  The length of reported sequence is dependent on the assay
            (see 'Sequencing technology') and the quality of the flanking
            sequence. This information is provided as part of the STR
            Sequencing Project (STRseq), a collaborative effort of the
            international forensic DNA community.  The purpose of this project
            is to facilitate the description of sequence-based STR alleles.
            Additional resources can be found at strseq.nist.gov.  For
            questions or feedback, please contact strseq@nist.gov.  Allele
            frequency data can be accessed in the strider.online database.

            ##HumanSTR-START##
            STR locus name          :: TPOX
            Length-based allele     :: 7
            Bracketed repeat        :: [AATG]7
            Sequencing technology :: ForenSeq, MiSeq FGx; PowerSeq Auto, MiSeq
            Coverage                :: >30X
            Length-based tech.      :: PowerPlex Fusion, ABI3500xl
            Assembly                :: GRCh38 (GCF_000001405)
            Chromosome              :: 2
            RefSeq Accession        :: NC_000002.12
            Chrom. Location         :: 1489532..1489698
            Repeat Location         :: 1489653..1489684
            Cytogenetic Location    :: 2p25.3
            ##HumanSTR-END##
FEATURES             Location/Qualifiers
     source          1..163
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
     misc_feature    1..163
                     /note="Promega PowerSeq Sequence"
     variation       25
                     /note="C/T SNP"
                     /db_xref="dbSNP:rs115644759"
     misc_feature    120..154
                     /note="Illumina ForenSeq Sequence"
     repeat_region   122..149
                     /rpt_type=tandem
                     /satellite="microsatellite:TPOX"
ORIGIN
        1 tggcctgtgg gtccccccat agattgtaag cccaggagga agggctgtgt ttcagggctg
       61 tgatcactag cacccagaac cgtcgactgg cacagaacag gcacttaggg aaccctcact
      121 gaatgaatga atgaatgaat gaatgaatgt ttgggcaaat aaa
//
```

GenBank record is accessible online by scanning this code:

Figure 6. An example STRSeq BioProject GenBank record for the TPOX locus, available online at https://www.ncbi.nlm.nih.gov/nuccore/1197990967. The custom comment section, bounded by ##HumanSTR-START## ... ##HumanSTR-END##, contains information of particular relevance to the forensic community such as the properly formatted bracketed repeat, quality information, and orientation on the human genome reference sequence. The FEATURES table toward the bottom of the record indicates the portion of the sequence was obtained from each assay, and the overlap therein.

EXAMPLE GRAPHICAL VIEW



Homo sapiens microsatellite TPOX 7 [AATG]7 rs115644759 sequence

GenBank: MF044247.1

GenBank FASTA

MF044247.1: 114..158 (45bp) Find:

Link To This View | Feedback
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
Tools
Tracks
C C T C A C T G A A T G A A T G A A T G A A T G A A T G A A T G A A T G T T T G G G C A A
G G A G T G A C T T A C T T A C T T A C T T A C T T A C T T A C T T A C A A A C C C G T T
Repeat region
[repeat_region]
misc_feature Features
Sequence
Promega PowerSeq Sequence
Illumina ForenSeq ...
Illumina ForenSeq Sequence
Tracks shown: 3/6

Figure 7. Example Graphics view of STRSeq Genbank record, available and interactive online at https://www.ncbi.nlm.nih.gov/nuccore/1197990967?report=graph.By selecting the Graphics link at the top of the GenBank record, users can access this graphical sequence view of the record, which shows the defined repeat region in the context of the reported sequence result from the assay(s) used. Polymorphisms in the flanking regions are also displayed, when present. This information will aide users understanding of the presence/absence of flanking polymorphisms across assays.

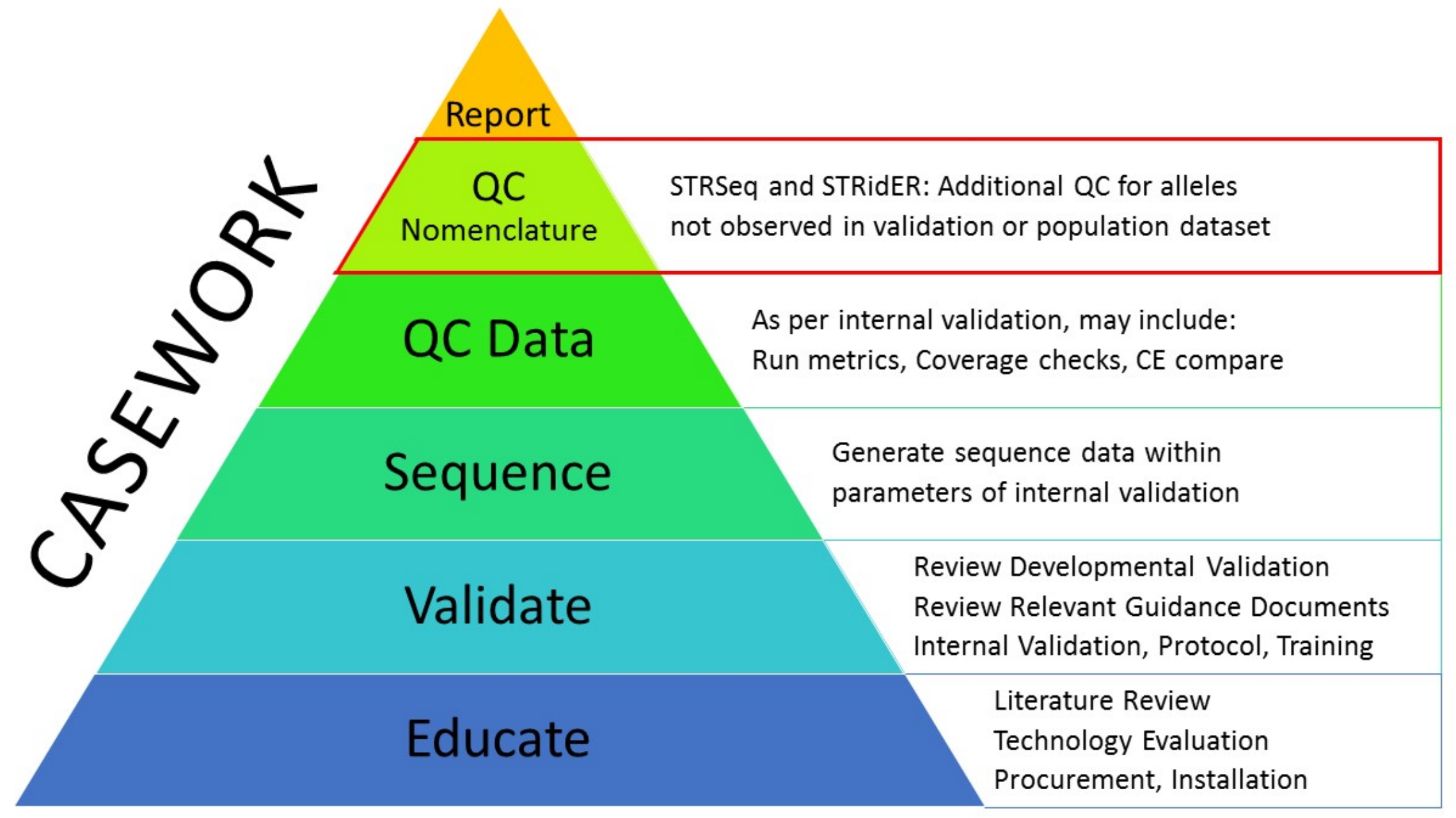
USE CASES

Several use cases for STRSeq have been identified based on feedback from the forensic community:

I. As a teaching tool to explore STR sequences. The STRSeq BioProject is expected to be useful to forensic operational, academic, and commercial laboratories interested in sequencing STRs as it allows the viewing and downloading of repeat region motifs, flanking region polymorphisms, and commercial assay overlap.

II. As the data backbone for software development. This catalog of sequences with associated forensic formatting and stable links to GenBank records facilitates development of STR sequencing methods and bioinformatic pipelines that conform to agreed variant data frameworks.

III. To provide a quality control function for the evaluation of rare sequences. When a sequence is observed in forensic casework that was not observed in initial validation studies or in the implemented allele frequency database, a STRSeq BLAST search determines if a similar or identical sequence has been recorded. When a link to previous data is identified, STRSeq provides nomenclature information and leads the analyst to published allele frequency data (see Figure 8).



CASEWORK

Report

QC Nomenclature — STRSeq and STRidER: Additional QC for alleles not observed in validation or population dataset

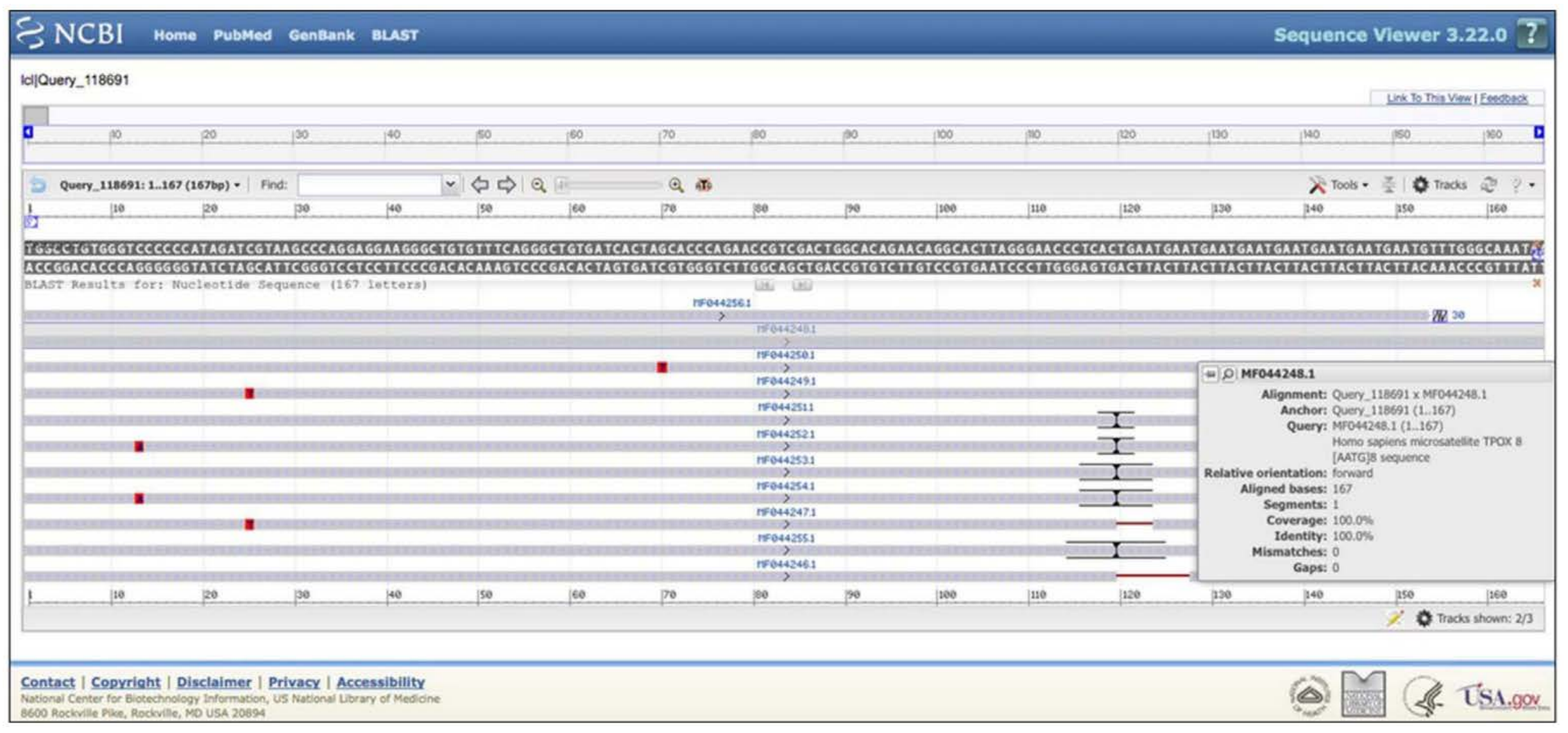
QC Data — As per internal validation, may include: Run metrics, Coverage checks, CE compare

Sequence — Generate sequence data within parameters of internal validation

Validate — Review Developmental Validation; Review Relevant Guidance Documents; Internal Validation, Protocol, Training

Educate — Literature Review; Technology Evaluation; Procurement, Installation

Figure 8. Outline of the anticipated STRSeq use case for evaluation of rare alleles in forensic casework, integrated into an overall quality assurance system. When a sequence variant is observed in casework which has not been previously observed, searching the STRSeq BioProject may provide the user with nomenclature information and lead the user to published allele frequency data through the connection which will be established between STRSeq and STRidER (an online database of allele frequencies maintained by the Institute of Legal Medicine, Medical University of Innsbruck, strider.online).



lcl|Query_118691

Query_118691: 1..167 (167bp) Find:

BLAST Results for: Nucleotide Sequence (167 letters)

MF044248.1

Alignment: Query_118691 x MF044248.1
Anchor: Query_118691 (1..167)
Query: MF044248.1 (1..167) Homo sapiens microsatellite TPOX 8 [AATG]8 sequence
Relative orientation: forward
Aligned bases: 167
Segments: 1
Coverage: 100.0%
Identity: 100.0%
Mismatches: 0
Gaps: 0

NCBI
Home
PubMed
GenBank
BLAST
Sequence Viewer 3.22.0
Link To This View | Feedback
Tools
Tracks
MF044256.1
MF044248.1
MF044250.1
MF044249.1
MF044251.1
MF044252.1
MF044253.1
MF044254.1
MF044247.1
MF044255.1
MF044246.1
Tracks shown: 2/3
Contact | Copyright | Disclaimer | Privacy | Accessibility
National Center for Biotechnology Information, US National Library of Medicine
8600 Rockville Pike, Rockville, MD USA 20894

Figure 9. NCBI Sequence Viewer display of alignments returned from a STRSeq BioProject BLAST search of a sequence generated at the TPOX locus. The sequence matching the query shows 100% coverage and 100% identity. Searching sequences with start/stop coordinates varying from the sequences contained in the BioProject has challenges, as described in [3].

Acknowledgements

The authors express gratitude to the NCBI staff who have facilitated development of the BioProject: Drs. Lori Black, Melissa Landrum, Ilene Mizrachi, Kim Pruitt, George Riley, and Steven Sherry. The authors also acknowledge the input of the European Commission project DNASEQEX (HOME/2014/ISFP/AG/LAWX/4000007135) and the support of the ENFSI DNA Working Group and thank the many practitioners and researchers who provided valuable feedback.

NIST funding sources and disclaimers: This work was funded in part by the National Institute of Justice (NIJ) interagency agreement 1609-602-18NIJ: "Forensic DNA Applications of Next Generation Sequencing". Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Departments of Commerce or Justice. Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

References

[1] Gettings, K.B., Borsuk, L.A., Ballard, D., Bodner, M., Budowle, B., Devesse, L., King, J., Parson, W., Phillips, C., and P.M. Vallone. STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci. Forensic Science International: Genetics 31, 111-117 (2017).

[2] Parson, W., Ballard, D., Budowle, B., Butler, J.M., Gettings, K.B., Gill, P., Gusmão, L., Hares, D., Irwin, J., King, J., de Knijff, P., Morling, N., Prinz, M., Schneider, P.M.,Van Neste, C., Willuweit, S., and C. Phillips. Massively Parallel Sequencing of forensic STRs: Considerations of the DNA Commission of the International Society of Forensic Genetics (ISFG) on minimal nomenclature requirements, Forensic Science International: Genetics 22, 54–63 (2016).

[3] Gettings, K.B., Borsuk, L.A., and P.M. Vallone. Performing a BLAST search of the STRSeq BioProject. Forensic Science International: Genetics Supplement Series 6, e372-e374 (2017).