STRSeq: A resource for sequence-based STR analysis

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

APPLIED GENETICS

L.A. Borsuk, K.B. Gettings, P.M. Vallone

National Institute of Standards and Technology, Applied Genetics Group, Gaithersburg, MD, USA

Abstract - The STR Sequencing Project (STRSeq)[1] was initiated to facilitate the description of sequence-based alleles at 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: 1) observed sequence of an STR region, 2) annotation of the repeat region ("bracketing" consistent with the guidance of the International Society for Forensic Genetics) and flanking region polymorphisms, 3) information regarding the sequencing assay and data quality, and 4) 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
- X-chromosomal STR loci

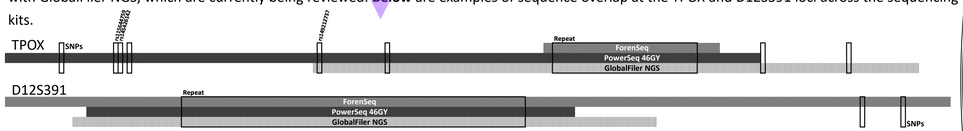
The BioProject will initially contain aggregate alleles across **4,612 samples** submitted by four laboratories: National Institute of Standards and Technology (NIST, the project organizer), Kings College London (KCL), University of North Texas Health Sciences Center (UNT), and University of Santiago de Compostela (USC)[2-6].

Currently, **1145** records have been uploaded into STRSeq. These submitted records represent the evaluation of approximately 25% of the current data set for 28 autosomal STR loci. **The plot to the right** is a representation of 35 autosomal STR loci from the four data sets representing **1739** unique sequences, the additional **594** autosomal sequences will potentially be uploaded to the STRSeq Project after additional evaluation and confirmation. STRSeq is meant to continually grow as new data sets are developed and new sequences are reported and confirmed.

Current participating laboratories and samples provided for evaluation

NIST	KCL	UNT HEALTH SCIENCE CENTER	USC
1786 Samples	1043 Samples	839 Samples	944 Samples

Method – Data from three Forensic STR sequencing kits are currently being evaluated for inclusion in STRSeq: 1) **ForenSeq** DNA Signature Prep Kit (Verogen), 2) **PowerSeq 46GY** (beta version, Promega), and 3) **GlobalFiler NGS** (Thermo Fisher). All of the samples currently represented in STRSeq have been sequenced with ForenSeq. NIST has sequenced subsets of >600 samples with PowerSeq (included in the current STRSeq records) and >200 samples with GlobalFiler NGS, which are currently being reviewed. **Below** are examples of sequence overlap at the TPOX and D12S391 loci across the sequencing kits.



NCBI houses the GenBank records. **To the left** are examples of records in the TPOX and D12S391 BioProjects. Orange boxes highlight differences in the records. All four TPOX allele "8" sequences have the same repeat but vary in the flanking regions, whereas D12S391 allele "27" varies in the repeat region but not in the flanking region.

To the right is an example of a D1S1656 allele "11" record. The goal is to provide the user with maximal information about the reported sequence. This includes confirmed length-based allele size, sequence coordinates corresponding to the GRCh38 reference sequence, and flanking region variants compared to the GRCh38 reference sequence.

TPOX allele 8

Homo sapiens microsatellite TPOX 8 [AATG]8 rs145426142 sequence
167 bp linear DNA
Accession: MG868075.1 GI: 1359120735
BioProject PubMed Taxonomy
GenBank FASTA Graphics

Homo sapiens microsatellite TPOX 8 [AATG]8 sequence
167 bp linear DNA
Accession: MF044248.1 GI: 1197990968
BioProject PubMed Taxonomy
GenBank FASTA Graphics

Homo sapiens microsatellite TPOX 8 [AATG]8 rs115644759 sequence
167 bp linear DNA
Accession: MF044249.1 GI: 1197990969
BioProject PubMed Taxonomy
GenBank FASTA Graphics

Homo sapiens microsatellite TPOX 8 [AATG]8 rs149212737 sequence
167 bp linear DNA
Accession: MF044250.1 GI: 1197990970
BioProject PubMed Taxonomy
GenBank FASTA Graphics

D12S391 allele 27

Homo sapiens microsatellite D12S391 27 AGATT[8 (AGAC)8 sequence
241 bp linear DNA
Accession: MH167199.1 GI: 1381324258
BioProject PubMed Taxonomy
GenBank FASTA Graphics

Homo sapiens microsatellite D12S391 27 AGATT[8 (AGAC)8 AGAT sequence
241 bp linear DNA
Accession: MH167200.1 GI: 1381324259
BioProject PubMed Taxonomy
GenBank FASTA Graphics

Homo sapiens microsatellite D12S391 27 AGGT [AGAT]17 (AGAC)8 AGAT sequence
241 bp linear DNA
Accession: MH167201.1 GI: 1381324260
BioProject PubMed Taxonomy
GenBank FASTA Graphics

Homo sapiens microsatellite D1S1656 11 CCTA [TCTA]10 sequence

GenBank: MH174836.1
FASTA Graphics

Go to:

```
LOCUS       MH174836                 124 bp    DNA     linear   PRI 04-SEP-2018
DEFINITION  Homo sapiens microsatellite D1S1656 11 CCTA [TCTA]10 sequence.
ACCESSION   MH174836
VERSION     MH174836.1
DBLINK      BioProject: PRJNA385553
KEYWORDS    STRSeq; STR; D1S1656.
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 124)
  AUTHORS   Gettings,K.B., Borsuk,L.A., Ballard,D., Bodner,M., Budowle,B.,
            Devesse,L., King,J., Parson,W., Phillips,C. and Vallone,P.M.
  TITLE     STRSeq: A catalog of sequence diversity at human identification
            Short Tandem Repeat loci
  JOURNAL   Forensic Sci Int Genet 31, 111-117 (2017)
  PUBMED    28881235
REFERENCE   2  (bases 1 to 124)
  AUTHORS   NIST,A.G.G.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-APR-2018) Applied Genetics Group, National Institute
            of Standards and Technology, 100 Bureau Drive, MS-8314,
            Gaithersburg, Maryland 20899, United States of America
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
            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          :: D1S1656
            Length-based allele     :: 11
            Bracketed repeat        :: CCTA [TCTA]10
            Sequencing technology   :: MiSeq FGx; MiSeq
            Coverage                :: >35X
            Length-based tech.      :: PowerPlex Fusion, 213fa1
            Assembly                :: GRCh38 (GCF_000001405)
            Chromosome              :: 1
            RefSeq Accession        :: NC_000001.11
            Chrom. Location         :: 230769555..230769704
            Repeat Location         :: 230769616..230769683
            Cytogenetic Location    :: 1q42.2
            ##HumanSTR-END##
FEATURES             Location/Qualifiers
     source          1..124
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
     misc_feature    1..105
                     /note="Illumina ForenSeq DNA Signature Prep Kit"
     misc_feature    6..124
                     /note="Promega PowerSeq 46GY System"
     repeat_region   60..105
                     /rpt_type=tandem
                     /satellite="microsatellite:D1S1656"

ORIGIN
        1 ttcagagaaa tagaatcact aggcaaacaa atatatatac atacaattaa acacacacac
       61 actatactat ctatctatct atctatctat ctatctatct atctacatca cacagttgac
      121 cctcga
//
```

Discussion –The next sets of autosomal, X, and Y STRSeq records representing the remainder of the current samples will be uploaded to the STRSeq BioProject over the next year. Future plans for this NIJ-funded effort include a pathway for researchers to submit additional alleles and customized interface tools, which would allow users to easily search the STRSeq data set. 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.

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.

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. Gettings, K.B., Borsuk, L.A., Steffen, C.R., Kiesler, K.M., and P.M. Vallone. **Sequence-based U.S. population data for 27 autosomal STR loci.** Forensic Science International: Genetics 37, 106-115 (2018).

3. Borsuk, L.A., Gettings, K.B., Steffen, C.R., Kiesler, K.M., and P.M. Vallone. **Sequence-based U.S. population data for the SE33 locus.** Electrophoresis 21, (2018).

4. Devesse, L., Ballard, D., Davenport, L., Riethorst, I., Mason-Buck, G., and D. Syndercombe Court. **Concordance of the ForenSeq system and characterization of sequence-specific autosomal STR alleles across two major population groups.** Forensic Science International: Genetics 34, 57-61 (2018).

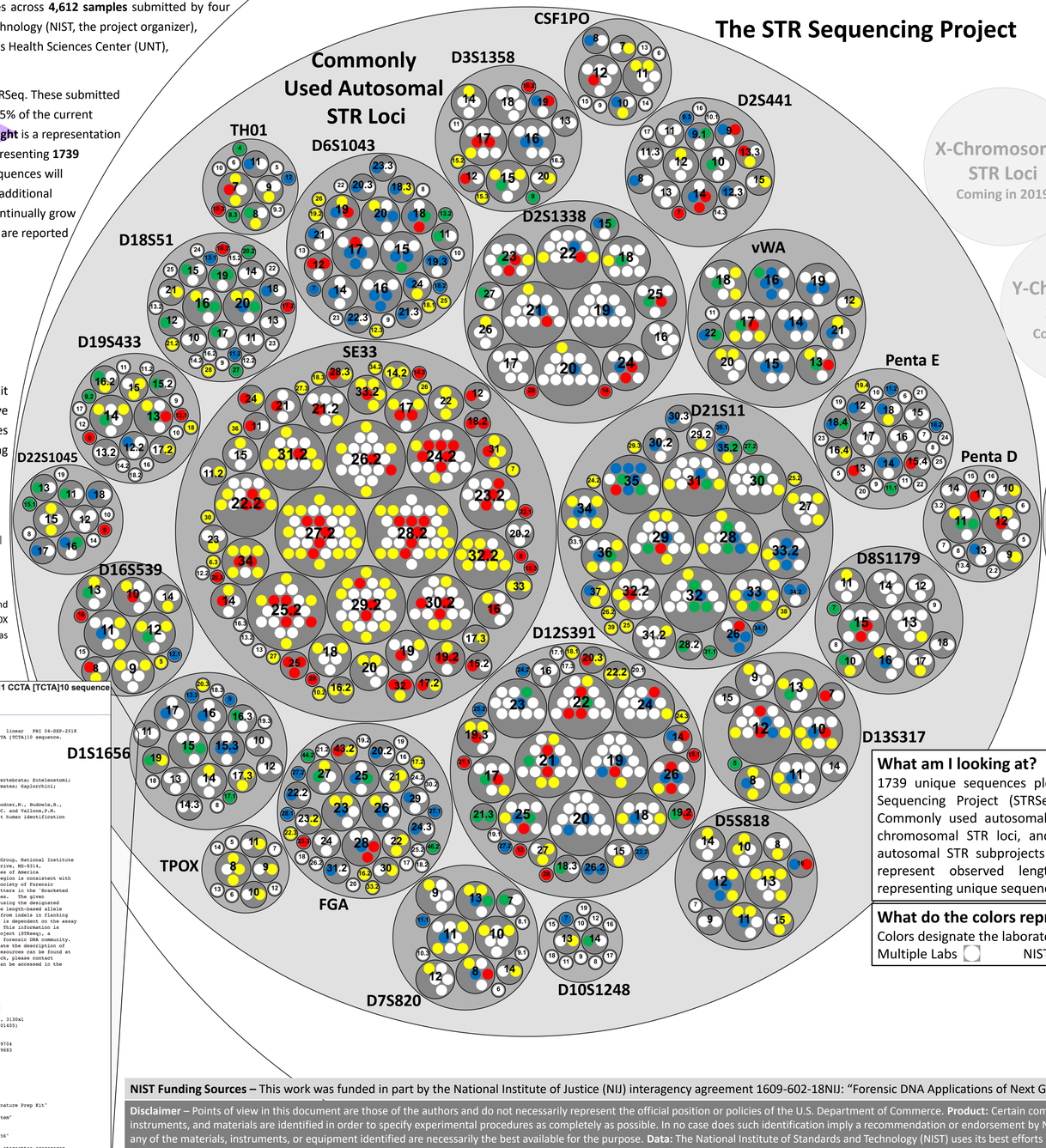
5. Novroski, N.M.M., King, J.L., Churchill, J.D., Seah, L.H., and B. Budowle. **Characterization of genetic sequence variation of 58 STR loci in four major population groups.** Forensic Science International: Genetics 25, 214-226 (2016).

6. Philips, C., Devesse, L., Ballard, D., van Weert, L., de la Puente, M., Melis, S. Alvarez Iglesias, V., Freire-Aradas, A., Oldroyd, N., Holt, C., Syndercombe Court, D., Carracedo, A., and M.V. Lareu. **Global patterns of STR sequence variation: Sequencing the CEPH human genome diversity panel for 58 forensic STRs using the Illumina ForenSeq DNA Signature Prep Kit.** Electrophoresis 21, (2018).

RawGraphic.io ref

(Circle Packing Plot) Mauri, M., Elli, T., Caviglia, G., Uboldi, G., and M. Azzi. **RAWGraphs: A Visualisation Platform to Create Open Outputs.** In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter* (p. 28:1–28:5). New York, NY, USA: ACM. (2017).

(ref) 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 37, 54–63 (2016).



What am I looking at?

1739 unique sequences plotted in a hierarchical representation of the STR Sequencing Project (STRSeq), which is divided into four subprojects: 1) Commonly used autosomal STR loci, 2) Alternate autosomal STR loci, 3) Y-chromosomal STR loci, and 4) X-chromosomal STR loci. Currently only the autosomal STR subprojects contain sequence records. Circles within the loci represent observed length-based alleles, which contain colored dots representing unique sequences.

What do the colors represent?

Colors designate the laboratory from which the unique sequence was identified:
Multiple Labs NIST KCL UNT USC

NIST Funding Sources – 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".

Poster # 61 at:
29th ISHI, Phoenix, AZ
September 24-27, 2018

Email: strseq@nist.gov