STRSeq: The evolution of the STR sequencing project

Lisa A. Borsuk, Peter M. Vallone, and Katherine B. Gettings
U.S. National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA

Email: lisa.borsuk@nist.gov

ISFG WASHINGTON, DC ISFG 2022 INTERNATIONAL SOCIETY FOR FORENSIC GENETICS

P112

Abstract
The STR Sequencing Project (STRSeq) began in 2017 to catalog sequences at the Short Tandem Repeat (STR) loci commonly used for human identification [1]. Working with NCBI and the forensic community, a GenBank record template was developed to include information of value to the forensic community. Records contain:

1) Complete **sequence strings** from commercial forensic sequencing assays
2) Genomic locations of the targeted STRs
3) Information provided by **length-based** assays
4) Nomenclature information that includes bracketing of the STR and identified flanking variations
5) General information about the locus

Over 2500 unique sequence records have been uploaded in the last five years to GenBank, including sequences from eleven publications covering 70 STR loci.

The movement toward implementing sequencing-based technology for STR loci requires that the new, sequence-based results are compatible with the standard, length-based results. Currently, an ISFG DNA Commission on STR Nomenclature is working to make recommendations for reporting forensic STR sequences. STRSeq records will be updated to incorporate the recommendations of the Commission to standardize the information reported.

COMMENT BLOCK

The comment block is being updated to reflect the new record format and information; draft wording is shown here. Individual record notes will now be found in the 'Notes' field in the HumanSTR section of the record.

```
COMMENT     On Jul 8, 2022 this sequence version
            replaced MG988075.2. 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 the reported
            sequence is dependent on the assay and the quality of
            the flanking sequence. Sequencing assays are coded as
            Verogen ForenSeq DNA Signature Prep Kit (FS), ForenSeq
            MainstAY (MS), Applied Biosystems Precision ID
            GlobalFiler NGS STR Panel v2 (GF), and Promega
            PowerSeq 46GY System (PS). All other methods of
            sequencing are coded as Targeted Sequencing (TS).
            Bracketing of the minimum range and full record
            sequences are performed by STRNaming PMID: 33607395,
            which is consistent with the guidance of the ISFG
            (International Society of Forensic Genetics) [In
            preparation]. The ISFG min. range code is currently
            under evaluation. 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. For questions or feedback, please contact
            strseq@nist.gov.
```

##HumanSTR

These fields are unique to STRSeq records and were developed in collaboration with the forensic community for the forensic community.

New fields added are highlighted in green. The 'Bracketed repeat' has been renamed to 'Historical bracketing', highlighted yellow. Fields removed from the report are 'Repeat location' and 'Cytogenetic location'.

```
##HumanSTR-START##
Sequence attribution    :: Applied Genetics Group,
                           NIST
STR locus name          :: TPOX
Length-based allele     :: 8
Minimum range bracket   :: TGAA[8]
Bracketed record seq.   :: TGAA[8]_-94G>A
Sequencing technology   :: MiSeq FGx
Sequencing assay code   :: FS,GF,PS
Coverage                :: >30X
Length-based tech.      :: PowerPlex Fusion,3130xl
Assembly                :: GRCh38 (GCF_000001405)
Chromosome              :: 2
Ref. seq. accession     :: NC_000002.12
Chrom. location         :: 1489529..1489732
ISFG minimum range      :: 1489647..1489692
ISFG min. range code    ::
Frequency reference     :: STRidER.online
STR locus alt name      :: hTPO, TPO
Historical bracketing   :: [AATG]8
Notes                   ::
##HumanSTR-END##
```

ORIGIN

This sequence can be highlighted by clicking on the links in the FEATURES section to identify kit ranges, the ISFG minimum recommended reporting range, and flanking region SNPs (if present).

```
ORIGIN
        1 cactggcctg tgggtccccc catagatcat aagcccagga ggaagggctg tgtttcaggg
       61 ctgtgatcac tagcacccag aaccgtcgac tggcacagaa caggcactta gggaaccctc
      121 actgaatgaa tgaatgaatg aatgaatgaa tgaatgtttg ggcaaataaa cgctgacaag
      181 gacagaaggg cctagcggga aggg
//
```

Conclusion

Work is ongoing to update existing records into this new format, corresponding to the development of ISFG STR nomenclature recommendations. Going forward, new STRSeq record submissions will follow this new format. Additionally, via collaboration with GMI, we aim to incorporate a STRSeq nomenclature check into STRidER [5] sequence-based STR population data QC.

For comments, questions, or concerns please contact us at **strseq@nist.gov** or email the author.

The border is a Treemap plot [4] of STRSeq records. Each color represents a locus. Each square represents a record. Allele length-based information is included in the square.

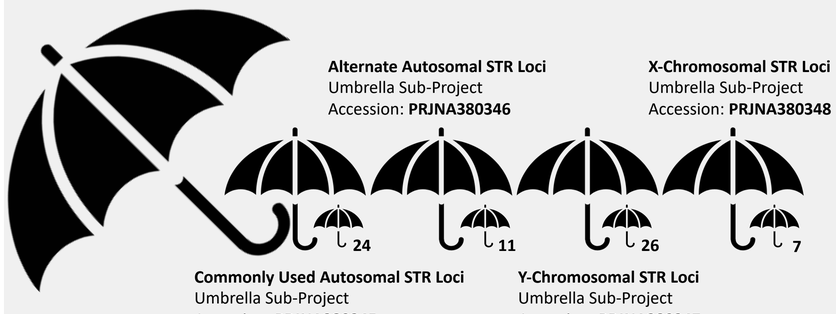
References

1. Gettings KB, Borsuk LA, Ballard D, Bodner M, Budowle B, Devesse L, King J, Parson W, Phillips C, and Vallone PM. STRSeq: A catalog of sequence diversity at the human identification Short Tandem Repeat loci. Forensic Science International: Genetics. 2017; 31:111-117
2. Hoogenboom J, Sijen T, and van der Gaag KJ. STRNaming: Generating simple, informative names for sequenced STR alleles in a standardized and automated manner. Forensic Science International: Genetics. 2021; 52:102473
3. Young B, Faris T, and Armogida L. A nomenclature for sequence-based forensic DNA analysis. Forensic Science International: Genetics. 2019; 42:14-20
4. Mauri M, Elli T, Caviglia G, Uboldi G, and Azzi M. RAWGraphs: A Visualisation Platform to Create Open Outputs. Proceedings of the 12th Biannual Conference on Italian SIGCHI 2017; Chapter (p. 28:1–28:5).
5. Bodner M, Bastisch I, Butler JM, Fimmers R, Gill P, Gusmão L, Morling N, Phillips C, Prinz M, Schneider P, and Parson W. Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER). Forensic Science International: Genetics. 2016; 24:97-102

Funding

- This work was in part supported by the NIST Special Programs Office: Forensic Genetics.
- This work was in part supported by 2016 and 2021 interagency agreements with the National Institute of Justice.

The STR Sequencing Project [Top Umbrella Project] **Accession: PRJNA380127**



DEFINITION LINE

All STRSeq definitions start with "Homo sapiens microsatellite". The rest of the definition line is generated from fields in the HumanSTR section of the report. This includes:

1) The STR locus name — **TPOX**
2) Length-based allele — **8**
3) Bracketed record seq. — **TGAA[8]_-94G>A**
4) Sequencing assay code — **FS,GF,PS**

This combination of information is designed to give each sequence a unique definition. The dbSNP accession numbers (rs#) are replaced by the STRNaming designation of flanking region polymorphisms; rs# are still reported and linked to dbSNP in the record FEATURES section.

GenBank ▾ Send to: ▾

Homo sapiens microsatellite TPOX 8 TGAA[8]_-94G>A FS,GF,PS sequence

GenBank: MG988075.3
FASTA Graphics

Go to: ▾

```
LOCUS       MG988075               204 bp    DNA     linear   PRI 08-JUL-2022
DEFINITION  Homo sapiens microsatellite TPOX 8 TGAA[8]_-94G>A FS,GF,PS sequence.
ACCESSION   MG988075
VERSION     MG988075.3
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 204)
 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     28888135
REFERENCE   2 (bases 1 to 204)
 AUTHORS    NIST, A.G.G.
 TITLE      Direct Submission
 JOURNAL    Submitted (26-FEB-2018) Applied Genetics Group, National Institute of
            Standards and Technology, 100 Bureau Drive, MS-8314, Gaithersburg,
            Maryland 20899, United States of America
```

NCBI BioProject

STRSeq records are organized within an NCBI BioProject hierarchical structure. Records are divided into categories from the top umbrella project (full set of records), into locus type sub-projects, then down to the individual locus records (base projects). This organization is illustrated on the left by the layers of umbrellas, with the number of locus-specific base projects indicated next to the smallest umbrellas.

GenBank

The STRSeq GenBank record is customized by the Applied Genetics Group at NIST to include information relevant to the forensic community, and NIST manages new record submission. If you are interested in adding published or soon-to-be published sequences to the STRSeq BioProject, please contact strseq@nist.gov.

FEATURES SECTION

This section includes the location of regions of interest within the record sequence: the ISFG minimum recommended reporting range, the vendor-indicated range for commercial kits (when available), and flanking sequence variants.

```
FEATURES             Location/Qualifiers
     source          1..204
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
     misc_feature    1..174
                     /note="Promega PowerSeq 46GY System"
     variation       29
                     /note="-94G>A"
                     /db_xref="dbSNP:rs145426142"
     misc_feature    77..204
                     /note="Applied Biosystems Precision ID GlobalFiler NGS
                     STR Panel v2"
     repeat_region   119..164
                     /note="minimum range"
                     /rpt_type=tandem
                     /satellite="microsatellite:TPOX"
     misc_feature    125..161
                     /note="Verogen ForenSeq DNA Signature Prep Kit"
```

misc feature*
- Information and location of a kit in the record sequence

variation*
- Information and location of a SNP or indel in the record sequence. It includes a link to the dbSNP record.

repeat region*
- Information and location of the minimum range of the locus in the record sequence.

*These are links that will highlight the specific region of the sequence reported in the record

Sequence attribution
- Institute credited with the publication of the sequence

Minimum range bracket
- The minimum range is based on the recommendations of the ISFG Nomenclature committee. The bracket is generated using STRNaming [2] on the minimum range sequence.

Bracketed record seq.
- The full sequence is also bracketed using STRNaming and reported here.

ISFG minimum range
- Reference genome GRCh38 coordinates for the minimum range as recommended by the ISFG Nomenclature committee

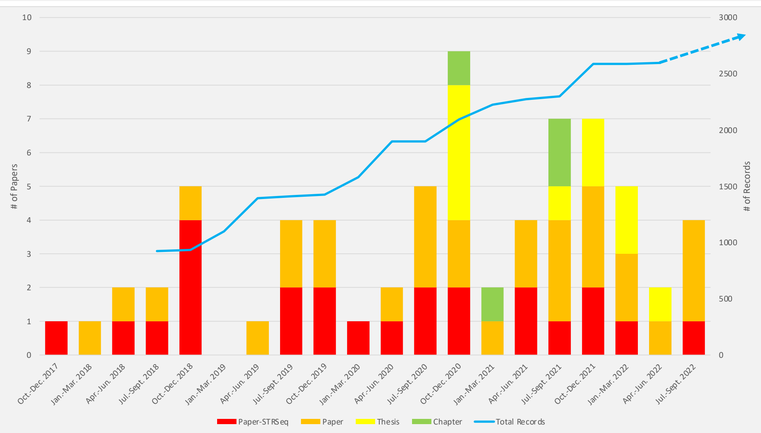
ISFG min. range code [In preparation]
- The goal of this field is to represent the minimum range sequence in a simplified code of a minimum number of characters. Under consideration is the sequence identifier (SID) from Young *et al.* [3] using 3 or more characters.

Frequency reference
- Present when available.

Notes
- Allows for additional information specific to the sequence or a subset of sequences.

GenBank Submission

The original test records were submitted in mid-2017. In mid-2018 the initial set of records became publicly available. The solid blue line in the plot below shows the increase in public STRSeq records. The dashed blue line represents future growth. As of September 2022, there are 2,597 publicly available records. These STRSeq records are associated with 11 publications from the authors of the original STRSeq paper [1]. Additional STRSeq records are in preparation and additional publications are being considered for potential STRSeq records.



Citing STRSeq

The foundational STRSeq paper [1] was published in Nov. 2017. Using Google Scholar and searching for publications that cite "STRSeq: A catalog of sequence diversity at the human identification Short Tandem Repeat loci" identified ≈75 documents. Above is the plot of the counts of papers, theses, and book chapters found in the list. A total of 68 documents are included in the bar plot. They are broken down into categories: Paper-STRSeq*, Paper, Thesis**, and Chapter. This plot indicates the uptake of the STRSeq resource in the forensic community.

*Paper-STRSeq are papers that have authors from the STRSeq paper.
** Some Theses only referenced a year and were counted at the end of that calendar year.

Disclaimer

The 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. Department of Commerce. Certain commercial software, 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 NIST, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose.

Research Protections Declaration

All work has been reviewed and approved by the National Institute of Standards and Technology Research Protections Office. This study was determined to be "not human subjects research" (often referred to as research not involving human subjects) as defined in U. S. Department of Commerce Regulations, 15 CFR 27, also known as the Common Rule (45 CFR 46, Subpart A), for the Protection of Human Subjects by the NIST Human Research Protections Office and therefore not subject to oversight by the NIST Institutional Review Board.

Download a PDF of this poster