



The author(s) shown below used Federal funding provided by the U.S. Department of Justice to prepare the following resource:

Document Title: Forensic DNA Applications of Massively Parallel / Next Generation Sequencing

Author(s): NIST

Document Number: 304646

Date Received: April 2022

Award Number: 2016-DNR-6150

This resource has not been published by the U.S. Department of Justice. This resource is being made publicly available through the Office of Justice Programs' National Criminal Justice Reference Service.

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Project design and methods

The STRSeq “BioProject” serves to organize these records within the GenBank repository and is divided into categories (as specified in Gettings 2015): commonly used autosomal STRs, alternate autosomal STRs, Y-chromosomal STRs, and X-chromosomal STRs. Each of these categories is divided further into locus-specific BioProjects. This BioProject hierarchy allows easier access to the GenBank records by browsing, BLAST searching, or ftp download. The sequence records in GenBank are flat files of specified format, such that they can be downloaded and parsed en masse as shown in Fig. 3, or explored via an interactive graphic display (Fig. 4).

Figure 3. An example DYS481 STRSeq BioProject GenBank record. a) The Definition Line uniquely identifies each allele with components of the record. b) Hyperlinks to FASTA sequence and Graphics view (see Figure 4). c) GenBank sequence identifiers and link to the parent BioProject (e.g. DYS481), records in which the sequence string is updated branch off as e.g. MW073994.2, coexisting with MW073994.1. d) References describe the BioProject and identify the submitting laboratory. e) Field-based information relevant to STRSeq records and orienting the sequence to the GRCh38 reference sequence. f) Position of the repeat region within the sequence, position and dbSNP rs number of variations in the flanking regions (when applicable), and subset of sequence observed with different commercial assays (when applicable). Selecting a feature highlights corresponding region in sequence string. Rs numbers are hyperlinked to dbSNP. g) Full sequence string reported by the submitting laboratory. Length of reported sequence is assay- and quality-dependent, generally consistent with assay-specific configuration files published in (Woerner 2017).

a **Homo sapiens microsatellite DYS481 19 [CTT]19 FS, PS sequence**

GenBank: MW073994.1

b [FASTA](#) [Graphics](#)

[Go to:](#)

c

LOCUS MW073994 104 bp DNA linear PRI 28-OCT-2020
 DEFINITION Homo sapiens microsatellite DYS481 19 [CTT]19 FS, PS sequence.
 ACCESSION MW073994
 VERSION MW073994.1
 DBLINK BioProject: PRJNA396135
 KEYWORDS STRSeq; STR; DYS481.
 SOURCE Homo sapiens (human).
 ORGANISM [Homo sapiens](#)
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
 Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
 Catarrhini; Hominidae; Homo.

d

REFERENCE 1 (bases 1 to 104)
 AUTHORS Gettings,K.D., Borsuk,I.A., Ballard,D., Bodner,M., Budowle,B.,
 Devesse,I., 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 [2888135](#)
 REFERENCE 2 (bases 1 to 104)
 AUTHORS NIST,A.G.G.
 TITLE Direct Submission
 JOURNAL Submitted (04-OCT-2020) Applied Genetics Group, National Institute
 of Standards and Technology, 100 Bureau Drive, MS-B314,
 Gaithersburg, Maryland 20899, United States of America

e

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. Sequencing assays are
 coded as Illumina ForenSeq DNA Signature Prep Kit (FS), Applied
 Biosystems Precision ID GlobalFiler NGS STR Panel v2 (GF), and
 Promega PowerSeq 460Y System (PS). 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.

f

```
##HumanSTR-START##
STR locus name      :: DYS481
Length-based allele  :: 19
Bracketed repeat    :: [CTT]19
Sequencing technology :: MISeq FGx; MISeq
Sequencing assay code :: FS, PS
Coverage            :: >30X
Length-based tech.  :: PowerPlex Y23, 3130x1
Assembly            :: GRCh38 (GCF_000001405)
Chromosome          :: Y
RefSeq Accession    :: NC_000024.10
Chrom. Location     :: 8558302..8558414
Repeat Location     :: 8558337..8558402
Cytogenetic Location :: Yp11.2
##HumanSTR-END##
```

g

```
1 taaaggaaat gttgctaacg ctgttcagca tgcctcttct tctcttctt ctctctcttc
61 tctcttctt tctcttctt ctctctctt ttttttgagt ctg
```

FEATURES Location/Qualifiers

source	1..104
misc_feature	1..104
misc_feature	12..98
repeat_region	36..92

ORIGIN

```
1 taaaggaaat gttgctaacg ctgttcagca tgcctcttct tctcttctt ctctctcttc
61 tctcttctt tctcttctt ctctctctt ttttttgagt ctg
```


