Forensic Science International: Genetics Supplement Series 8 (2022) 245–247

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

journal homepage: www.elsevier.com/locate/fsigss

STRSeq: FAQ for submitting

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

US National Institute of Standards and Technology, Applied Genetics Group, 100 Bureau Drive, Gaithersburg, MD 20899–8314, USA

ARTICLE INFO

Keywords:
STRSeq
NGS/MPS
Short tandem repeats
GenBank
Sequencing
Human identification

ABSTRACT

The STR sequencing project was developed due to the necessity of publicly sharing sequencing information about Short Tandem Repeats (STR) associated with human identification. Next Generation Sequencing (NGS) or Massively Parallel Sequencing (MPS) is becoming more prevalent in forensics. Having a publicly accessible centralized location to maintain this forensic-specific type of sequencing information is useful for the community. It is also important that the community participate in providing information to strengthen this resource. Here we answer common questions about the STRSeq BioProject and encourage community engagement.

1. Introduction

The STR Sequencing Project (STRSeq) began in 2017 to catalog sequences at the Short Tandem Repeat (STR) loci commonly used for human identification [1]. In collaboration with NCBI and the forensic community, a GenBank record template was developed to include information of value to the forensic community. Over 2500 unique sequence records have been uploaded to GenBank in the last five years, 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. All future records will be structured in this new format. Expanding the STRSeq BioProject is an ongoing endeavor.

2. How do you decide which loci to include?

The loci included in the STRSeq BioProject are present in commercial forensic sequencing kits available for purchase in the U.S. The commercial manufacturers provide information about the targeted sequences, including reportable sequence ranges. Our goal is to provide the entire reportable sequence range per kit in the STRSeq record. Sequences that are not associated with known, evaluated commercial kits could introduce issues such as non-target results, leading to chromosomal location errors in the STRSeq records, especially for the less common loci. Sanger sequencing has been used to augment results derived from a commercial STR sequencing kit on a case-by-case basis after evaluation and consideration of the utility of this additional information. For example, Sanger sequencing has been provided in cases where there are known null or discordant alleles from commercial kits, in order to provide clarification about potential issues when developing primers for these loci (e.g. a "22" allele for the locus SE33 (GenBank: MH232764.2) included Sanger sequence information).

3. How is the STRSeq BioProject organized?

The STRSeq BioProject has a tiered structure that allows for additional BioProjects to be added as needed when new kits, with additional loci, become available. At the top of the structure is the umbrella BioProject STR sequencing project (BioProject accession: PRJNA380127 https://www.ncbi.nlm.nih.gov/bioproject/380127). There are four sub-BioProjects: commonly used autosomal STR loci (BioProject accession: PRJNA380345), alternate autosomal STR loci (BioProject accession: PRJNA380346), Y-chromosomal STR loci (BioProject accession: PRJNA380347), and X-chromosomal STR loci (BioProject accession: PRJNA380348). There are 68 base BioProjects. This organization allows for a record to be added to a base BioProject and then be accessible through all higher-level BioProjects that include that base BioProject.

4. Why do some loci have more records than others?

All STRSeq base BioProjects have records associated with them, and

* Corresponding author.
E-mail address: lisa.borsuk@nist.gov (L.A. Borsuk).

https://doi.org/10.1016/j.fsigss.2022.10.050
Received 16 September 2022; Accepted 18 October 2022
Available online 20 October 2022
1875-1768/Published by Elsevier B.V.

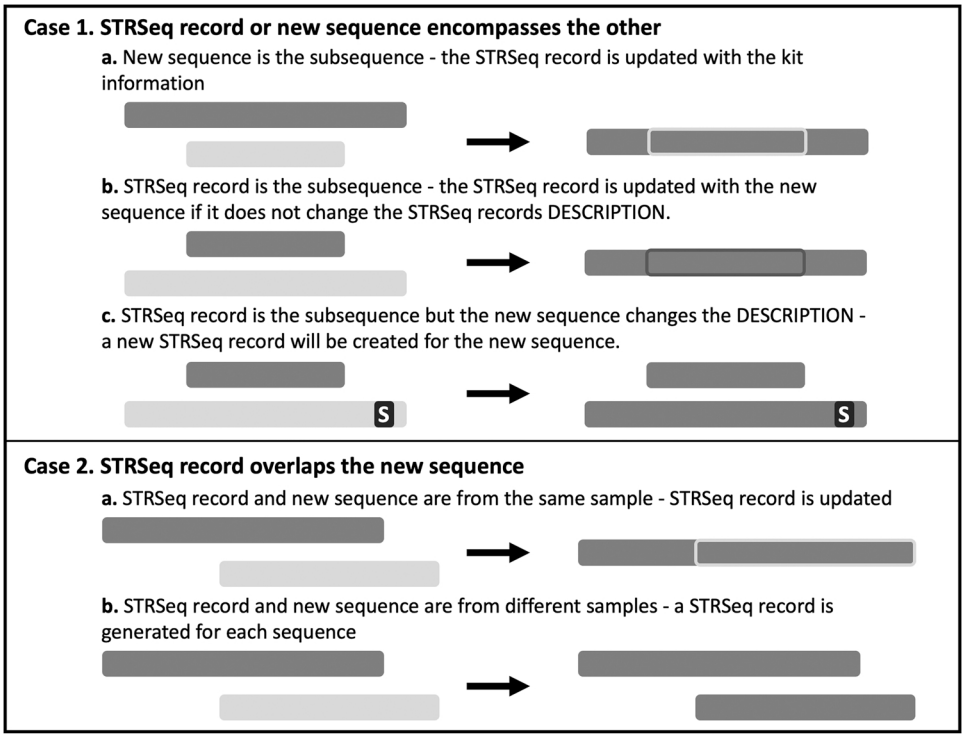

Fig. 1. Incorporating new sequences into STRSeq. The dark gray bars represent STRSeq records. The light gray bars represent new sequences. The black box with an 'S' represents a SNP or other sequence change. Case 1a. the sequence range of the STRSeq record encompasses the entire new sequence range; the additional (smaller range) sequence information can be added to an existing record for the original (larger range) STRSeq record. Case 1b. the new sequence encompasses the current STRSeq record. The sequence string will be replaced with the updated sequence if no new flanking region polymorphisms are included in the new sequence range. Case 1c. the additional sequence with flanking polymorphism(s) will be added as a new STRSeq record. Case 2a. sequences from multiple assays which overlap incompletely will only be merged when both sequences are derived from the same sample. Case 2b. the sequences are reported separately if they are not from the same sample, regardless of the content of the flanking sequence. The DESCRIPTION will differ based on the kits used to generate each sequence.

the number of records is generally associated with the level of sequence polymorphism present at the locus. Some of the alternate autosomal STR loci BioProjects are minimally populated, due to their presence in a subset of commercially available kits. Information for all the loci continues to be evaluated, collected, and added from published literature and collaborating laboratories.

5. Where does the data come from?

Currently, the STRSeq records derive from the five laboratories associated with the STRAND working group [2]. This includes 11 publications and over 4600 samples sequenced. There are currently approximately 30 additional publications from a variety of laboratories published between 2016 and 2022 that are being considered as a source of new STRSeq records.

6. Where are the STR sequence allele frequencies?

It is not possible to include comprehensive allele frequency data in the GenBank record format. A single publication containing each unique sequence is associated with each record, but most sequences are found in multiple populations and have been reported across publications. Maintaining accurate allele frequency information in the STRSeq records exceeds the scope of this project. STRidER (STRs for identity ENFSI Reference database [3]) is specifically designed for collecting, maintaining, and serving out STR allele frequency information, and planned updates to this resource will expand its scope to include STR sequence-based allele frequency data. A frequency reference field has been added to the new STRSeq record format in order to direct users to STRidER or other applicable resources.

7. Can I send you a sequence?

Please contact strseq@nist.gov to find out more about submitting sequences. Sequences that are published or are in the process of being published are candidate sequences for STRSeq. Sequences can be submitted to STRSeq and held for public release until the publication has been accepted. Collaborative efforts are ongoing to make a publicly available search tool for identifying sequences in STRSeq and associated STRidER frequency data. In the future, when a user searches an STR sequence with no match in the STRSeq BioProject, we aim to provide users with a pathway for submitting appropriate sequences.

8. Can I submit my sequences directly to STRSeq through GenBank at NCBI?

No, the sequences that are incorporated into STRSeq are evaluated by the NIST Applied Genetics Group, and additional information is added to the record. In some cases, existing STRSeq records are updated with new information rather than creating a new STRSeq record, in order to avoid creating duplicate records. In cases where multiple kits target the same locus, the ranges of the kits' sequences are taken into consideration, see Fig. 1.

9. Conclusion

In the last five years, STRSeq has evolved and grown. It continues to develop and expand to bring value to the forensic community. Feedback is essential and greatly appreciated and can be sent to strseq@nist.gov.

Role of funding

This work was funded by the NIST Special Programs Office: Forensic Genetics and by 2016 and 2021 Interagency Agreements with the National Institute of Justice.

Conflict of interest

None.

Acknowledgments

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. Specific commercial equipment, instruments, and materials are identified to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by NIST, nor does it mean that any materials, instruments, or equipment identified are necessarily the best available for the purpose. The NIST Human Subjects Protection Office approved this work. We wish to thank the GenBank group at NCBI, specifically Dr. Lori Black.

References

[1] K.B. Gettings, et al., STRSeq: a catalog of sequence diversity at the human identification Short Tandem Repeat loci, Forensic Sci. Int. Genet. 31 (2017) 111–117.

[2] K.B. Gettings, et al., Report from the STRAND Working Group on the 2019 STR sequence nomenclature meeting, Forensic Sci. Int. Genet. 43 (2019), 102165.

[3] M. Bodner, et al., 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 Sci. Int. Genet. 24 (2016) 97–102.