Final Summary Overview: STRSeq BioProject (Award # 2016-DNR-6150)

Abstract

From 2017 to present[1], NIST scientists have worked with staff scientists at NCBI and leveraged existing infrastructure to develop standardized STR (short tandem repeat) DNA sequence records which are now available to the forensic community. This project is called STRSeq, and the basic framework can be accessed at https://www.ncbi.nlm.nih.gov/bioproject/380127. Using published population data, over 2100 records have been created to-date, and this catalog serves as a backbone for software development and as a future resource to facilitate interlaboratory communication. Overall, the transition from the current capillary electrophoresis typing technology to STR sequencing results in a more discriminating marker system, which is expected to improve public safety by providing more powerful statistics and greater ability to differentiate individual contributors to DNA mixtures.

Purpose

As the forensic DNA community evaluates the potential of sequencing applications for short tandem repeat (STR) loci, an analogous need has arisen to characterize the allelic diversity in these regions of the human genome. Large-scale sequencing projects within the broader genomics community have historically used methods which result in sequences that are of insufficient length to capture both ends of these repetitive regions and, further, often avoided analyzing repetitive regions entirely due to their complexity and non-conformity to typical alignment parameters (Willems 2017). Additionally, knowledge of the forensic literature is needed to report STR sequences in the same manner historically established by the forensic community. Even within forensic sequencing studies, there have been differences in the reporting of sequence-based STR alleles. Names of convenience have not been standardized and may create confusion about the specific allele being reported. More importantly, there have been differences in DNA strand reporting: some STR loci were originally

[1] This award period was three years, 2017-2019. Year one included several STR sequencing and sequence nomenclature tasks, whereas the STRSeq Bioproject described in this summary report was the singular focus of years two and three. This report was requested and prepared in Q1 of 2021.

characterized on the reverse strand of the modern human genome reference sequence, because this was the "coding" strand of a neighboring gene. The DNA Commission of the ISFG on minimal nomenclature considerations in 2016 advocated reporting all sequences in the forward strand orientation (Parson 2016). In particular, STRs for which the reported strand has changed over time could differ in the start position of the repeat region. This can result in shifted (different) allele number designations for the same sequence. Lastly, the recovery and reporting of varying lengths of regions adjacent to the repeat (and hence, the detection of variants in these "flanking" regions) is inherent to differences in kit designs and bioinformatic pipelines.

A 2016 survey was published by the European Network of Forensic Science Institutes (ENFSI) DNA Working Group (Alonso 2017), in which over half of the 33 responding laboratories had already purchased at least one sequencing instrument. Additionally, the respondents (primarily composed of government forensic laboratories across 25 countries) reported lack of nomenclature and reporting standards as the highest ranking scientific and legal challenge for the implementation of new sequencing technologies in forensic genetics. Also in 2016, the Applied Genetics Group of the U.S. National Institute of Standards and Technology (NIST) queried forensic laboratories to assess the utility of STR reference sequences for loci of forensic interest. The feedback received from 22 U.S. and international forensic laboratories (representing 11 countries) mirrored the ENSFI survey with strong support for the development of STR sequence nomenclature resources.

In response to this need, NIST obtained funding through the NIJ and worked with the U.S. National Center for Biotechnology Information (NCBI) to develop an STR nomenclature resource. This work leveraged NIST's over 20-year history supporting the forensic STR typing community (Ruitberg 2001) and NCBI's extensive infrastructure for accepting, maintaining and serving DNA sequence data. Thus, the STR Sequencing Project (STRSeq) was initiated to facilitate the description of sequence-based alleles at the STRs targeted in human identification assays. This resource now consists of a curated catalog of sequence diversity at forensic STR loci, along with the key elements of nomenclature conforming to current guidelines (Parson 2016) and serves as a foundation during these years of transition as well as a stable resource for the future.

Project subjects

The initial data used to populate STRSeq are the aggregate alleles observed in targeted sequencing studies of single source samples across four laboratories: NIST, Kings College London (KCL), University of North Texas Health Sciences Center (UNT), and University of Santiago de Compostela (USC), for a total of 4673 individuals. Since only aggregate alleles are displayed, the individual source of each allele is anonymized. Upload of data into the STRSeq BioProject has been approved by the NIST Research Protections Office, whereas sample sequencing within each laboratory was subject to review by the home institution. Records were added to the STRSeq BioProject in sets, largely coinciding with associated publications, as follows:

NIST: N=1786 samples from multiple sources: 1) N=665 liquid blood samples purchased from Interstate Blood Bank (Memphis, TN) and Millennium Biotech, Inc. (Ft. Lauderdale, FL) with self-declared ancestries from four different US population groups: Caucasian, African American, Asian and Hispanic (population group names as collected by blood bank); 2) N=781 buccal swabs provided by DNA Diagnostics Center (Fairfield, OH) from paternity testing samples with self-declared ancestries from three different US population groups: Caucasian, African American and Hispanic (population group names as collected by parentage testing laboratory); 3) N=297 buccal swabs collected from anonymous volunteers of self-reported, diverse ancestries, provided by the George Washington University; and 4) N=43 control samples and reference materials. For most of these samples, capillary electrophoresis (CE) data have been published previously at all sequenced STR loci (Hill 2008, Hill 2013). The sequence based STR data from the ForenSeq system (Verogen) for 1036 of these samples has been published for the 27 autosomal STR loci (Gettings 2018), the "hidden" SE33 locus (Borsuk 2018), and the 7 X-STR loci (Borsuk 2020), with the 24 Y-STR loci publication forthcoming in 2021. For a subset of these samples, STR sequence data has also been generated with the PowerSeq Auto-Y prototype assay (Promega, >600 samples) and/or the Precision ID GlobalFiler NGS STR Panel v2 (Thermo Fisher, >200 samples).

KCL: N=1100 samples were obtained from consenting adult volunteers. The samples relate to six different UK population groups with self-declared ancestries of: White British, West African, North East African, South Asian,

Chinese and Middle Eastern. All samples were sequenced with the ForenSeq system (Verogen) and additionally genotyped with at least two commonly available CE kits (Devesse 2020).

UNT: N=839 samples which have been described in associated sequence-based allele frequency publications and were sequenced with the ForenSeq system (Verogen) (Novroski 2016).

USC: N=944 samples from the HGDP-CEPH diversity panel cell-line DNAs from 51 diverse populations were sequenced with the ForenSeq system (Verogen) (Phillips 2018).

The association of a submitting laboratory with a record does not imply "discovery" of a sequence variant; rather the designation is simply the organization that initially provided the sequence and maintains the supporting data. Duplicate records were not created, resulting in a decreasing number of new sequence records as successive sample sets were added. Fig. 1 outlines an example submission strategy of non-duplicate allele records from a highly polymorphic STR, while Fig. 2 contextualizes the large sample numbers published by the STRSeq collaborators.

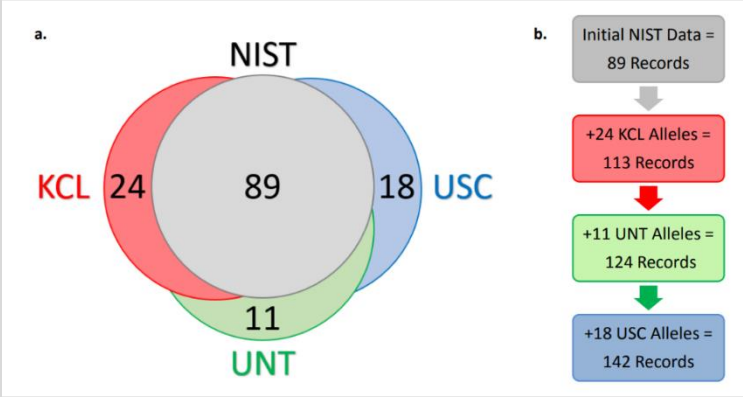

Figure 1. (left) a. Submission strategy for 142 unique sequences observed at the D12S391 locus. The unique sequences generated at NIST (N=89), KCL (N=24), UNT (N=11), and USC (N=18) form the basis of STRSeq records. b. Duplicate records are not created, resulting in a declining number of alleles submitted from each subsequent laboratory; however, the addition of the biogeographically diverse USC sample set yielded a relatively high number of new records.

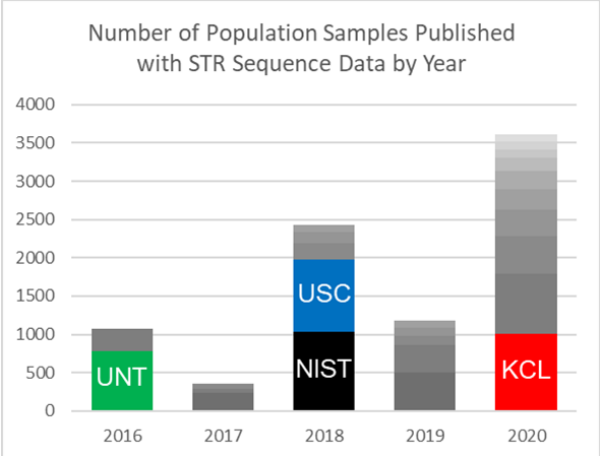

Figure 2. (right) Graph detailing number of population samples for which STR sequence data was published from 2016 to 2020 (2016 being the first year such studies were published due to availability of commercial STR sequencing assays). Data sets from STRSeq consortium labs are labeled accordingly, and represent the largest sequencing studies published during this timeframe. Additional studies are shown in grayscale and sorted by number of samples included. Data courtesy Jonathan King, UNT.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. 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).

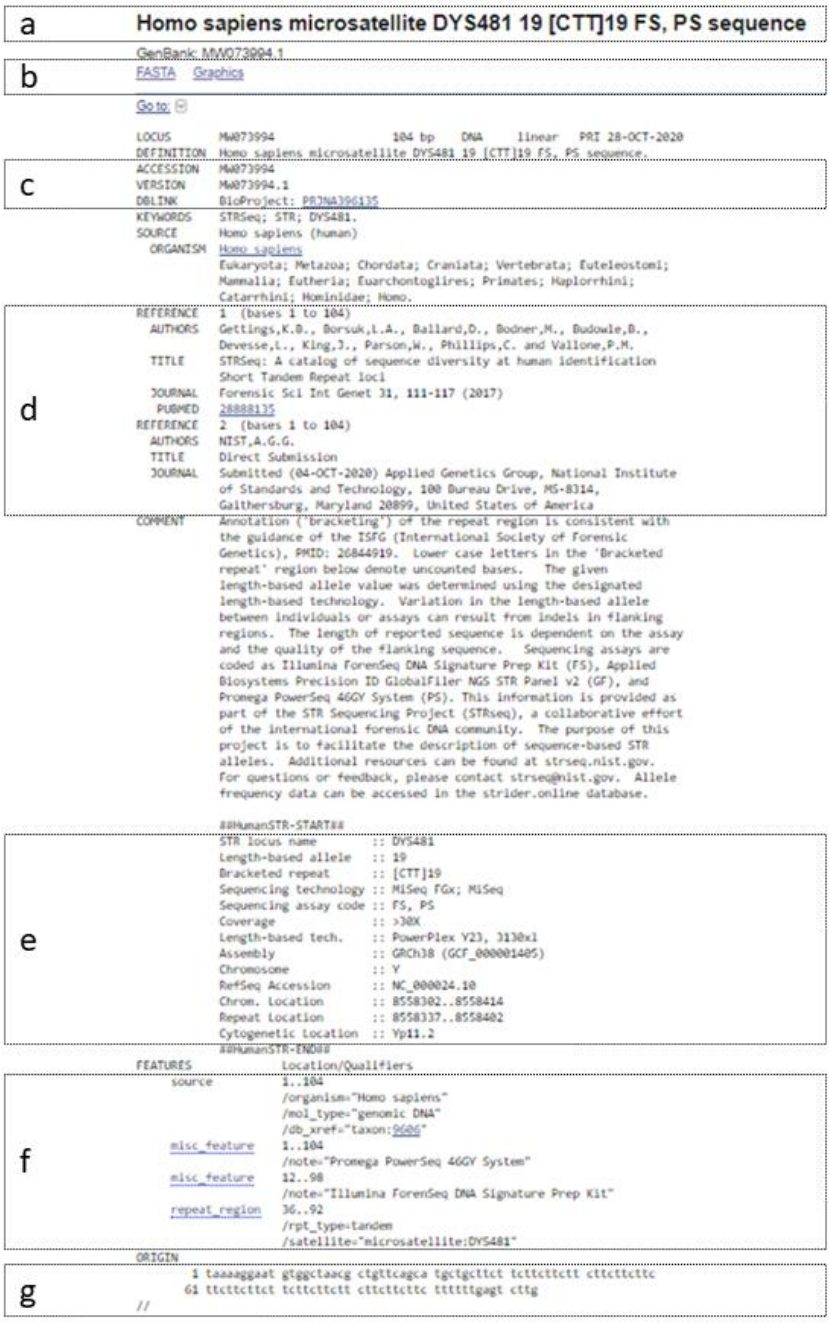

Figure 3. An example DYS481 STRSeq BioProject GenBank record. a) The Definition Line uniquely identifies each allele with components of the record. b) Hyperlinks to FASTA sequence and Graphics view (see Figure 4). c) GenBank sequence identifiers and link to the parent BioProject (e.g. DYS481), records 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).

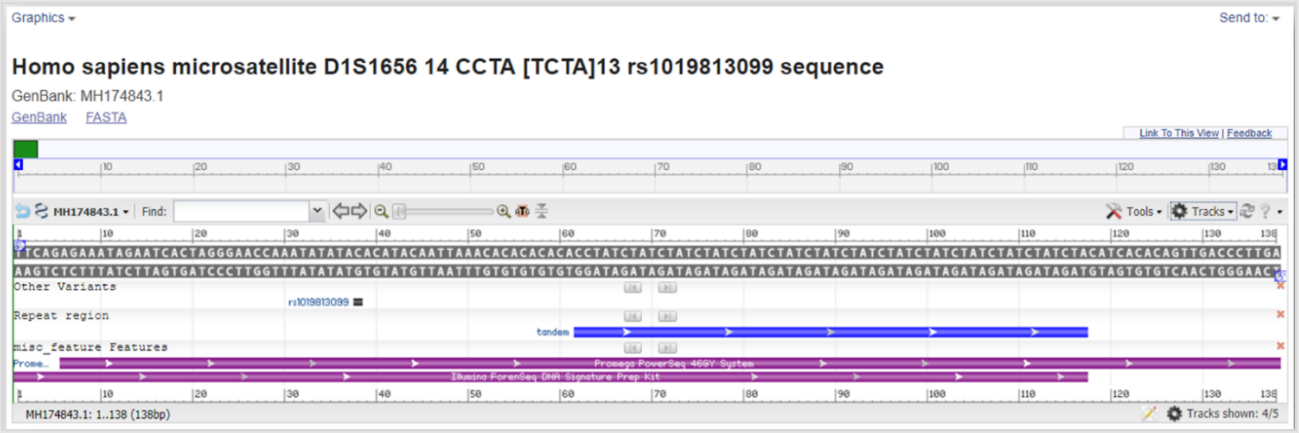

Figure 4. Graphics link in GenBank records presents an interactive version of the sequence with features from the record oriented onto GRCh38. An example from D1S1656 shows the repeat region (blue track), region(s) reported from available sequencing technologies (purple tracks), and associated flanking region polymorphisms (black square). More information and tutorials on the NCBI Sequence Viewer can be found at https:// www.ncbi.nlm.nih.gov/tools/sviewer.

Data analysis and findings

As previously described, sample sets and STRs are added iteratively, allowing the BioProject records to be released in phases. Therefore, while the GenBank records are stable, STRSeq should be viewed as a dynamic resource. Fig. 5 shows the cumulative number of STRSeq records submitted by quarter from 2017 to present, divided into locus categories. The earliest large submissions of commonly used autosomal STR loci correspond to the publication of the associated allele frequencies for the NIST "1036" population samples in 2018 (Gettings 2018, Borsuk 2018), with subsequent increases in this category resulting from the addition of collaborator data. More recently, X-STR and Y-STR records have been added, corresponding to the publication of the X-STR NIST population data in 2020 (Borsuk 2020) and preparation for the publication of the Y-STR NIST population data in 2021. These numbers are further broken down by locus in Fig. 6. With each new population study evaluated, it is expected that novel sequences will be observed, particularly at loci with complex sequence motifs, such as SE33 and D21S11. These increases will depend both on the number of individuals included in the study and the divergence of the population being studied from those already evaluated. However, it is expected that many of the less-polymorphic loci shown in Fig. 6 have plateaued, such that the addition of new records will be rare.

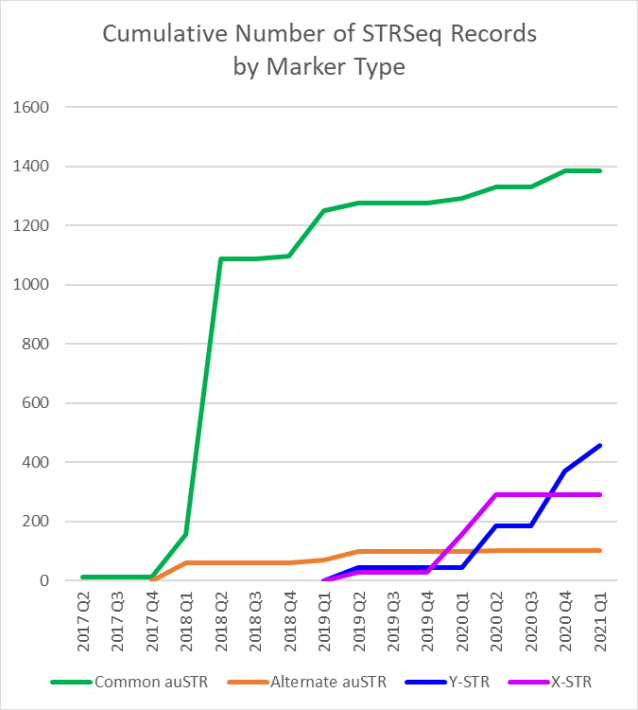

Figure 5 (above). Cumulative number of STRSeq records added to the BioProject categories of common autosomal STR loci, alternate autosomal STR loci, Y-STR loci and X-STR loci. Large increases correspond to publications of NIST population studies.

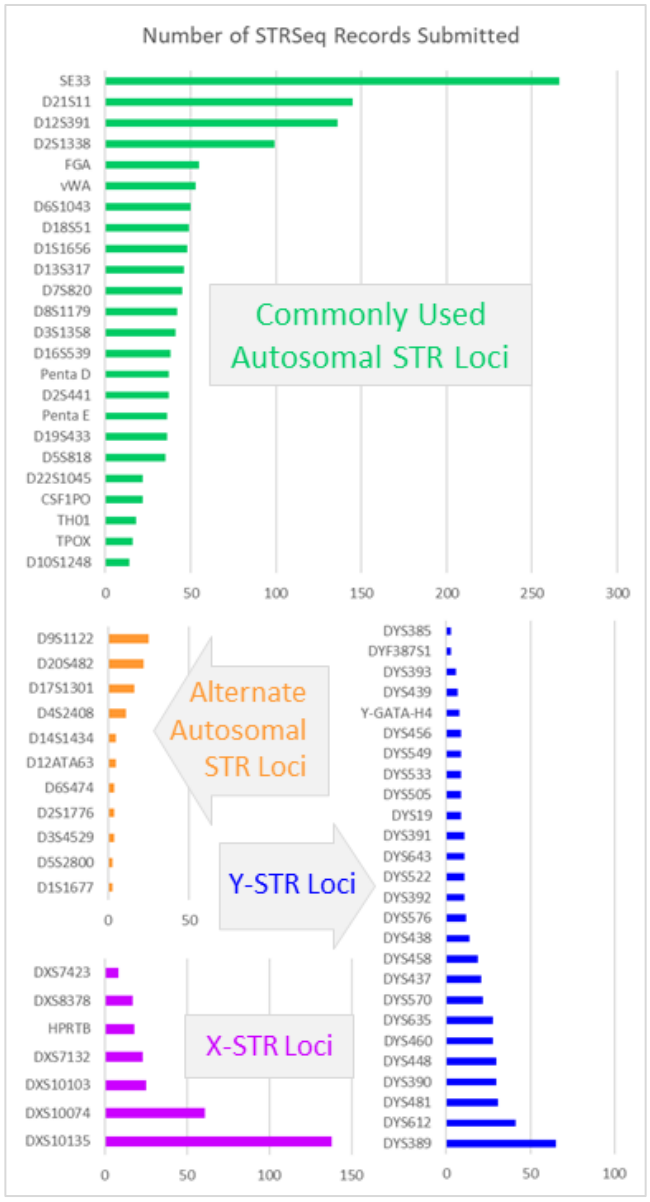

Figure 6 (right). Number of STRSeq records submitted by category and locus. Large increases are still expected for Y-STR loci as this NIST population data continues to be prepared for publication.

When this project was conceived, several use cases for STRSeq records were identified based on feedback from the forensic community: 1) As a teaching tool for forensic operational, academic, and commercial laboratories interested in sequencing STRs to explore STR sequences; 2) As the data backbone for development of STR sequencing methods and bioinformatic pipelines that conform to agreed variant data frameworks; 3) As a tool for the evaluation of rare sequences in casework or population sequencing studies, providing nomenclature information and leading the casework analyst to published allele frequency data.

While it is difficult to determine the level of uptake for all use cases, citation rates can provide a window into the successful uptake for certain use cases. Fig. 7 represents categorized citations of the original STRSeq publication

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. 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.

(Gettings 2017) by quarter, compared to the cumulative number of STRSeq records. While early citations consist of general references to the existence of this BioProject (e.g. in review articles) and self-citations by NIST or collaborators, more recent years show an increase in citations for both bioinformatic method development and evaluation of sequences in population studies.

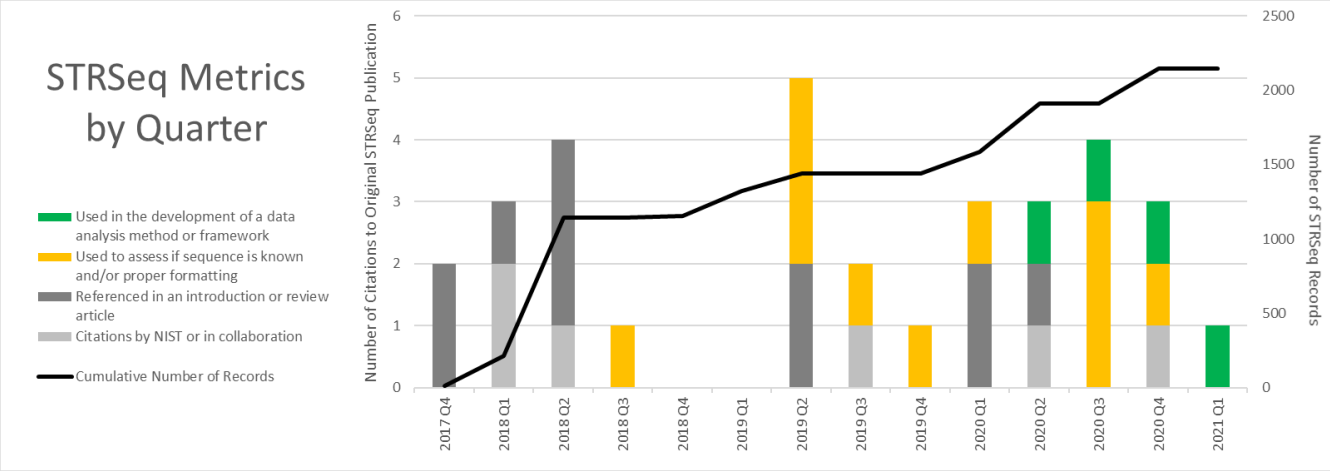

Figure 7. Cumulative number of STRSeq records by quarter (black line) compared to citations of the original STRSeq publication (32 citations, list available upon request). Yellow and green bars support successful uptake for identified use cases.

Future directions for STRSeq

In 2021, the International Society for Forensic Genetics convened a Commission on STR Nomenclature Recommendations, with a goal of finalizing a requirements document within a year. It is expected that the STRSeq BioProject will feature prominently, and continued efforts will be dedicated to updating STRSeq records according to these recommendations. The eventual goal for the STRSeq BioProject is to develop a pathway for submission of new sequence records from laboratories performing population sample sequencing. Currently, researchers wishing to publish STR population studies (CE or sequence based) in the journal *Forensic Science International: Genetics* are required to submit data files to STRidER (strider.online, Bodner 2016) for quality control evaluation (checks for duplicate samples, allele frequency percentages for a locus sum to 100%, etc.). Ideally, there will be an integrated, seamless process whereby users upload population sample sequencing data to STRidER for quality control, STRidER queries STRSeq for a matching sequence accession number, and the

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. 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.

STRSeq record formatting/nomenclature is compared to the laboratory submitted formatting/nomenclature. In cases where the STRidER query finds no match in STRSeq, a process would be initiated to evaluate the sequence (expanded checks for sequence coverage and range, polymorphisms in flanking regions, and phylogenetic context) and, for sequences which pass this evaluation, a new GenBank record would be created. Such a process would strengthen the STRidER quality control function and expand STRSeq, while harmonizing nomenclature between both resources. This is particularly important for novel sequence variants likely to be encountered as population studies extend their geographic scope and/or sample numbers. This collaborative effort will be the subject of a future funding proposal.

Implications for criminal justice policy and practice in the United States

It has now been over five years since the first commercially available assay for sequencing forensic STR markers was released; subsequently, the associated developmental validation was published in 2017 (Jager 2017), and length-based STR profiles generated from this assay are eligible for NDIS submission (FBI Laboratory 2021). Two additional STR sequencing assays have been released, with developmental validations underway. Probabilistic genotyping software vendors are currently evaluating CE-based models for suitability of use on STR sequence data (Cheng 2021), and multiple bioinformatic methods, including agnostic freeware (King 2021), are now available. Several U.S. forensic laboratories have brought sequencing of STR markers online for missing persons (e.g. CAL-DOJ, Ohio BCI) and forensic casework (e.g. DC DFS), in addition to a more widespread uptake in the European forensic community. While this shift may be more subtle than some had originally hoped, and the technology may continue to serve a complementary role to traditional CE testing, it is crucial that laboratories using STR sequencing technology have a standard format for reporting and exchanging data which maintains backward-compatibility with CE profiles. The STRSeq BioProject has now matured to a level where it can serve as the resource to provide this standard format.

References

A. Alonso, P. Muller, L. Roewer, S. Willuweit, B. Budowle, W. Parson, European survey on forensic applications of massively parallel sequencing, Forensic Sci Int Genet 29 (2017) e23-e25.

M. Bodner, I. Bastisch, J.M. Butler, R. Fimmers, P. Gill, L. Gusmao, N. Morling, C. Phillips, M. Prinz, P.M. Schneider, W. Parson, 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.

L.A. Borsuk, K.B. Gettings, C.R. Steffen, K.M. Kiesler, P.M. Vallone, Sequence-based US population data for the SE33 locus, Electrophoresis 39(21) (2018) 2694-2701.

L.A. Borsuk, C.R. Steffen, K.M. Kiesler, P.M. Vallone, K.B. Gettings, Sequence-based U.S. population data for 7 X-STR loci, Forensic Sci Int Reports (2020) 100160.

K. Cheng, M.H. Lin, L. Moreno, J. Skillman, S. Hickey, D. Cuenca, W.R. Hudlow, R. Just, J. Bright, J. Buckleton, J.M. Curran, Modeling allelic analyte signals for aSTRs in NGS DNA profiles, J Forensic Sci 00 (2021) 1-12.

L. Devesse, L. Davenport, L. Borsuk ,K. Gettings, G. Mason-Buck, P.M. Vallone, D. Syndercombe Court, D. Ballard, Classification of STR allelic variation using massively parallel sequencing and assessment of flanking region power, Forensic Sci Int Genet 48 (2020) 102356.

FBI Laboratory, National DNA Index System (NDIS) Operational Procedures Manual, Version 9 Rev2 (accessed from https://www.fbi.gov/file-repository/ndis-operational-procedures-manual.pdf/view on March 11, 2021).

K.B. Gettings, R.A. Aponte, P.M. Vallone, J.M. Butler, STR allele sequence variation: Current knowledge and future issues, Forensic Sci Int Genet 18 (2015) 118-30.

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

K.B. Gettings, L.A. Borsuk, C.R. Steffen, K.M. Kiesler, P.M. Vallone, Sequence-based U.S. population data for 27 autosomal STR loci, Forensic Sci Int Genet 37 (2018) 106-115.

C.R. Hill, D.L. Duewer, M.C. Kline, M.D. Coble, J.M. Butler, U.S. population data for 29 autosomal STR loci, Forensic Sci Int Genet 7(3) (2013) e82-3.

C.R. Hill, M.C. Kline, M.D. Coble, J.M. Butler, Characterization of 26 MiniSTR Loci for Improved Analysis of Degraded DNA Samples, J Forensic Sci 53(1) (2008) 73-80.

A.C. Jager, M.L. Alvarez, C.P. Davis, E. Guzman, Y. Han, L. Way, P. Walichiewicz, D. Silva, N. Pham, G. Caves, J. Bruand, F. Schlesinger, S.J. Pond, J. Varlaro, K.M. Stephens, C.L. Holt, Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories, Forensic Sci Int Genet 28 (2017) 52-70.

J.L. King, A.E. Woerner, S.N. Mandape, K.B. Kapema, R.S. Moura-Neto, R. Silva, B. Budowle, STRait Razor Online: An enhanced user interface to facilitate interpretation of MPS data, Forensic Science Int Genet 52 (2021) 102463.

N.M. Novroski, J.L. King, J.D. Churchill, L.H. Seah, B. Budowle, Characterization of genetic sequence variation of 58 STR loci in four major population groups, Forensic Sci Int Genet 25 (2016) 214-226.

W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, L. Gusmao, D.R. Hares, J.A. Irwin, J.L. King, P. Knijff, N. Morling, M. Prinz, P.M. Schneider, C.V. Neste, S. Willuweit, C. Phillips, Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements, Forensic Sci Int Genet 22 (2016) 54-63.

C. Phillips, L. Devesse, D. Ballard, L. van Weert, M. de la Puente, S. Melis, V. Alvarez Iglesias, A. Friere-Aradas, N. Oldroyd, C. Holt, D. Syndercombe Court, A. Carracedo, 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 39(21) (2018) 2708-2724.

C.M. Ruitberg, Reeder, D.J., and J.M. Butler, STRBase: A short tandem repeat DNA database for the human identity testing community, Nucleic acids research 29(1) (2001) 320-322.

T.Z. Willems, D.; Yuan, J.; Gordon, A.; Gymrek, M.; and Y. Erlich, Genome-wide profiling of heritable 228 and de novo STR variations, Nat Methods 14(6) (2017) 590-592.

A.E. Woerner, J.L. King, B. Budowle, Fast STR allele identification with STRait Razor 3.0, Forensic Science International: Genetics (2017).