Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on short tandem repeat sequence nomenclature

Gettings, Katherine B.; Bodner, Martin; Borsuk, Lisa A.; King, Jonathan L.; Ballard, David ...

2024-01

Elsevier Ireland Ltd.

http://hdl.handle.net/10138/566895

Gettings, K B, Bodner, M, Borsuk, L A, King, J L, Ballard, D, Parson, W, Benschop, C C G, Børsting, C, Budowle, B, Butler, J M, van der Gaag, K J, Gill, P, Gusmão, L, Hares, D R, Hoogenboom, J, Irwin, J, Prieto, L, Schneider, P M, Vennemann, M & Phillips, C 2024, 'Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on short tandem repeat sequence nomenclature', Forensic Science International: Genetics, vol. 68, 102946. https://doi.org/10.1016/j.fsigen.2023.102946

Forensic Science International: Genetics 68 (2024) 102946

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on short tandem repeat sequence nomenclature

Katherine B. Gettings [a,*], Martin Bodner [b], Lisa A. Borsuk [a], Jonathan L. King [c], David Ballard [d], Walther Parson [b,e], Corina C.G. Benschop [f], Claus Børsting [g], Bruce Budowle [h,i], John M. Butler [a], Kristiaan J. van der Gaag [f], Peter Gill [j], Leonor Gusmão [k], Douglas R. Hares [l], Jerry Hoogenboom [f], Jodi Irwin [l], Lourdes Prieto [m,n], Peter M. Schneider [o], Marielle Vennemann [p], Christopher Phillips [m]

[a] National Institute of Standards and Technology, Gaithersburg, MD, USA
[b] Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria
[c] Center for Human Identification, University of North Texas Health Science Center, Fort Worth, TX, USA
[d] King's Forensics, Department of Analytical, Environmental and Forensic Sciences, King's College London, London, United Kingdom
[e] Forensic Science Program, The Pennsylvania State University, University Park, PA, USA
[f] Division of Biological Traces, Netherlands Forensic Institute, The Hague, the Netherlands
[g] Section of Forensic Genetics, Department of Forensic Medicine, University of Copenhagen, Denmark
[h] Department of Forensic Medicine, University of Helsinki, Helsinki, Finland
[i] Radford University Forensic Science Institute, Radford University, Radford, VA, USA
[j] Forensic Genetics Research Group, Oslo University Hospital, Oslo, Norway
[k] DNA Diagnostic Laboratory, State University of Rio de Janeiro, Rio de Janeiro, Brazil
[l] FBI Laboratory, Quantico, VA, USA
[m] Forensic Sciences Institute Luis Concheiro. University of Santiago de Compostela, Santiago de Compostela, Spain
[n] Comisaría General de Policía Científica, Madrid, Spain
[o] Institute of Legal Medicine, University of Cologne, Cologne, Germany
[p] Institute of Legal Medicine, University of Münster, Münster, Germany

ARTICLE INFO

Keywords:
Short tandem repeat
DNA sequencing
Nomenclature
Quality control
Forensic genetics

ABSTRACT

The DNA Commission of the International Society for Forensic Genetics (ISFG) has developed a set of nomenclature recommendations for short tandem repeat (STR) sequences. These recommendations follow the 2016 considerations of the DNA Commission of the ISFG, incorporating the knowledge gained through research and population studies in the intervening years. While maintaining a focus on backward compatibility with the CE data that currently populate national DNA databases, this report also looks to the future with the establishment of recommended minimum sequence reporting ranges to facilitate interlaboratory comparisons, automated solutions for sequence-based allele designations, a suite of resources to support bioinformatic development, guidance for characterizing new STR loci, and considerations for incorporating STR sequences and other new markers into investigative databases.

1. Introduction

In 2016, the DNA Commission of the International Society for Forensic Genetics (ISFG) outlined eight considerations for early adopters of short tandem repeat (STR) massively parallel sequencing (MPS) technologies [1], summarized as: *1) STR sequence analysis software should allow export/storage of sequence strings; 2) Sequences should be aligned to the forward strand of the reference genome; 3) The current reference genome GRCh38 should be used; 4) Loci historically reported on the reverse strand need to be translated to the forward strand, with defined anchor points; 5) Early adopters should employ comprehensive nomenclature to ensure compatibility with a future nomenclature system, and maintain backward compatibility to the repeat-based nomenclature derived from Capillary Electrophoresis (CE); 6) Stored STR sequence strings should*

* Correspondence to: Applied Genetics Group, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA.
E-mail address: katherine.gettings@nist.gov (K.B. Gettings).

https://doi.org/10.1016/j.fsigen.2023.102946
Received 27 September 2023; Accepted 14 October 2023
Available online 18 October 2023

include flanking sequence and genomic start/stop nucleotide coordinate metadata should be maintained; 7) Sequence-based worldwide STR allele frequency databases using a unified nomenclature system are necessary to take full advantage of the increased power of discrimination offered by MPS generated STR data; and 8) Future STR sequencing multiplexes should retain past STR markers, new marker selection should incorporate sequence-based population data.

In the years following the 2016 DNA Commission report, more laboratories in the forensic DNA community have evaluated and applied the potential of sequencing STRs to improve human identification. Multiple commercial sequencing kits targeting autosomal, Y-, and/or X-STR loci are currently available and have been well characterized in the literature (including [2–8] and reviewed in [9,10]). Kit-specific and agnostic (*i.e.*, not specific to a kit) bioinformatic methods have been developed, both commercial and open-source options (reviewed in [11]). Additionally, numerous population studies have been performed to-date to characterize the allelic diversity in these regions of the human genome.

Following the publication of the 2016 DNA Commission report, a subset of the authors representing five laboratories formed an *ad hoc* working group and began collaborating to harmonize STR nomenclature-related efforts across respective laboratories, listed here and detailed below: the Forensic Sequence STRucture Guide (FSSG) for genome reference and bracketing guidance [12], the STR Sequencing Project (STRSeq) catalog of sequences [13], and the STRs for identity European Network of Forensic Science Institutes (ENFSI) Reference database (STRidER) for sequence quality control [14,15]. To address the more broadly reaching issue of STR sequence nomenclature, this group of researchers formalized in 2018 as the STRAND Working Group (Short Tandem Repeat: Align, Name, Define) and subsequently received the endorsement of the ISFG Executive Board to organize an STR sequence nomenclature meeting. This meeting was held in April 2019, with 26 researchers in attendance representing eight countries. Attendees included researchers developing STR sequence-based nomenclature schemata, scientific representatives from vendors developing STR sequence bioinformatic methods, DNA investigative database curators, and academic experts in STR genomics. Many of the topics and concepts included in this Commission report were discussed during this meeting and introduced to the larger forensic community in the resulting STRAND Meeting report [16]. Since this 2019 meeting, the five laboratories comprising the STRAND Working Group have continued to coordinate development of their respective resources, as described here.

The Forensic Sequence STRucture Guide (FSSG) [12] contains human genome reference sequences for forensically relevant autosomal STR, Y-STR, and X-STR loci, annotated with bracketing following the 2016 ISFG guidance [1] and placement of nearby flanking region polymorphisms. The curated, comprehensive file is hosted at https://strider.online/nomenclature together with an update log recording changes made since the original version was published. A parallel version of this file focusing on commercial STR sequencing kit ranges was published as a supplemental file to the 2019 STRAND Meeting report [16]. Subsequently, minor changes in kit sequence ranges were needed based on vendor information; therefore, up-to-date kit ranges have now been merged into the FSSG.

The STR Sequencing Project (STRSeq) [13] consists of a curated catalog of sequence diversity at forensic STR loci, along with key elements of nomenclature, originally designed to conform with the 2016 ISFG DNA Commission's considerations [1]. The initial data used to populate STRSeq are the aggregate alleles observed in targeted sequencing studies comprising > 4500 single-source samples (one GenBank record is created for each unique STR sequence observed among these samples). The STRSeq "BioProject" serves to organize these records within the GenBank repository (https://www.ncbi.nlm.nih.gov/bioproject/380127) and is divided into categories: 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. The sequence records in GenBank are flat files of specified format, such that they can be downloaded and parsed *en masse* or explored via an interactive graphic. Additionally, application programming interfaces (API) are available, which facilitate the creation of other resources leveraging these data (*e.g.*, bioinformatic pipelines). Answers to commonly asked questions about the STRSeq BioProject have recently been published [17].

The STRs for identity ENFSI Reference database (STRidER) is a freely accessible online forensic autosomal STR frequency database and platform for interpretation and quality control (QC) of such data, and rarity estimation of genotypes. Observations of erroneous STR allele frequency data in the literature led to launching STRidER in 2017 with the aim to reduce such problematic data [14]. STRidER builds upon the ENFSI STRbase (2004–2016) [18,19] that was further developed with endorsement by the ISFG, where the two forensic databases EMPOP [20] and YHRD [21] for mitochondrial DNA and Y-chromosomal haplotypes, respectively, had previously shown the importance of QC in haploid DNA marker data generation and databasing [22] [23]. Though the need for autosomal STR data QC had been previously demonstrated (*e.g.*, [24–27]), prior to STRidER, no publicly available pre-publication QC for autosomal STR allele frequency data existed. Now, QC is mandatory before allele frequency data upload to the online database and publication of associated articles in the ISFG-endorsed journals of FSI: Genetics [23] and FSI: Reports. STRidER has received over 300 autosomal STR population datasets since its inception and evaluated their quality before publication. At the time of writing, 43 of the datasets contained STR sequence data for a total of > 11,000 genotypes. STR sequence data are submitted to STRidER as nucleotide string files and scrutinized by a suite of software tools. QC is optimized for the detection of common discrepancies, idiosyncrasies and implausibility in the data. A broad range of error types and rates is encountered [15].

In parallel to the efforts of the STRAND working group, several additional resources have been developed to address aspects of a sequence-based forensic STR nomenclature. In 2020, a method for converting DNA sequences into sequence identifier (SID) codes was published [28]. The method uses the hash function SHA-256 to convert a DNA sequence string into a 50+ character SID which can be truncated, depending on the application. Additionally, in 2021, an algorithm called STRNaming was made publicly available, and a corresponding manuscript was published [29]. This program standardizes and automates the conversion of an STR string into a bracketed format based on a defined set of parameters, which were informed by STR sequencing-user surveys. Similar to genomic sequence alignment methods, points are assigned for desirable features (*e.g.*, length of repetitive run), and penalties are levied for undesirable features (*e.g.*, introduction of gaps).

In 2021, the ISFG convened a DNA Commission on sequence-based STR nomenclature, composed of researchers and operational laboratory representatives with relevant subject matter expertise. The charge of this Commission, which built upon STRAND and parallel STR nomenclature activities, was to review the 2016 Commission's considerations in light of the information gained over the intervening five years and to finalize a set of recommendations. The results of this Commission are described herein, including five recommendation topics: 1) STR sequence string alignment and minimum reporting range, 2) STR sequence visualization (bracketing) and sequence codes, 3) resources for recommended format and allele frequency data, 4) resources for characterization of new STR loci, and 5) databasing considerations.

2. Recommendations

2.1. Sequence Strings

A forensic laboratory generating STR allele sequences may convey results in a variety of ways in reports, database records, and/or publications. Additional recommendations in this document address formats which may be more user-friendly and databasing software-friendly than

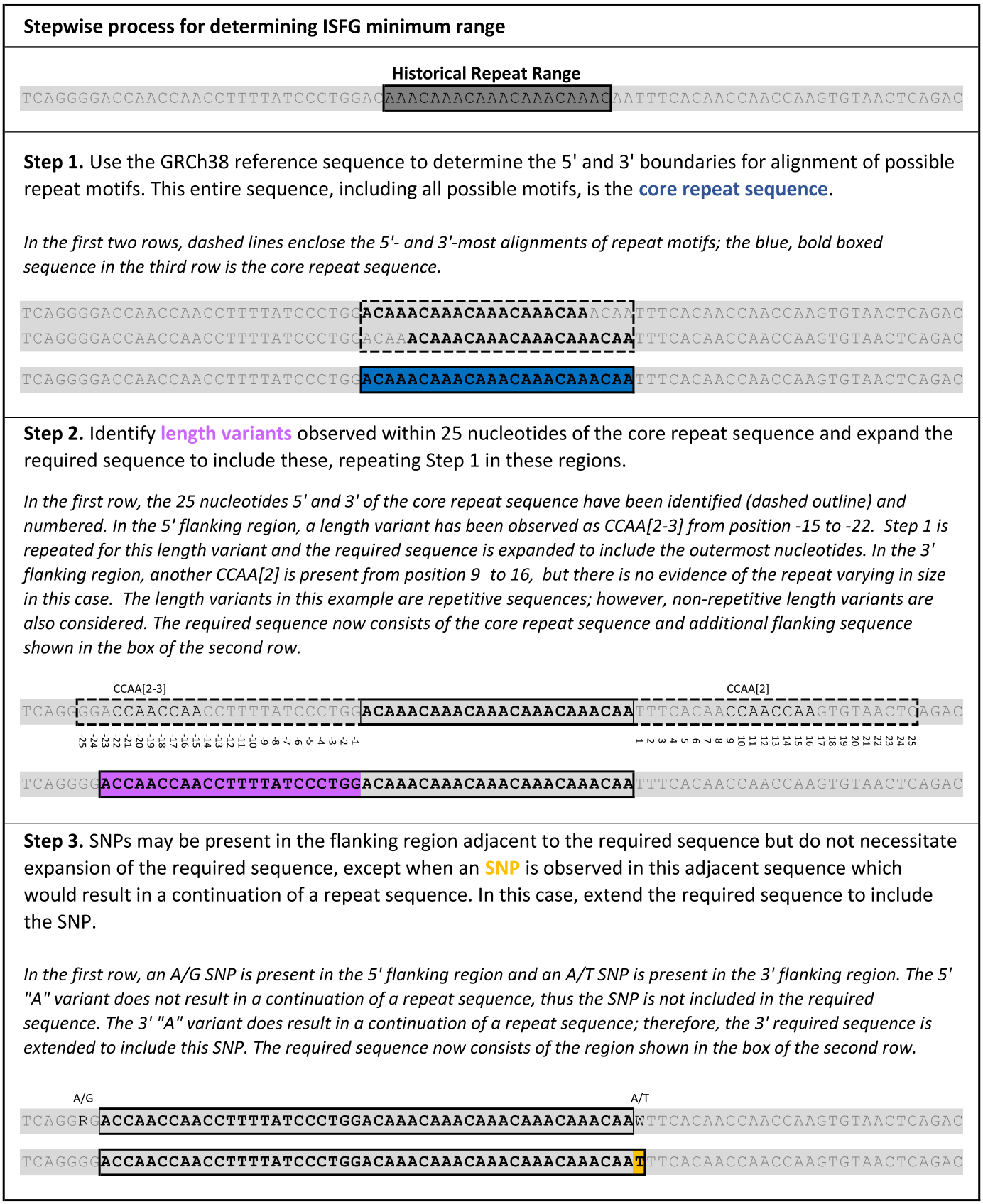
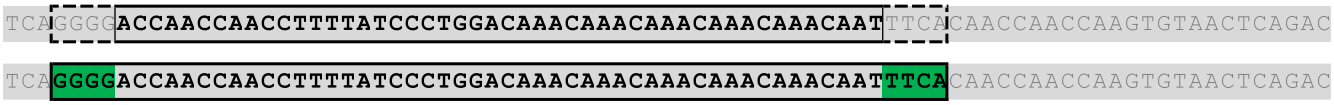


Fig. 1. The stepwise process for determining the ISFG minimum range, shown on an idealized locus created to incorporate examples of each step in the process. Each row of sequence shown is identical throughout the figure. Bottom-most is the final ISFG Minimum Range (bolded, boxed) aligned with the reportable sequence ranges of four commercial kits with varying overlap (note that the Kit 4 sequence range does not meet the ISFG minimum range). Color in online version only.

Stepwise process for determining ISFG minimum range (continued)

Step 4. Add a four-nucleotide **cap** to the 5' and 3' ends of the required sequence. The sequence within and including the 5' and 3' caps is the ISFG minimum range, barring any additional expansion required for parameterized bracketing.

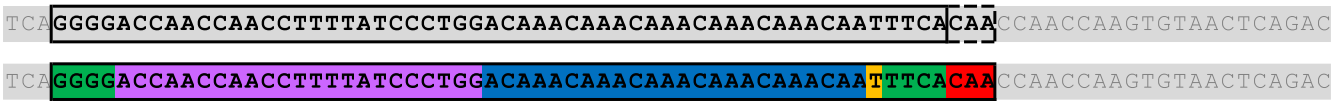
In the first row, 5' and 3' caps are added (dashed outline) and the resulting required sequence (second row, box) is processed with the parameterized bracketing algorithm, yielding the result shown. Lower case nucleotides indicate the algorithm has inferred nucleotides from the reference sequence.



Parameterized Bracketing Result: **ACAA[6]TTTC[1]Acaa[1]**

Step 5. Evaluate the **parameterized bracketing** result to determine whether the algorithm incorporates nucleotides beyond the required sequence due to the presence of a repeat unit. In such cases, the required sequence is expanded to complete this repeat unit. This new range is the ISFG minimum range.

In the first row of the example, the dashed outline indicates three nucleotides outside of the required sequence determined in Step 4, but included in the Acaa[1] shown in the parameterized bracketing result of Step 4. To prevent the algorithm from inferring reference sequence nucleotides, the required sequence is expanded to include these three additional nucleotides. The final ISFG minimum range is shown in the second row, with the associated parameterized bracketing result below.



Parameterized Bracketing Result: **ACAA[6]TTTC[1]ACAA[1]**

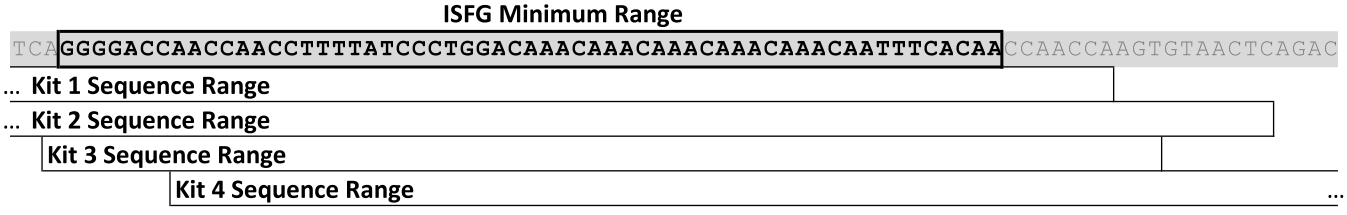


Fig. 1. (*continued*).

sequence strings. Regardless of how the sequence information is conveyed, the originating laboratory should maintain a record of the sequence string. Additionally, a standard (minimum) reporting range is useful to provide sufficient flanking region to distinguish the termini of the core repeat sequence, which determines the numerical STR allele designation, and to unify results across kits, software, and laboratories.

Recommendation 1:. *Sequenced STR alleles should be maintained as sequence strings, oriented to the forward strand of the current human genome assembly. Sequence strings should include the genomic coordinates of the minimum reporting range defined herein.*

2.1.1. Human Genome Assembly

The current human genome assembly is GRCh38 and the Genome Reference Consortium (GRC) regularly releases patches which do not affect genomic coordinates (*e.g.,* GRCh38.p14). At this time, the GRC has indefinitely postponed the next coordinate-changing update (*i.e.,* GRCh39; for more information, see https://www.ncbi.nlm.nih.gov/grc/human). If the GRC publishes a coordinate-changing build in the future, we anticipate this Commission will reconvene to evaluate the impact of the changes and provide guidance on whether the forensic community should migrate to the new coordinate system. Online tools are readily available for remapping genomic coordinates between builds (https://www.ncbi.nlm.nih.gov/genome/tools/remap) and multiple coordinate sets are already provided in resources specific to the forensic community

(*e.g.*, in the FSSG).

2.1.2. ISFG Minimum Range

During the course of this Commission, steps for determining a standard (minimum) range were established (see below). The Commission has applied these steps to the STR loci in current use, and the implementation of this standard (minimum) range (hereafter referred to as "ISFG minimum range") is further described in subsequent recommendations. It is important to note that the following steps are applied *one time only* for a locus so that the range will remain static for bioinformatic and databasing purposes. Through the course of this Commission, the ISFG minimum range has been established for many STR loci; this range is included in files associated with Recommendation 3. The steps for determining the ISFG minimum range are described here to exemplify the logic and to aid researchers in evaluating new STR loci (the process of which is further described in Recommendation 4).The following steps are applied sequentially to determine the ISFG minimum range, as exemplified on a fictitious locus in Fig. 1.

Step 1 Consider whether ambiguity exists regarding the start and stop of the repetitive sequence used to determine the numerical STR allele designation (the initial "CE-counted repeat"), depending on which repeat motif is chosen. For such loci, use the GRCh38 reference sequence to determine the 5′ and 3′ boundaries for alignment of possible repeat motifs. This entire sequence, including all possible motifs, is the "core repeat sequence".

Step 2 The numerical STR allele designation is affected by insertions or deletions contained in the flanking regions of the targeted sequence.

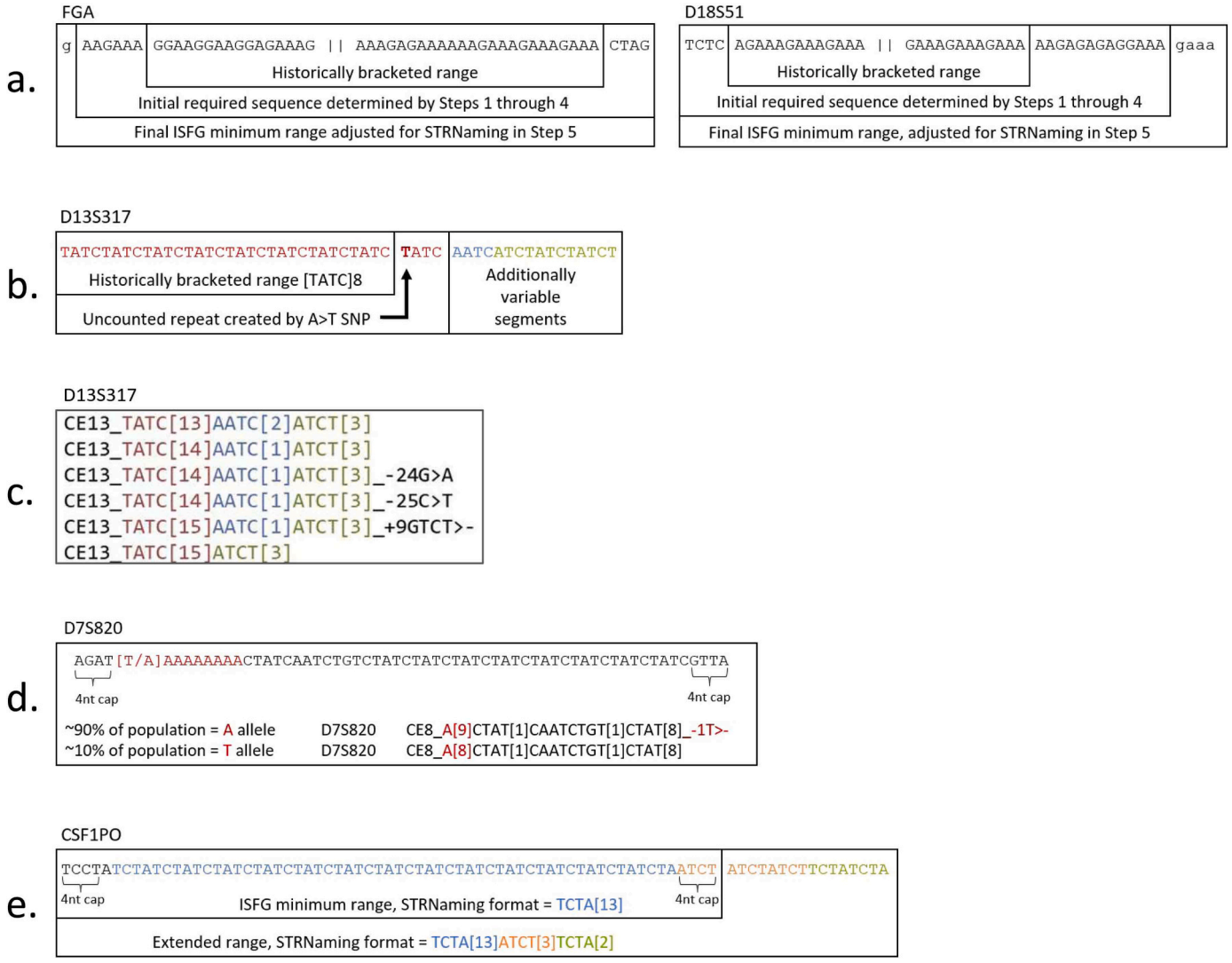


Fig. 2. a. Following Step 5, the recommended minimum ranges for FGA and D18S51 were adjusted to prevent STRNaming from including inferred nucleotides, shown lower case in this figure. "||" represent repeat units removed for display purposes. b. D13S317 has a common A>T SNP (shown as a bolded T, rs9546005) in the 3′ flanking region that results in an additional TATC repeat which is not counted toward the length-based allele designation and has historically been excluded from bracketing. Color coding differentiates repeat motifs identified by STRNaming. c. Examples of STRNaming output for various D13S317 alleles. The length-based allele number is explicitly represented (CE13); however, this number does not always match the number of TATC repeats. Flanking region polymorphisms are suffixed; identified by the change from and oriented relative to the GRCh38 reference sequence. d. D7S820 has a common SNP in the 5′ flanking region, rs7789995, shown as [T/A] in this figure. Approximately 90% of global population chromosomes contain the A allele in this position, whereas the GRCh38 reference genome contains the T allele. Because STRNaming designates flanking region variants relative to GRCh38, it reports the major allele A as a variant. Note that the A allele is reported as an additional ninth A in a homopolymer stretch, and the "missing" T allele is suffixed as "– 1 T > -". e. CSF1PO sequence exemplifies submitting a range of sequence larger than the ISFG minimum range into STRNaming v1.2 may result in bracketing of additional repetitive units. The ISFG minimum range is not expanded to include these repeats; this is shown for user awareness only. Color coding differentiates repeat motifs identified by STRNaming. See text for more details. Color in online version only.

Identify reported length variants in the nearby flanking region and determine the 5′ and 3′ boundaries for alignment of the length variants. If the length variant alignment boundary begins within 25 nucleotides of the core repeat sequence, expand the required sequence to encompass this flanking variation.

Step 3 The numerical STR allele designation is usually unaffected by SNPs contained in the flanking regions of the targeted sequence (rare instances of CE mobility shifts have been attributed to SNPs [30,31]); therefore, SNPs are not generally considered in determining the required minimum sequence. When SNPs are present in the flanking region adjacent to the core repeat sequence, consider whether the variant SNP alleles can result in a continuation of the core repeat sequence. In such cases, extend the required minimum sequence to include the SNP.

Step 4 Add a four-nucleotide cap to both the 5′ and 3′ ends of the sequence determined in Steps 1 through 3. These caps are included in the required minimum sequence to distinguish the boundaries of the elements defined in Steps 1 through 3, which generally determine the numerical STR allele designation.

Step 5 The final step in determining the ISFG minimum range includes evaluation of the parameterized bracketing "STRNaming" [29], further described in Recommendation 2. Briefly, this method relies on the automated application of a set of parameters to a reference sequence to determine a bracketed motif, which then is applied to sample sequences. In some cases, the resulting bracketed repeat motif may include additional, adjacent nucleotides beyond the minimum sequence determined in Steps 1 through 4. In such instances, STRNaming fills in (infers) these "missing nucleotides" from the reference sequence. Among the 35 autosomal STR loci currently included in sequence-based commercial kits, the inference of missing nucleotides occurred in two instances (see Fig. 2a). To avoid the addition of inferred sequence into the bracketed repeat, such adjacent nucleotides are included in the final ISFG minimum range.

Notes on *Step 2* and *Step 3*: The STRSeq catalog [13] as it existed at the end of calendar year 2021 was used for evaluating length variants and SNPs for the STR loci largely covered by the > 4500 population samples included at that time. For the purposes of applying this rule set to STR loci in current use (defined in the resources provided in Recommendation 3), a length variant or SNP was required to have been observed in two different samples/individuals: either as multiple observations in one or more forensic population sequence study(ies) (as submitted to STRSeq [13] through 2021) or one observation in a forensic population sequence study (in STRSeq through 2021) along with a confirming rs-number record in dbSNP build 155 [32] and/or observation in gnomAD version 3.1.2 [33]. Nine STR loci are currently included in only one commercial STR sequencing kit and were not well-covered in STRSeq population studies by the end of 2021: D1S1677, D2S1776, D3S4529, D5S2800, SE33, D6S474, D12ATA63, D14S1434, and DYS627. For the purpose of applying this rule set to these and future loci, a resource other than STRSeq must be applied in order to provide timely guidance. In such cases, gnomAD version 3.1.2 was used, and two observations with a minimum frequency of 0.02% (frequency level chosen to approximate STRSeq sample numbers) was required.

Additionally, the distance of 25 nucleotides in Step 2 was chosen operationally because this distance generally encompassed the flanking region length variation observed in the STRSeq catalog for autosomal loci in current STR sequencing kits [13]. This distance is expected to help kit designers avoid placing PCR primers in regions that contain length variants, which may cause null alleles and/or CE discordances. Finally, SNPs and Indels in the flanking region combined with the STR sequence form a haplotype (meaning they are expected to be inherited together); in this document the flanking region polymorphisms are described separately from the STR sequence for clarity.

A note on sequence strings: Any reported sequence should consist entirely of sequenced nucleotides from the analyzed sample. When existing STR sequencing kits cannot currently meet the recommended minimum reporting range (*e.g.,* PCR primer placement overlaps the minimum reporting range and/or the core repeat), the kit should be redesigned to meet this recommendation. In the interim, it is recommended that laboratories report the entire ISFG minimum range and clearly denote *inferred* nucleotides and their source (*i.e.,* GRCh38). Affected loci and nucleotide positions in current commercial STR sequencing kits are identified in the FSSG, further described in Recommendation 3.

2.2. Sequence Proxies

While the sequence string is the primary format for maintaining STR sequence data, other shorthand formats or "sequence proxies" may be useful for reporting, databasing, presenting/describing data, and bioinformatic software analyses. The 2016 DNA Commission report exemplified a comprehensive nomenclature for early adopters of STR sequencing methods; now, more streamlined formats are available.

Recommendation 2:. *To provide a unified system with minimal human intervention, STRNaming [29] should be used for bracketed repeat formatting. Additionally, any system of sequence codes should be generated based on the ISFG minimum range if these codes are employed for general inter-laboratory comparisons.*

2.2.1. Bracketed Repeat. For condensing the repeat region of a sequence string into a descriptive, "human readable" format, the so-called bracketed repeat is useful for reporting and other applications, such as the characterization of stutter (as shown in [34]). Historically, the original publications characterizing the STR regions for forensic use defined this format, in which the repeat region of the sequence is represented by the repetitive pattern of a motif and the number of repeats. Efforts were made to standardize the start/stop positions and inclusion/exclusion of neighboring repetitive elements on a per-locus basis [35–40]; however, many exceptions exist either due to historical legacy (the locus was characterized before guidance was published) or the inability of a rule set to encompass all scenarios [1,41].

In the simplest cases, the bracketed sequence encompassed the start/stop points of the "counted" repeat region (*i.e.*, the repetitive sequence counted toward the length-based, numerical allele designation). This approach maximized the ability to visually discern the numerical allele designation from the bracketed repeat; however, some loci are challenging. For example, a common SNP in the 3′ flanking region of D13S317 creates the appearance of an additional repeat which has historically been excluded from bracketing, instead it was represented as part of the flanking region.

In addition, from a practical point of view, this lack of a unified rule set hampers developers from automating STR sequence bracketing, instead requiring a look-up database and manual formatting/curation when alleles novel to the database are encountered. This system introduces the possibility of variable approaches among users when sequences are not present in a database and multiple bracketing options are possible, particularly at more complex loci such as D21S11 or SE33. A goal of this Commission was to determine a unified system of bracketing that could be implemented with minimal human intervention. The publication of the STRNaming program [29] largely fulfilled this goal.

Following the 2021 STRNaming publication and throughout the course of this Commission, the developers have further evaluated settings and solicited additional feedback. With the recent publication of STRNaming version 1.1 [42], the developers have established universal parameters which optimize arrangements of the repeat region structures across most of the STR loci in the human genome. These changes, along with a minor update to version 1.2, have already been incorporated into the python packages FDSTools (seqconvert) and STRNaming (further

explained on fdstools.nl), and the interactive version of the software (fdstools.nl/strnaming) is expected to include these changes (*i.e.*, output v1.2) within the second half of 2023. No additional changes which would affect the bracketing are expected at this time. If a future version of STRNaming results in changes to STR locus bracketing (as compared to version 1.2 output used in developing these recommendations), it is anticipated that this Commission will reconvene, coordinate with the developers to evaluate the impact of the changes, and provide guidance on whether the forensic community should migrate to the new software version (and if so, coordinate with the developers on a transition plan). Additionally, any such changes will be registered in the FSSG.

A series of figures (Figs. 2a to 2e) are included in this document to exemplify various phenomenon which arise from applying the STRNaming method of bracketing sequences. Fig. 2a shows two loci in which STRNaming would infer nucleotides from the reference sequences, if they were not included in the input sequence; thus, the ISFG minimum range is expanded to include these nucleotides for these two loci. Fig. 2b shows a common flanking region SNP resulting in the appearance of an additional repeat, which was excluded from historical bracketing and is included in STRNaming bracketing. Fig. 2c gives examples of STRNaming outputs for various sequences of the same length. Fig. 2d demonstrates the effect on STRNaming when the reference genome contains the less frequent allele of a flanking region SNP. Finally, Fig. 2e exemplifies the effect of submitting different sequence ranges into STRNaming.

Several challenges exist in implementing this algorithmic approach for repeat region bracketing. First, the STRNaming approach results in a change from the previous bracketed designation for some commonly used loci and, as noted previously, substantial population sequence data have been published in recent years. The challenge of translating recently published data can be ameliorated by updating the STRSeq BioProject to link the *2016 ISFG considerations* formatting [1] with these *2023 ISFG recommended* bracketed designations (updates to STRSeq are further detailed in Recommendation 3). Additionally, implementing this algorithm resolves the issue exemplified in Fig. 2b; however, this solution creates apparent discrepancies between the length-based allele number traditionally used in CE and the bracketed repeat, as shown in Fig. 2c. This challenge necessitates a change in conceptualization: the bracketed sequence is simply a visual aid rather than an explanation of the numerical allele. The allele number is inferred from the input sequence length, similar to historical allele designation for some complex loci (*e.g.*, D21S11 and SE33) and is prefixed to the bracketed sequence (as shown in Fig. 2c, *e.g.*, "CE13″).

In STRNaming, flanking region variant calls and placement are relative to the GRCh38 reference sequence. Placement is indicated by negative or positive numbers corresponding to their 5′ or 3′ positions, respectively, on either side of the bracketed repeat region of GRCh38. For example (as shown in Fig. 2c), − 25 C > T indicates that a "T" nucleotide was observed in the sample where a "C" nucleotide is located in GRCh38: 25 nucleotides 5′ of the STRNaming bracketed repeat region of GRCh38. Although this particular variant is cataloged in dbSNP as rs73250432, the STRNaming nomenclature does not use rs numbers to avoid the issue of novel variants and the dependency on a database; additionally, multiple rs numbers would be cumbersome to use in nomenclature. Necessarily, the use of the GRCh38 reference genome for flanking region variant calling means that when the reference genotype is the less-frequent allele at a variant position, the alternate allele will be reported with a proportionately high frequency, as shown in Fig. 2d.

When STRNaming is used to format a sequence range that is different from the ISFG minimum range, the bracketed repeat structure may vary, as shown in Fig. 2e. Interestingly, this CSF1PO example differs from the FGA and D18S51 examples in Fig. 2a in that the additional, adjacent repeat units, which are excluded from minimum range STRNaming bracketing, are not inferred from the reference for CSF1PO. This is because STRNaming will fill in *at most* one full repeat unit beyond the sequence range provided. It is important to note that the STRNaming result can be reversed back to the original sequence string by the same version and settings (including reference sequence) of the program, when the reference genome coordinate range is indicated. Thus, when sequence data with different reported ranges are to be compared, the sequence strings can be trimmed to a common range prior to comparison. While laboratories are encouraged to maximize the informativeness of sequences by reporting as large a sequence as possible with a given kit (*e.g.*, the kit-specific ranges shown in the FSSG), it is recommended that the ISFG minimum range and associated STRNaming output are also provided in allele frequency publications and any other applications where inter-laboratory comparisons are expected. The developers of STRNaming are considering improvements to facilitate access to the ISFG minimum range, including automated trimming of larger input sequences in order to provide multiple STRNaming outputs (*e.g.*, full range and ISFG minimum range).

Lastly, extended ranges of complex loci with repetitive regions adjacent to the counted repeat (such as SE33 and DYS385) can produce lengthy STRNaming results. While STRNaming is still the recommended approach, the bracketed format may not be as useful for these loci, particularly when formatting ranges beyond the ISFG minimum range. Arguably, the historical bracketing was not as useful for these more complex loci either, and manually bracketing such loci is more prone to inconsistencies and errors. Furthermore, the complexity and variety of motifs found at SE33 can result in different repeat units being bracketed by STRNaming, depending on the sequence and the input sequence range. In addition to the publications, useful information regarding the technical properties of STRNaming (particularly relevant to the more complex loci) can be found at fdstools.nl/strnaming. It is important to note that, generally, STRNaming produces a visually intuitive format and, often, the STRNaming format is the same as the historical bracketing. The exceptions are exemplified here for user awareness, and any tradeoffs are considered worthwhile in order to create a sustainable, automated system for developing user-friendly representations of the sequence strings.

2.2.2. Sequence Codes. A consistently short minimal "code" sequence proxy may be useful when analyzing data in casework, databasing when character space is limited, and for a common simple reference in discussion. Any system of codes necessitates a predetermined sequence range because increasing or decreasing the sequence range could reveal or mask polymorphisms, respectively. Therefore, it is recommended that sequence codes should be generated based on the ISFG minimum range if these codes are employed for general inter-laboratory comparisons. In specific scenarios, greater regions of overlap could be compared to maximize information. For example, two laboratories using different kits and both reporting the maximum possible flanking region could compare the entire overlapping range, and calculate statistics based on this hybrid range [43]. In order to implement a consistent system of codes, there are two options: 1) a repository of sequences and corresponding codes or 2) an algorithm which produces codes for input sequences. There are significant challenges to developing a repository of sequence codes, such as the need for committed resources to maintain it and sequence data handling considerations.

In 2020, an algorithmic method amenable to generating STR sequence codes was published using the hash function SHA-256 and additional calculations to convert a DNA sequence into a 50+ character sequence identifier (SID) using the 26 letters in the Modern English alphabet [28]. This SID can be truncated, depending on the application, as described below and shown in Fig. 3. Also exemplified in Fig. 3 is the importance of a predetermined sequence range in order for sequence codes to be used comparatively; any change in sequence, including sequence length, will result in a different SID. A web interface for generating SIDs is available at https://nichevision.github.io/sid.js/. In cooperation with the goals of this Commission, the developer has recently released an open-source version of the JavaScript for sequence

a.

TH01 Sequence String	ATGGTGAATGAATGAATGAATGAATGAATGAGGGA
Full SID	UCYLMZAZZPQEEOSMFIEIRFFSYHYOLNKVYAVUYKYWEUVDINCBXFDDMXB
Locus Name_SID	TH01_UCYLM
Locus Name_CE Allele_SID	TH01_7_UCYL

b.

TH01 Sequence String	ATGGTGAATGAATGAATGAATGAATGAATGAGG■
Full SID	JGUVGCAHRCSKPGAMNHEEASITTGYEOZCFQGZMLBRXMHIVBPTGDYFAYSB
Locus Name_SID	TH01_JGUVG
Locus Name_CE Allele_SID	TH01_7_JGUV

Fig. 3. a. An example TH01 39-nucleotide sequence string (ISFG minimum range) and its corresponding 55-character full SID, followed by a five-character SID added to the locus name, and a four-character SID added to the locus name and CE (length-based) allele. b. A 38-nucleotide sequence string, identical to the sequence in 3a. except for removal of one A nucleotide at the 3′ end (indicated by the black square). As shown, this shortened sequence generates a different SID.

transformation into SID, which can be incorporated into other open-source bioinformatic programs (full license and code available at https://github.com/nichevision/sid). Alternative licensing agreements are available by working directly with the developer. Due to the potential for these licensing restrictions to limit scientific development, the SID is described but not "recommended" by this Commission. To further improve accessibility, SID codes for the ISFG minimum range will be included in all STRSeq BioProject GenBank records and can be incorporated into a bioinformatic program as a library-style lookup rather than directly incorporating the SID-generating script into a bioinformatic program. In this case, when users encounter ISFG minimum range sequences novel to STRSeq, they could be directed to the web-based SID generator and/or strseq@nist.gov for assistance and potential GenBank record creation.

The number of characters needed to uniquely identify a sequence depends on the number of characters in the sequence coding system, the level of metadata accompanying the sequence, and the use case. For example, within a case with the locus name and CE (length-based) allele number prefixed, two sequence code characters may be sufficient to uniquely identify sequence-based alleles, as shown in [28]. If it were desirable to assign unique sequence codes to all sequence artifacts in a sample, many more characters would be needed. Similarly, for a use case such as databasing, more characters are needed to uniquely identify all STR sequences within a locus. This Commission considered the number of SID characters needed to uniquely identify a sequence within a locus when at least the locus name is included as metadata. Using the ISFG minimum range to evaluate the catalog of STR sequences currently contained in the STRSeq BioProject [13] and considering the potential for additional sequence discovery, five characters are recommended to uniquely identify all sequence-based alleles within a locus, while four characters are recommended if both locus name and CE allele are designated (as shown in Fig. 3). If this approach proves to be insufficient, it is anticipated this Commission will reconvene, assess and make further recommendations. Additional discussion regarding the use of sequence codes is included in subsequent sections (see below).

2.3. Resources

Easily accessible and properly formatted STR sequence exemplars and allele frequency data are expected to encourage the incorporation of ISFG STR nomenclature and sequence-based frequency calculations into bioinformatic software; therefore, the following STR sequence nomenclature resources will be updated and harmonized to facilitate the implementation of Recommendations 1 and 2: The FSSG, STRSeq, and STRidER.

Recommendation 3:. *Standardized resources should be used in bioinformatic software to promote consistency in nomenclature.*

2.3.1. FSSG

An updated version of the FSSG (version 6) has been created in the course of this Commission, annotating GRCh38 with the following information for the autosomal, Y- and X- chromosomal STR loci currently included in commercial STR sequencing kits: A) historical bracketing (2016–2022), B) STRNaming bracketing (2023 onward), C) ISFG minimum range, and D) commercial kit range(s). In an effort to simplify the schematic view of the STR loci included in the FSSG and provide more information, two new tabs were created: Common Locus Information and Variant Frequencies. The Common Locus Information tab lists the reference genome coordinates and sequence strings shown on the schematic view, as well as common STR motifs, and locus-specific notes. Additionally, the Variant Frequencies tab contains updated polymorphism information based on gnomAD v3.1.2, with different criteria for inclusion in this table based on the STR region. For example, within the repeat region, only the variants which contribute to common STR motifs are cataloged but in the extended flanking regions, variants are cataloged if the minor allele frequency (MAF) reaches at least 0.1% in one or more of these gnomAD v3.1.2 super-populations: African/African American, East Asian, European (non-Finnish), Latino/Admixed American, South Asian. Full details are included in the FSSG, available at https://strider.online/nomenclature.

2.3.2. STRSeq

The STRSeq BioProject is recommended as the primary resource to

provide the sequence nomenclature formats recommended herein. Fig. 4 shows an existing STRSeq GenBank record, with record-specific changes aligned with these recommendations indicated in blue: 1) sequence ranges updated to match both the manufacturer recommended ranges and ISFG minimum range, 2) bracketed repeat structure updated to align with these ISFG recommendations, and 3) SID code added corresponding to the ISFG minimum range. Efforts are underway to update the > 2500 existing STRSeq records. In addition, due to the flexibility for making changes and the forensic community's familiarity with this resource, STRBase (https://strbase.nist.gov) will be leveraged to help connect users with STR sequence information by incorporating STRSeq accession numbers and links to GenBank records. This resource will provide users with a simpler, more familiar mode by which to access this growing body of information.

As previously noted, many STR population sequence studies have been published in recent years; however, only a subset have been evaluated by the National Institute of Standards and Technology Applied Genetics Group (NIST AGG) for unique STR sequences to include in the STRSeq BioProject. To address recent and future publications (*i.e.*, 2016 onward), NIST AGG scientists are evaluating and will continue to evaluate published STR sequences and will create GenBank records for alleles novel to the STRSeq BioProject, when data quality criteria are met (see [14,17]). These additional records will further strengthen this resource by encompassing a larger share of the sequence variation found in human forensic STR loci.

2.3.3. STRidER

Currently, STR sequence allele nomenclature is not assessed during STRidER QC, and a specific sequence range beyond the conventional repeat region (as specified in the FSSG up to version 5) is not required. With the publication of these ISFG recommendations, the development of this additional QC step is underway and will involve a query of the STRSeq catalog.

The developers aim to introduce an integrated process whereby users upload population sample STR sequence data (including at least the ISFG minimum range) to STRidER for QC, STRidER queries STRSeq for a matching sequence, and the STRSeq record formatting and nomenclature are compared to the submission. In cases of no match in STRSeq, scientists at the Institute of Legal Medicine Medical University of Innsbruck and the NIST AGG will evaluate the unknown allele, including expanded checks for sequence coverage and range, polymorphisms in flanking regions, and phylogenetic context. For sequences which pass this evaluation, a new GenBank record will be created, containing both the STRNaming bracketed format and the SID for the ISFG minimum range. Such a process will strengthen the STRidER QC function and expand STRSeq, while harmonizing these resources. This process is particularly important for novel sequence variants likely to be encountered as population studies are extended in geographic scope and/or sample numbers.

Additionally, updates to the STRidER platform are underway to include the capability of providing sequence-based STR allele frequency data which have passed QC, as it currently does for length-based CE STR data. With more added functionality, users will ultimately be able to generate sequence-based STR profile frequency estimates for the ISFG minimum range in STRidER. The recommended changes have been initiated, and new capabilities will be announced via the STRidER website (https://strider.online) and newsletter.

2.4. New Loci

Unified characterization of new STR loci is vital for inter-laboratory comparability and databasing.

Recommendation 4:. *Researchers exploring new STR loci for forensic assays should evaluate whether a historical locus designation exists, follow Recommendations 1 and 2 to determine locus formatting, and establish genotype concordance with quality control data.*

2.4.1. New Locus Designations

Currently, when a researcher evaluates an STR locus not previously characterized for forensic use, there is no robust way for the researcher to obtain the historical locus designation [44,45]. During early gene mapping efforts, D#S# designations (*e.g.,* D21S11) were applied to "arbitrary DNA fragments and loci", where "D" stood for DNA, a number or letter corresponded to the chromosomal assignment, a letter indicated the complexity of the locus ("S" for a unique DNA fragment, "Z" for repetitive DNA segments at a single chromosomal site, "F" for families of homologous sequences found on multiple chromosomes), and a sequential number gave uniqueness to the prior designations [46]. These D#S# designations were cataloged in reports of the Human Gene Mapping (HGM) workshops held from 1973 to 1991. The final (1991) HGM workshop report contains nearly 5000 D#S# names [47]. Afterwards, efforts were aligned with the physical mapping efforts of The Human Genome Project (initiated in 1990), under the umbrella of the Human Genome Organisation (HUGO). HUGO remains an active organization, and a "gene nomenclature committee" still exists to develop and catalog gene names across vertebrates; however, their database does not appear to catalog D#S# designations (https://www.genenames.org). This Commission has not identified any active efforts to maintain historical locus designations alongside the modern human genome reference sequence (*e.g.*, GRCh38), nor is there a current authority or agency for establishing new STR locus designations for regions that were not historically designated.

Researchers and manufacturers interested in developing forensic assays for new STR loci (*i.e.*, those not currently included in commercial kits) can contact members of the STRAND working group and/or strseq@nist.gov for assistance in determining historical STR locus name designations, leveraging various resources: HGM Workshop 11 report [47], NCBI Probe database archive (https://ftp.ncbi.nih.gov/pub/ProbeDB), publications related to the Marshfield set [48], information in the catalog of STR variation described in [49], and the associated WebSTR database (http://webstr.ucsd.edu), information in the International Society of Genetic Genealogy (ISOGG) Wiki and associated resources (https://isogg.org/wiki/Wiki_Welcome_Page) and other relevant publications (*e.g.*, [50,51]).

2.4.2. New Locus Formatting

Formatting the repeat region of new loci should follow the recommendations described herein; researchers can contact the STRAND working group and/or strseq@nist.gov for assistance in setting the ISFG minimum range, evaluating STRNaming results, and determining associated sequence codes and thereby initiating potential inclusion of the loci in the resources described above. Additionally, when *ad hoc* locus names have been used in these interim years (*e.g.* [44]) and the historical locus name is subsequently identified, the *ad hoc* locus name will be maintained in any STRSeq BioProject GenBank records created through the efforts described herein.

2.4.3. New Locus Quality Control

An additional consideration in developing new STR loci is QC, specifically the location of the target locus (*i.e.*, *does the assay generate sequence data for the intended genomic coordinates?*). This assessment is difficult for assays of new loci, which are typically characterized in research studies, because neither high-quality positive control data nor standard reference material data (*e.g.*, NIST 2391d PCR-based DNA profiling standard) may be available. One solution described in a proof-of-concept study [52] leverages the Genome in a Bottle (GIAB, http://genomeinabottle.org) project, a public-private-academic consortium which provides an authoritative characterization of human genomes from broadly consented Coriell samples for use in clinical analytical validation and technology development. Researchers

DEFINITION
This is automatically generated by merging information from the **HumanSTR** fields: **STR locus name, Length-based allele, Bracketed record seq.**, and **Sequencing assay code**. Flanking region polymorphisms were previously identified by including the **FEATURES** variation **db_xref** field when present. This information is included in the STRNaming designation of flanking region polymorphisms found in the **Bracketed record seq.** field.

COMMENT
Text will be updated and made uniform across all records. Individual comments will be moved to the **HumanSTR Notes** field.

HumanSTR Fields
These fields have been updated and reordered. New fields are highlighted with **blue text**. Fields that have not changed are in black text. The **Historical bracketing** field, highlighted in **orange text**, is the **Bracketed repeat** field renamed. The content of this field has not changed. Removed fields are highlighted with **red text**.

```
LOCUS       MH085118               222 bp    DNA     linear   PRI 10-DEC-2019
DEFINITION  Homo sapiens microsatellite TH01 7 TGAA[7]_-12C>G FS,PS
            sequence.
ACCESSION   MH085118
VERSION     MH085118.2
DBLINK      BioProject: PRJNA380567
KEYWORDS    STRSeq; STR; TH01.
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 222)
  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 222)
  AUTHORS   NIST,A.G.G.
  TITLE     Direct Submission
  JOURNAL   Submitted (19-MAR-2018) Applied Genetics Group, National Institute
            of Standards and Technology, 100 Bureau Drive, MS-8314,
            Gaithersburg, MD 20899, USA
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##
            Sequence attribution  :: Applied Genetics Group, NIST
            STR locus name        :: TH01
            Length-based allele   :: 7
            Minimum range bracket :: TGAA[7]
            Bracketed record seq. :: TGAA[7]_-12C>G
            Sequencing technology :: MiSeq FGx
            Sequencing assay code :: FS,PS
            Coverage              :: >30X
            Length-based tech.    :: PowerPlex Fusion, 3130xl
            Assembly              :: GRCh38 (GCF_000001405)
            Chromosome            :: 11
            Ref. seq. accession   :: NC_000011.10
            Chrom. location       :: 2171056..2171277
            ISFG minimum range    :: 2171082..2171120
            ISFG min. range code  :: UCYLM
            Frequency reference   :: STRidER.online
            STR locus alt name    :: HUMTH01, TC11
            Historical bracketing :: [AATG]7
            Notes                 ::
            Repeat Location       :: 2171088..2171115
            Cytogenetic Location  :: 11p15.5
            ##HumanSTR-END##
FEATURES             Location/Qualifiers
     source          1..222
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
     misc_feature    1..72
                     /note="Verogen ForenSeq DNA Signature Prep Kit"
     variation       19
                     /note="-12C>G"
                     /db_xref="dbSNP:rs1051822965"
     misc_feature    23..222
                     /note="Promega PowerSeq 46GY System"
     repeat_region   27..65
                     /note="minimum range"
                     /rpt_type=tandem
                     /satellite="microsatellite:TH01"
ORIGIN
        1 tgcaggtcac agggaacaga gactccatgg tgaatgaatg aatgaatgaa tgaatgaatg
       61 agggaaataa gggaggaaca ggccaatggg aatcacccca gagcccagat acccttgaa
      121 ttttgccccc tatttgccca ggacccccca ccatgagctg ctgctagagc ctgggaaggg
      181 ccttggggct gcctccccaa gcaggcaggc tggttggggt gc
//
```

Fig. 4. Existing STRSeq GenBank record example with updates indicated in light blue text to the left of the record, and color coded in light blue, orange and red within the example record. The light gray portion of the record is unchanged. Color in online version only.

interested in this option can contact strseq@nist.gov for assistance in obtaining high-quality GIAB STR sequence data for these samples. Researchers can then obtain the associated samples to run as positive controls, comparing their own results to the GIAB expected sequences. Additionally, when new STR loci are included in published assays, the sequences extracted from GIAB genome data can be cataloged in the STRSeq BioProject, making the STR sequence information available to the entire forensic community.

2.4.4. General Considerations of Forensic Locus Selection

Microsatellites associated with disease have been identified in coding regions, adjacent untranslated regions, and introns, where expansions beyond a normal range are increasingly deleterious [53]. This kind of direct relationship is unlikely to be found for the STRs commonly used in human identification, which are typically located outside of gene regions (early characterized loci such as TH01 and FGA are intronic but have not been linked to any positive predictive values of phenotypic consequence [54]), although alternative relationships have been proposed [55]. The future sequencing and exploration of non-core STRs should be carefully examined for the presence of variants with positive predictive power that could impact privacy or personal health.

2.5. Databases

Several types of *databases* are used by the forensic DNA community. As described in Recommendation 3, STRidER contains STR allele frequency information, and the STRSeq BioProject catalogs and characterizes unique sequence strings, with both resources drawing information from academic population studies. These resources are allele frequency and variation databases and, as such, are different from investigative databases searched in the course of forensic DNA investigations, both in their purpose and the composition of individual profiles contained therein. It is far simpler to provide recommendations regarding community resources (particularly those managed by members of this Commission), whereas it exceeds the scope of this Commission to recommend changes to investigative databases. The information provided within this section is intended as "food for thought" for jurisdictions considering how to make existing STR investigative databases amenable to STR sequences and for future database design. Furthermore, jurisdictions will need to consider specific legislation or policies regarding the information which may be stored and searched in their respective databases, as well as generally applicable genetic privacy laws.

The challenge of updating investigative DNA databases as a result of technological improvements is something the forensic DNA community has faced before. In the early 1990s, jurisdictions around the world began building investigative databases composed of restriction fragment length polymorphism (RFLP)-based variable number of tandem repeat (VNTR) profiles. However, within a few years, the increased sensitivity realized via PCR amplification and the ability to discretely identify alleles, combined with the possibility of multiplexing STRs, presented a superior technology. For these reasons, in 1996, the U.S. National Research Council Committee on DNA Forensic Science recommended that forensic laboratories should implement PCR-based STR typing [56]. While all technology transitions present hurdles, this change was made easier by two factors. First, as RFLP analysis required significantly more sample than PCR testing, with PCR-based STR analyses there tended to be enough remaining sample to retest evidence that had been used to generate RFLP profiles, when needed. Second, due to the larger sample input requirement, lower throughput of testing, and the short time period during which RFLP had been used in forensic investigations, the RFLP investigative databases were in the order of tens of thousands of samples [56]. While the change was not trivial, it was tractable. Today, there may be as many as 100 million length-based STR profiles in investigative databases worldwide. Retesting all of these samples is not reasonable, nor is it needed for this technology transition. By implementing backward-compatible STR sequence nomenclature guidance, new STR sequence-based profiles are compatible with existing STR length-based investigative databases (barring length-based differences due to variable primer placement, as has been observed historically between commercial CE kits [57]).

Recommendation 5:. *It is the general recommendation of this Commission that a universal sequence coding and nomenclature is implemented across: community resources (described in Recommendation 3), bioinformatic software used by the forensic community for STR sequence data analysis, and software used to manage investigative databases if such software is updated to accept STR sequence data.*

2.5.1. Existing Databases and Compatibility

Current investigative databases configured for length-based STR profiles could be adapted to capture information generated from STR sequencing kits in multiple ways. The most straightforward system is the development of a length-based (numerical allele) STR profile from STR sequence data for the loci which overlap with the existing database. This route may simply require review and approval of the kit by the organization managing the database, including an evaluation of compatibility between the two data types (*e.g.*, the same sample generally produces the same length-based STR profile via either technology, other than differences related to primer placement, as noted previously).

While including only numerical allele profiles from the sequenced STR will result in a loss of any additional STR sequence information for database matching purposes, there are several potential ways to ameliorate this loss, depending on the database design. If STR sequences are databased as numerical alleles only, the database records should include commercial kit information at all levels of the database so that users are aware additional information is available when hits are returned to numerical allele profiles derived from sequence data. Further, laboratories should have policies and procedures for obtaining this additional information when needed.

Additionally, if the database software configuration allows for comments to be added to database records, information about the STR sequence could be included. Examples of such information include the presence of isoalleles at specific length-based homozygous loci or that the uploaded sole source evidence profile was deduced from a mixture profile based on STR sequence data. Database design may limit the use of comment fields if comments are applied to the entire profile, if the character length of comments is limited, or if comments are only stored at particular levels of the database (*e.g.*, the local level but not the national level).

When a database hit involves a length-based profile and a length-converted sequence-based profile, additional sequencing could be performed. For example, in the case of a sequenced evidence profile matching a length-based offender profile, the new suspect standard collected for direct comparison could be sequenced and compared to the evidence profile sequence data. This is analogous to the hit verification process used in some jurisdictions during past expansions of core database STR loci: when a hit resulted from a match to an older offender sample profile with fewer loci than the searched evidence profile, the new suspect standard was considered for typing at the new core loci for comparison to the evidence sample. The same "upgrade" is possible for any stored CE profile where DNA is still available.

2.5.2. Future Databases

There are many considerations for future investigative database development as forensic DNA sequencing evolves. For STR sequences, future database design could include the storage of sequence strings and/or established codes that unambiguously represent the sequence strings. The primary drawback to storing sequence codes is the prerequisite to establish and maintain a short designator system that is accessible to software vendors and forensic DNA databases worldwide. A suitable system is described under Recommendation 2, and

Recommendation 3 describes the development of a catalog of STR sequence strings with these corresponding sequence codes in the ISFG minimum range of Recommendation 1.

Storing sequence codes or STRNaming designators based on the ISFG minimum range rather than the actual sequences may offer the benefit of reduction in required character space. However, a consistent, defined sequence range is vital to inter-database comparisons of sequence proxies. Therefore, if these proxies are used as a primary means of comparison in databasing, reference genome coordinates (*e.g.*, GRCh38) must be established across the database or stored alongside each sequence proxy. Storing sequence codes rather than full sequences may appear to improve data privacy; however, actual data privacy is only improved if access to the coding system is restricted. If character space permits, storing the ISFG minimum range of DNA sequence for search and data exchange purposes is preferred to eliminate the ambiguity caused by proxies; additionally, the use of the fixed ISFG minimum range removes the need for sequence alignment which is expected to facilitate search method development and improve database searching speed.

It is important that STR sequence-based databases are compatible with existing length-based databases for the foreseeable future. Database developers should consider the cost-benefit of incorporating STR sequence data fields into the current length-based database design or developing a parallel sequence-based STR database that can interface with the current length-based database.

With the additional information available from DNA sequencing, there exists an opportunity to develop capacity for additional STR markers as well as other forensic markers and data types, such as SNPs and microhaplotypes (multiple SNPs in a phased sequence string). Ideally, new data models would be created for SNP and microhaplotype genotypes. To date, no core set of SNPs or microhaplotypes has been determined for databasing. A core set of SNPs (such as those proposed in [58,59]) will comprise many more markers than core sets of STRs due to the lower discriminating capacity of SNPs. Determining a core set of markers will require a concerted effort involving the development of marker criteria, review of literature and published data, and inter-laboratory studies. With STRs being the common thread in all national forensic DNA databases, expanding or changing current DNA databasing models and establishing a new core set of forensic markers requires an international effort to ensure alignment across national databases.

Database input requirements are likely to be different for other forensic markers compared to STRs and will need to be determined by the forensic DNA community. In addition to direct sequencing of microhaplotypes producing phased SNP alleles, such loci may be assayed with SNP arrays for example, generating individual SNP genotypes without phase data. If microhaplotypes are assayed in phase on some samples and as single SNPs on others, the phase becomes additional information that could be verified when a potential match arises, analogous to sequences for length-based STRs. A consistent framework of nomenclature for SNPs and microhaplotypes (including a clear indication of phase, when applicable) will be important.

3Conclusions

This report of the ISFG DNA Commission on STR Sequence Nomenclature incorporates current knowledge into five recommendations covering 1) sequence strings and a minimum reporting range, 2) additional sequence formats, 3) resources for researchers, bioinformaticians, and practitioners, 4) information on characterizing new STR loci, and 5) considerations for incorporating STR sequences and other new markers into investigative databases. STR sequence interpretation is beyond the scope of this Commission but may be considered by future Commissions. Relevant topics include evaluation of stutter and other artifacts, guidance on population database sample size, and calculation of frequency estimates for rare STR sequence alleles.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Bruce Budowle: Currently Consultant with Verogen.

Acknowledgements

The authors are grateful to Peter de Knijff and Sascha Willuweit for their early contributions to this DNA Commission. Additionally, Peter M. Schneider served on this DNA Commission until his passing in 2022, and we honor his contribution with posthumous authorship.

Disclaimers

NIST: This DNA Commission of the ISFG has sole responsibility for the contents of this report and the questions, findings, and recommendations within. The views expressed in this report do not necessarily represent the views of the U.S. Department of Commerce or the National Institute of Standards and Technology. Certain commercial entities are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that any of the entities identified are necessarily the best available for the purpose.

FBI: This DNA Commission of the ISFG has sole responsibility for the contents of this report and the questions, findings, and recommendations within. Names of commercial manufacturers are provided for identification purposes only, and inclusion does not imply endorsement of the manufacturer, or its products or services by the FBI. The views expressed are those of the author(s) and do not necessarily reflect the official policy or position of the FBI or the U.S. Government.

References

[1] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, et al., 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.

[2] X. Zeng, J. King, S. Hermanson, J. Patel, D.R. Storts, B. Budowle, An evaluation of the PowerSeq Auto System: a multiplex short tandem repeat marker kit compatible with massively parallel sequencing, Forensic Sci. Int. Genet. 19 (2015) 172–179.

[3] A.C. Jager, M.L. Alvarez, C.P. Davis, E. Guzman, Y. Han, L. Way, et al., 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.

[4] Z. Wang, D. Zhou, H. Wang, Z. Jia, J. Liu, X. Qian, et al., Massively parallel sequencing of 32 forensic markers using the Precision ID GlobalFiler NGS STR Panel and the Ion PGM System, Forensic Sci. Int. Genet. 31 (2017) 126–134.

[5] S. Kocher, P. Muller, B. Berger, M. Bodner, W. Parson, L. Roewer, et al., Inter-laboratory validation study of the ForenSeq DNA Signature Prep Kit, Forensic Sci. Int. Genet. 36 (2018) 77–85.

[6] R. Tao, W. Qi, C. Chen, J. Zhang, Z. Yang, W. Song, et al., Pilot study for forensic evaluations of the Precision ID GlobalFiler NGS STR Panel v2 with the Ion S5 system, Forensic Sci. Int. Genet. 43 (2019), 102147.

[7] K.M. Stephens, R. Barta, K. Fleming, J.C. Perez, S.F. Wu, J. Snedecor, et al., Developmental validation of the ForenSeq MainstAY kit, MiSeq FGx sequencing system and ForenSeq universal analysis software, Forensic Sci. Int. Genet. 64 (2023), 102851.

[8] K. Elwick, P. Rydzak, J.M. Robertson, Evaluation of library preparation workflows and applications to different sample types using the powerSeq((R)) 46GY System with Massively Parallel Sequencing, Genes 14 (5) (2023) 977.

[9] A. Alonso, P.A. Barrio, P. Muller, S. Kocher, B. Berger, P. Martin, et al., Current state-of-art of STR sequencing in forensic genetics, Electrophoresis 39 (21) (2018) 2655–2668.

[10] B. Bruijns, R. Tiggelaar, H. Gardeniers, Massively parallel sequencing techniques for forensics: a review, Electrophoresis 39 (21) (2018) 2642–2654.

[11] T.I. Huszar, K.B. Gettings, P.M. Vallone, An introductory overview of open-source and commercial software options for the analysis of forensic sequencing data, Genes 12 (11) (2021) 1739.

[12] C. Phillips, K.B. Gettings, J.L. King, D. Ballard, M. Bodner, L. Borsuk, et al., The devil's in the detail": release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide, Forensic Sci. Int. Genet. 34 (2018) 162–169.

[13] K.B. Gettings, L.A. Borsuk, D. Ballard, M. Bodner, B. Budowle, L. Devesse, et al., STRSeq: a catalog of sequence diversity at human identification short tandem repeat loci, Forensic Sci. Int. Genet. 31 (2017) 111–117.
[14] M. Bodner, I. Bastisch, J.M. Butler, R. Fimmers, P. Gill, L. Gusmao, 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.
[15] M. Bodner, W. Parson, The STRidER report on two years of quality control of autosomal STR population datasets, Genes 11 (8) (2020) 901.
[16] K.B. Gettings, D. Ballard, M. Bodner, L.A. Borsuk, J.L. King, W. Parson, et al., Report from the STRAND Working Group on the 2019 STR sequence nomenclature meeting, Forensic Sci. Int. Genet. 43 (2019), 102165.
[17] L.A. Borsuk, P.M. Vallone, K.B. Gettings, STRSeq: FAQ for submitting, Forensic Sci. Int. Genet. Suppl. Ser. 8 (1) (2022) 245–247.
[18] P. Gill, L. Foreman, J.S. Buckleton, C.M. Triggs, H. Allen, A comparison of adjustment methods to test the robustness of an STR DNA database comprised of 24 European populations, Forensic Sci. Int. 131 (2–3) (2003) 184–196.
[19] L.A. Welch, P. Gill, C. Phillips, R. Ansell, N. Morling, W. Parson, et al., European Network of Forensic Science Institutes (ENFSI): Evaluation of new commercial STR multiplexes that include the European Standard Set (ESS) of markers, Forensic Sci. Int. Genet. 6 (6) (2012) 819–826.
[20] W. Parson, A. Dur, EMPOP–a forensic mtDNA database. Forensic Sci. Int. Genet. 1 (2) (2007) 88–92.
[21] S. Willuweit, L. Roewer, Y chromosome haplotype reference database (YHRD): Update, Forensic Sci. Int. Genet. 1 (2007) 83–87.
[22] W. Parson, A. Brandstatter, A. Alonso, N. Brandt, B. Brinkmann, A. Carracedo, et al., The EDNAP mitochondrial DNA population database (EMPOP) collaborative exercises: organisation, results and perspectives, Forensic Sci. Int. 139 (2–3) (2004) 215–226.
[23] L. Gusmao, J.M. Butler, A. Linacre, W. Parson, L. Roewer, P.M. Schneider, et al., Revised guidelines for the publication of genetic population data, Forensic Sci. Int. Genet. 30 (2017) 160–163.
[24] P. Gill, L. Foreman, J.S. Buckleton, C.M. Triggs, H. Allen, A comparison of adjustment methods to test the robustness of an STR DNA database comprised of 24 European populations, Forensic Sci. Int. 131 (2) (2003) 184–196.
[25] S.M. Gomes, M. Bodner, L. Souto, B. Zimmermann, G. Huber, C. Strobl, et al., Human settlement history between Sunda and Sahul: a focus on East Timor (Timor-Leste) and the Pleistocenic mtDNA diversity, BMC Genom. 16 (2015), 70.
[26] T.R. Moretti, B. Budowle, J.S. Buckleton, Erratum for population data on the thirteen CODIS core short tandem repeat loci in African Americans, US Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians, J. Forensic Sci. 60 (4) (2015) 1114–1116.
[27] J. Buckleton, J. Curran, J. Goudet, D. Taylor, A. Thiery, B.S. Weir, Population-specific F.S.T. values for forensic STR markers: a worldwide survey, Forensic Sci. Int. Genet. 23 (2016) 91–100.
[28] B. Young, T. Faris, L. Armogida, A nomenclature for sequence-based forensic DNA analysis, Forensic Sci. Int. Genet. 42 (2019) 14–20.
[29] J. Hoogenboom, T. Sijen, K.J. van der Gaag, STRNaming: generating simple, informative names for sequenced STR alleles in a standardised and automated manner, Forensic Sci. Int. Genet. 52 (2021), 102473.
[30] E.Y. Lee, H.Y. Lee, K.J. Shin, Off-ladder alleles due to a single nucleotide polymorphism in the flanking region at DYS481 detected by the PowerPlex((R)) Y23 System, Forensic Sci. Int. Genet. 24 (2016) e7–e8.
[31] D.Y. Wang, R.L. Green, R.E. Lagace, N.J. Oldroyd, L.K. Hennessy, J.J. Mulero, Identification and secondary structure analysis of a region affecting electrophoretic mobility of the STR locus SE33, Forensic Sci. Int. Genet. 6 (3) (2012) 310–316.
[32] E.W. Sayers, E.E. Bolton, J.R. Brister, K. Canese, J. Chan, D.C. Comeau, et al., Database resources of the national center for biotechnology information, Nucleic Acids Res. 50 (D1) (2022) D20–D26.
[33] S. Gudmundsson, M. Singer-Berk, N.A. Watts, W. Phu, J.K. Goodrich, M. Solomonson, et al., Variant interpretation using population databases: lessons from gnomAD, Hum. Mutat. 43 (8) (2022) 1012–1030.
[34] R.S. Just, J.A. Irwin, Use of the LUS in sequence allele designations to facilitate probabilistic genotyping of NGS-based STR typing results, Forensic Sci. Int. Genet. 34 (2018) 197–205.
[35] A. Urquhart, C.P. Kimpton, T.J. Downes, P. Gill, Variation in short tandem repeat sequences–a survey of twelve microsatellite loci for use as forensic identification markers, Int. J. Leg. Med. 107 (1994) 13–20.
[36] C. Puers, H.A. Hammond, L. Jin, C.T. Caskey, J. Schumm, Identification of repeat sequence heterogeneity at the polymorphic short tandem repeat locus HUMTH01 [AATG]n and reassignment of alleles in population analysis by using a locus-specific allelic ladder, Am. J. Hum. Genet. 53 (4) (1993) 953–958.
[37] W. Bar, B. Brinkmann, P. Lincoln, W.R. Mayr, U. Rossi, DNA recommendations–1994 report concerning further recommendations of the DNA Commission of the ISFH regarding PCR-based polymorphisms in STR (short tandem repeat) systems, Int. J. Leg. Med. 107 (3) (1994) 159–160.
[38] W. Bar, B. Brinkmann, B. Budowle, A. Carracedo, P. Gill, P. Lincoln, et al., DNA recommendations: Further report of the DNA Commission of the ISFH regarding the use of short tandem repeat systems, Int. J. Leg. Med. 110 (4) (1997) 175–176.
[39] P. Gill, C. Brenner, B. Brinkmann, B. Budowle, A. Carracedo, M.A. Jobling, et al., DNA commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs, Int. J. Leg. Med. 114 (6) (2001) 305–309.
[40] L. Gusmao, J.M. Butler, A. Carracedo, P. Gill, M. Kayser, W.R. Mayr, et al., DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis, Int. J. Leg. Med. (2005) 1–10.
[41] C. Santos, M. Fondevila, D. Ballard, R. Banemann, A.M. Bento, C. Borsting, et al., Forensic ancestry analysis with two capillary electrophoresis ancestry informative marker (AIM) panels: results of a collaborative EDNAP exercise, Forensic Sci. Int. Genet. 19 (2015) 56–67.
[42] J. Hoogenboom, N. Weiler, L. Busscher, L. Struik, T. Sijen, K.J. van der Gaag, Advancing FDSTools by integrating STRNaming 1.1, Forensic Sci. Int. Genet. 61 (2022), 102768.
[43] B. Young, M. Marciano, K. Crenshaw, G. Duncan, L. Armogida, B. McCord, Match statistics for sequence-based alleles in profiles from forensic PCR-mps kits, Electrophoresis 42 (6) (2021) 756–765.
[44] N.M.M. Novroski, F.R. Wendt, A.E. Woerner, M.M. Bus, M. Coble, B. Budowle, Expanding beyond the current core STR loci: an exploration of 73 STR markers with increased diversity for enhanced DNA mixture deconvolution, Forensic Sci. Int. Genet. 38 (2019) 121–129.
[45] C. Phillips, W. Parson, J. Amigo, J.L. King, M.D. Coble, C.R. Steffen, et al., D5S2500 is an ambiguously characterized STR: Identification and description of forensic microsatellites in the genomics age, Forensic Sci. Int. Genet. 23 (2016) 19–24.
[46] H.F. Willard, M.H. Skolnick, P.L. Pearson, J.L. Mandel, Report of the committee on human gene mapping by recombinant DNA techniques, Cytogenet Cell Genet. 40 (1–4) (1985) 360–489.
[47] Human gene mapping 11. London Conference, 1991. Eleventh International Workshop on Human Gene Mapping. London, UK, August 18–22, 1991. Cytogenet Cell Genet. 1991;58(1–2):1–984.
[48] T.J. Pemberton, C.I. Sandefur, M. Jakobsson, N.A. Rosenberg, Sequence determinants of human microsatellite variability, BMC Genom. 10 (2009) 612.
[49] T. Willems, M. Gymrek, G. Highnam, C. Genomes Project, D. Mittelman, Y. Erlich, The landscape of human STR variation, Genome Res. 24 (11) (2014) 1894–1904.
[50] C. Phillips, A genomic audit of newly-adopted autosomal STRs for forensic identification, Forensic Sci. Int. Genet. 29 (2017) 193–204.
[51] E.K. Hanson, J. Ballantyne, Comprehensive annotated STR physical map of the human Y chromosome: Forensic implications, Leg. Med. 8 (2) (2006) 110–120.
[52] K.B. Gettings, L.A. Borsuk, J. Zook, P.M. Vallone, Unleashing novel STRS via characterization of genome in a bottle reference samples, Forensic Sci. Int. Genet. Suppl. Ser. 7 (1) (2019) 218–220.
[53] J.R. Brouwer, R. Willemsen, B.A. Oostra, Microsatellite repeat instability and neurological disease, Bioessay.: N. Rev. Mol., Cell. Dev. Biol. 31 (1) (2009) 71–83.
[54] N. Wyner, M. Barash, D. McNevin, Forensic autosomal short tandem repeats and their potential association with phenotype, Front. Genet. 11 (2020), 884.
[55] B.F. Algee-Hewitt, M.D. Edge, J. Kim, J.Z. Li, N.A. Rosenberg, Individual identifiability predicts population identifiability in forensic microsatellite markers, Curr. Biol.: CB 26 (7) (2016) 935–942.
[56] National Research Council (U.S.). Committee on DNA Forensic Science: an Update., National Research Council (U.S.). Commission on DNA Forensic Science: an Update. The evaluation of forensic DNA evidence. Washington, D.C.: National Academy Press; 1996. xv, 254 p.p.
[57] C.R. Hill, M.C. Kline, J.J. Mulero, R.E. Lagace, C.W. Chang, L.K. Hennessy, et al., Concordance study between the AmpFlSTR MiniFiler PCR amplification kit and conventional STR typing kits, J. Forensic Sci. 52 (4) (2007) 870–873.
[58] A.J. Pakstis, W.C. Speed, R. Fang, F.C. Hyland, M.R. Furtado, J.R. Kidd, et al., SNPs for a universal individual identification panel, Hum. Genet. 127 (3) (2010) 315–324.
[59] J.J. Sanchez, C. Phillips, C. Borsting, K. Balogh, M. Bogus, M. Fondevila, et al., A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27 (9) (2006) 1713–1724.