Author Manuscript

Accepted for publication in a peer-reviewed journal

NIST National Institute of Standards and Technology • U.S. Department of Commerce

Published in final edited form as:

Forensic Sci Int Genet. 2019 November ; 43: 102165. doi:10.1016/j.fsigen.2019.102165.

Report from the STRAND Working Group on the 2019 STR Sequence Nomenclature Meeting

Katherine Butler Gettings[a], **David Ballard**[b], **Martin Bodner**[c], **Lisa A. Borsuk**[a], **Jonathan L. King**[d], **Walther Parson**[c,e], **Christopher Phillips**[f]

[a)]U.S. National Institute of Standards and Technology, Biomolecular Measurement Division, 100 Bureau Drive, Gaithersburg, MD, 20899, USA

[b)]King's Forensics, King's College London, Franklin-Wilkins Building, 150 Stamford Street, London, UK

[c)]Institute of Legal Medicine, Medical University of Innsbruck, Austria

[d)]Center for Human Identification, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX, 76107, USA

[e)]Forensic Science Program, The Pennsylvania State University, USA

[f)]Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain

Abstract

This report summarizes topics discussed at the STR sequence nomenclature meeting hosted by the STRAND Working Group in April 2019. Invited attendees for this meeting included researchers known-to-us to be developing STR sequence-based nomenclature schemata, scientific representatives from vendors developing STR sequence bioinformatic methods, DNA intelligence database curators, and academic experts in STR genomics. The goal of this meeting was to provide a forum for individuals developing nomenclature schemata to present and discuss their ideas, encouraging mutual awareness, identification of differences in approaches, opposing aspects, and opportunities for parallelization while some approaches are still under development.

Introduction

Since 2016, the *ad hoc* formed STR Sequence Working Group (the authorship of this publication) has been collaborating to harmonize related efforts across our respective laboratories, consisting of: STRidER STR sequence quality control [1], STRSeq catalog of sequences [2], STRait Razor bioinformatic freeware [3], the Forensic STR Sequence Structure Guide [4, 5], and large-scale population sample sequencing efforts [6–9] (see [10] for a comprehensive review).

Disclaimer: Identification of commercial assays and/or software is not intended to imply recommendation or endorsement by the U.S. National Institute of Standards and Technology.

To address the more broadly reaching issue of STR sequence nomenclature, we formalized our group in 2018 as the STRAND Working Group (Short Tandem Repeat: Align, Name, Define). Subsequently, we received the endorsement of the ISFG Executive Board to organize an STR sequence nomenclature meeting, which was held in London on April 11th and 12th, 2019. Invited attendees for this meeting included researchers known-to-us to be developing STR sequence-based nomenclature schemata, scientific representatives from vendors developing STR sequence bioinformatic methods, DNA intelligence database curators, and academic experts in STR genomics. Attendees and affiliations were as follows:

David Ballard, King's College London, UK.

Pedro A. Barrio, National Institute of Toxicology and Forensic Science, Spain.

Martin Bodner, Medical University of Innsbruck, Austria.

Claus Børsting, University of Copenhagen, Denmark.

Lisa Borsuk, National Institute of Standards and Technology, US.

Laurence Devesse, King's College London, UK.

Kristiaan van der Gaag, Netherlands Forensic Institute, Netherlands.

Sebastian Ganschow, LABCON-OWL, Germany.

Katherine Gettings, National Institute of Standards and Technology, US.

Peter Gill, Norwegian Institute of Public Health, Norway.

Theresa Gross, University of Cologne, Germany.

Douglas Hares, Federal Bureau of Investigation, US.

Cydne Holt, Verogen, US.

Jerry Hoogenboom, Netherlands Forensic Institute, Netherlands.

Tunde Huszar, University of Leicester, UK.

Jodi Irwin, Federal Bureau of Investigation, US.

Rebecca Just, Federal Bureau of Investigation, US.

Jonathan King, University of North Texas Health Science Center, US.

Peter de Knijff, Leiden University, Netherlands.

Robert Lagacé, Thermo Fisher, US.

Walther Parson, Medical University of Innsbruck, Austria.

Christopher Phillips, University of Santiago de Compostela, Spain.

Peter Schneider, University of Cologne, Germany.

Christian Sell, BKA Wiesbaden, Germany.

Sascha Willuweit, Charité Berlin University of Medicine, Germany.

Brian Young, NicheVision, US.

The goal of this meeting was to provide a forum for individuals developing nomenclature schemata to present and discuss their ideas. Thus, the first day of the meeting was dedicated to attendee presentations, and the second day consisted of group discussion (agenda and presentations permitted for distribution are included in Supplemental File 1). This forum encouraged mutual awareness, identification of differences in approaches, opposing aspects, and opportunities for parallelization while some approaches are still under development. The primary topics are outlined, and related discussions are summarized in this report, which we hope will advance this conversation toward the ultimate goal of an official (ISFG) recommendation on STR sequence nomenclature.

1. **Formats for STR sequences**: The first outcome of this meeting was consensus on the utility of three formats for STR sequences. The formats are described below, and the relevant presentations are summarized.

    1.1. **Short designator**: For analyzing data within a case, databasing, and for common simple reference in discussion, a minimal code may be useful. Methods for generating such a code were presented and applications were discussed as follows:

        1.1.1. Brian Young presented a process using the hash function SHA-256 that converts a DNA sequence into a 55 letter sequence identifier (SID) [11]. This SID can be truncated, depending on the application (e.g., identifying sequences within a sample/case may only require two letters). This method is available on GitHub (https://nichevision.github.io/sid.js/) and has been incorporated into ArmedXpert-MixtureAce software (NicheVision), where the SID is appended to the length-based allele and the locus name (e.g., TPOX 12 KG). Linking SIDs together with ticks or dots serves to identify artifacts and stutter, respectively, to primary alleles in the software.

        1.1.2. Sascha Willuweit presented NOMAUT, short for Nomenclature Authority, which is an online repository accessed at nomaut.org. The service allows users to upload a sequence, which is assigned a lower-case letter designator (e.g., TPOX 12+b) when the submitted sequence is new to the database or is converted to upper-case if already submitted from another source (TPOX 12+B). NOMAUT seeks to serve as a centralized repository for STR sequence alleles; it can also be used offline, with periodic updates.

        1.1.3. Rebecca Just presented on using the LUS (longest uninterrupted stretch) to represent sequence alleles and

stutter in existing probabilistic genotyping applications [12], and Peter Gill demonstrated the use of LUS-based allele designations in EuroForMix [13]. The designator consists of the locus name, length-based allele, and LUS (e.g., D12S391 23_13 represents an [AGAT]13 [AGAC]9 AGAT sequence/allele). Some loci regularly exhibit multiple alleles which would have the same designator, as in the aforementioned D12S391 23_13 which also describes [AGAT]13 [AGAC]10; however, by extending the designation to secondary or tertiary reference regions, nearly all known alleles can be differentiated. An example locus with rarely non-differentiable alleles under this system is D21S11, at which five subunits of the most common motif have shown variability (indicated by bolded n): [TCTA]**n** [TCTG]**n** [TCTA]**n** TA [TCTA]**n** TCA [TCTA]2 TCCATA [TCTA]**n**.

1.1.4. *Included for completeness/context,* Lisa Borsuk presented on the STRSeq BioProject [2](www.ncbi.nlm.nih.gov/bioproject/380127), which is a catalog of sequences maintained as GenBank records at NCBI, where each sequence has a unique accession number (e.g., MH167243.1). STRSeq records are created for sequences published in population studies after quality control. Many STRSeq records represent sequencing results for a single sample across multiple assays, with different ranges of flanking sequence overlap. When a flanking region polymorphism is present outside of the range of one assay, different accession numbers may be assigned to the same sequence in that assay. For example, MH167243.1 and MH167244.1 are both 205 nucleotide (nt) D16S539 sequences with repeat region [GATA]9. These records are differentiated by rs11642858, present 20 nt from the 3′ end of the reported string, included in the ForenSeq range and not in the PowerSeq range. Therefore, the 173 nt PowerSeq sequence is identical for these two accession numbers. If a designator system is recommended by the ISFG DNA Commission, the unique designators could be added and maintained within STRSeq records, connecting such parallel records for easier comparison.

1.2. **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 (e.g., interpretation of stutter). Historically, the original publication characterizing the repeat region for forensic use defined this format, in

which the repeat region of the sequence is represented by the repeated motif and the number of repeats. Efforts were made to standardize the start/stop and inclusion/exclusion of neighboring repetitive elements on a per-locus basis [14–19]; however, many exceptions exist due either to historical legacy (locus was characterized before guidance was published), or the inability of a rule set to encompass all scenarios [4, 5].

Historically, the bracketed sequence encompassed the start/stop points of the “counted” repeat region. This maximizes the ability to visually discern the length-based allele from the bracketed repeat; however, this approach is not well-suited to some situations (e.g., a 10 allele at D13S317 with the common rs9546005 A>T would be bracketed as [TATC]10 TATC... rather than [TATC]11). In addition, practically speaking, this approach precludes coding programs for automatic bracketing; instead requiring a look-up database. This introduces the possibility of variable approaches among laboratories when sequences are encountered which are not present in the database, particularly at more complex loci such as D21S11 or SE33.

Jerry Hoogenboom and Kristiaan van der Gaag presented a program called STRNaming (manuscript in preparation), which standardizes and automates conversion of the STR string into a bracketed format, based on a defined set of parameters. 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). At the time of the meeting, the developers were evaluating settings and preparing to engage users for feedback, with an eventual goal of establishing universal parameters that yield the most coherent arrangement of the repeat region structure and overall data display regarding any locus in present or future use.

Challenges to this approach include a likely change in bracketed designation for some commonly used loci, where significant sequence data have already been published in recent years. Additionally, implementing an algorithm such as this is likely to result in apparent discrepancies between the length-based CE allele number and the bracketed repeat. While STRNaming results in a more inclusive user-friendly representation of the sequence string, the length-based allele number would still be inferred from the full sequence length and is maintained as part of the allele name.

Fig. 1 demonstrates parameterized bracketing for various D13S317 alleles. The length-based CE allele number is explicitly represented in the name, as the bracketed sequence includes additional repeats outside the originally “counted” repeat region. Some length variation can be observed in this “extra” bracketed sequence. The allele name

format accommodates sequence variation outside the repeat region by means of variant calls, where variations 5′ or 3′ of the repeat region have negative or positive position numbers, respectively. For example, −25C>T indicates that a T nucleotide was encountered 25 bases 5′ of the repeat region, whereas the reference sequence has a C in that position. Although this particular variant is also known as rs73250432, the nomenclature does not use rs numbers to avoid potential issues with novel variants and the dependency on database lookups.

1.3. **Full string**: as stated in the 2016 considerations paper [4], the unformatted, entire reported sequence and associated genomic coordinates serve as an unequivocal record of results. The way in which this information is stored (e.g., in the case report, case file, or as a database with corresponding short designators applied per case), falls under the purview of each laboratory.

At this time, forensic DNA databasing software (e.g., CODIS) is generally not equipped to store or search STR sequence strings. Such databases primarily contain convicted offender samples; therefore, enabling STR sequence storage or search capabilities may be of limited use until laboratories begin routinely sequencing this sample type. In the interim, length based (numerical allele) profiles can be developed via STR sequencing assays. Profiles generated with one such assay have recently been approved for upload to the U.S. National DNA Index System (see *CODIS and NDIS Fact Sheet* at https://www.fbi.gov/services/laboratory/biometric-analysis/codis/codis-and-ndis-fact-sheet#NDIS, accessed May 30, 2019). Analysts confirming interlaboratory matches could compare sequence data, when applicable.

2. **Defined coordinates**: A second outcome of the meeting is the need for a recommended start and stop per locus, oriented to a reference genome. This is prerequisite to a short designator system. Four possible definitions were discussed; these are described below and applied to the D13S317 locus in Fig. 2.

2.1. **Assay Specific**: Coordinates designed to maximize flanking region sequence per assay/software. Maximizing reported flanking region is desirable for research purposes, to detect private mutations and assess potential association of flanking region polymorphisms with repeat number alleles or a motif. For casework purposes, at some loci, it may be challenging to obtain high quality/high read depth flanking region data for larger alleles. Removing reads because they do not contain high quality flanking region sequence would likely be an undesirable trade-off in low-level samples. A recent analysis of ForenSeq SNP data showed reporting the flanking region nominally decreased read depth (>95 % of reduced region) [20]; however, the effect of these bounds has yet to be reported for the longer amplicons of STRs.

Additionally, assay-centric coordinates would require changes in concert with assay design changes, and the need to establish new coordinate sets for future assays. A key piece of information needed for such coordinates is the "analyzable range" per assay, which has been released for the three existing commercial STR sequencing assays. To facilitate the nomenclature discussion, these ranges have been compiled into Supplemental File 2, a single spreadsheet formatted similarly to the STR Sequence Guide.

2.2. **Informative universal coordinates**: Coordinates designed to maximize informative polymorphisms in flanking regions across existing assays. Maximizing informative SNPs and indels would lead to increased differentiation of alleles. The above indicated trade-off in quality would still apply. Additionally, considering information gain without regard to current assay design may result in a recommended set of coordinates requiring significant re-design of current manufactured assays (and repeated validation experiments for early adopters).

2.3. **Unambiguous universal coordinates**: The minimum range of coordinates, which provide unambiguous termination of the designated repeat region. For multiple loci, additional tetranucleotides similar to the repeat motif are present adjacent to the "counted" region. In such cases, a single change may create the appearance of an additional repeat, and often, this change has been observed at measurable frequencies (e.g., D13S317: rs9546005 [adjacent to the repeat in Fig. 2] and vWA: rs199970098). Ambiguous regions such as these would be included/reported under this coordinate definition; the range would terminate when at least two substitutions (not previously observed in tandem) would be needed to create the appearance of an additional repeat.

2.4. **Repeat region only**: Coordinates defining the "counted" repeat region only. While this approach would work for many loci, there are examples where it would lead to ambiguous sequence reporting (as discussed in section 2.3) and could result in increased challenges for string searching.

Several considerations regarding defined coordinates were discussed in the meeting, as follows.

For the coordinate definitions in 2.2, 2.3 and 2.4, the concept of a "recommended" range pertains to unifying results across laboratories/assays; high quality data may be present outside of this range. If the eventual recommended range lies within the extent of high quality data, it is expected that some laboratories will continue to interpret flanking region polymorphisms beyond these bounds. It would be the laboratory's own decision to determine how this information is applied. One relevant analogy may be the use of

STR allele(s) below analytical threshold on an electropherogram to exclude contributors; however, it is important to distinguish that the analytical threshold is determined based on data quality whereas coordinate definitions 2.2, 2.3 and 2.4 are not directly related to data quality.

One issue pertinent to establishing ranges is that different countries have varied legislation regarding forensic applications of SNP data. As this discussion expands and progresses, it will be useful to understand existing legislation which may prohibit a laboratory from reporting SNPs in these non-coding STR flanking regions.

Any future recommended ranges will exclude the primer sequences, meaning bases reported within these ranges should reflect the genomic sequence of the sample donor rather than the primer sequence used in its amplification. For example, if the recommended range is “repeat region only”, the STR sequencing assay primers must bind entirely outside of the repeat region. It is expected some current assay redesign will be required in order to meet this criterion, due to existing examples where the primer binding site appears to extend into the repeat region. Inference of genomic sequence based upon the incorporation of primers is not considered a rigorous scientific approach.

Finally, it has come to the attention of the STRAND Working Group that some researchers have considered the flanking sequence included in the Forensic STR Sequence Structure Guide [5] to be the recommended range. This is not a recommended range, but rather a neutral, arbitrary setting of currently 100 base pairs on either side of the repeat region, designed to highlight significant flanking region sequence features that may only be relevant to some forensic primer designs.

3. **Forensic-specific reference**: A significant point of discussion in the meeting was the possibility of designating a forensic-specific reference genome (as opposed to, e.g., GRCh38 human genome reference sequence). Three advantages of creating such a reference genome are: a) Elimination of rare SNP alleles in STR flanking regions and incorporation of known insertions; b) Stability, i.e., the forensic community would control changes/updates; c) Ability to create repeat regions most representative of worldwide populations, or representative of maximal complexity. Three arguments against creating such a reference genome are: a) Significant effort would be required for curation, maintenance, version control, and enforcement of general use within the forensic community, b) Duplication of existing effort/infrastructure, c) Impact on established bioinformatic methods.

   If it is useful to have forensic-specific references for loci/regions of interest, this can be accomplished by designating STRSeq GenBank records as representative of characteristics, e.g. most common flanking region sequence or most complex repeat region. The annotated reference alleles could be provided in the “STR Seq Nomenclature” page of STRidER, where the Forensic STR Sequence Structure guide is currently made available (https://strider.online/nomenclature).

4. **Resources:** To ensure all interested parties have access to existing resources, we provide the following tables of population STR sequence data and STR sequence software/tools.

   4.1. STR Sequence Population Data

   Table 1 contains publications which include at least 50 population samples, with citations ordered by publication date. Populations listed are as defined in the publication.

   4.2. STR Sequence Analysis Software

   Table 2 contains a list of software currently available for STR sequence analysis and citations or links to additional information.

A final topic, on which a philosophical discussion focused, was that of thresholds; specifically, how thresholds may be implemented more intelligently for sequence data than has been possible for traditional CE methods. Sequencing STR loci allows users to differentiate erroneous sequences of the same length as genomic alleles. With traditional CE methods, amplification errors are incorporated into the RFU intensity of the allele. The discussion centered on the possibility of incorporating into the allele read depth a validated level of sequences determined to have originated from the parent allele, rather than attempting to exclude such sequences via thresholds. This approach could clarify when additional contributors are present in mixed DNA samples and might allow for lower analytical thresholds in general. Furthermore, the possibility of integrating a validated level of sequence-based stutter into the parent allele read depth, was raised. These forward-thinking concepts are presented to encourage discussion; as more thorough exploration of such ideas is beyond the scope of this paper.

Lack of nomenclature is often named as a roadblock to STR sequencing implementation; therefore, our ultimate goal is an official (ISFG) recommendation on STR sequence nomenclature. This follows the tradition of STR allele designation guidelines coming from the ISFG [16, 17] and further evolving as the technology expanded (e.g. Y-STRs [18, 19]). Such an approach encourages a rigorous, science-based system. We view this meeting as the first step towards STR nomenclature recommendations; the STRAND WG is committed to facilitating continued dialogue among practitioners, researchers, vendors, and database representatives.

With this communication, we invite the broader forensic community to actively contribute in these discussions. Individuals interested in receiving future communications and/or meeting invitations from the STRAND Working Group may register by email strandwg@gmail.com (please include a brief description of your work in STR sequencing/bioinformatics). Feedback emailed to strandwg@gmail.com will be distributed and discussed at future STRAND Working Group meetings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

[1]. Bodner M, Bastisch I, Butler JM, Fimmers R, Gill P, Gusmao L, Morling N, Phillips C, Prinz M, Schneider PM, Parson W, Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER), Forensic Sci Int Genet 24 (2016) 97–102. [PubMed: 27352221]

[2]. Gettings KB, Borsuk LA, Ballard D, Bodner M, Budowle B, Devesse L, King J, Parson W, Phillips C, Vallone PM, STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci, Forensic Sci Int Genet 31 (2017) 111–117. [PubMed: 28888135]

[3]. King JL, Wendt FR, Sun J, Budowle B, STRait Razor v2s: Advancing sequence-based STR allele reporting and beyond to other marker systems, Forensic Sci Int Genet 29 (2017) 21–28. [PubMed: 28343097]

[4]. Parson W, Ballard D, Budowle B, Butler JM, Gettings KB, Gill P, Gusmao L, Hares DR, Irwin JA, King JL, Knijff P, Morling N, Prinz M, Schneider PM, Neste CV, Willuweit S, Phillips C, 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. [PubMed: 26844919]

[5]. Phillips C, Gettings KB, King JL, Ballard D, Bodner M, Borsuk L, Parson W, "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. [PubMed: 29486434]

[6]. Novroski NM, King JL, Churchill JD, Seah LH, Budowle B, Characterization of genetic sequence variation of 58 STR loci in four major population groups, Forensic Sci Int Genet 25 (2016) 214–226. [PubMed: 27697609]

[7]. Devesse L, Ballard D, Davenport L, Riethorst I, Mason-Buck G, Court DS, Concordance of the ForenSeq™ system and characterisation of sequence-specific autosomal STR alleles across two major population groups, Forensic Science International: Genetics (2017).

[8]. Phillips C, Devesse L, Ballard D, van Weert L, de la Puente M, Melis S, Alvarez Iglesias V, Freire-Aradas A, Oldroyd N, Holt C, Syndercombe Court D, Carracedo A, Lareu MV, 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. [PubMed: 30101987]

[9]. Gettings KB, Borsuk LA, Steffen CR, Kiesler KM, Vallone PM, Sequence-based US population data for 27 autosomal STR loci, Forensic Science International: Genetics 37 (2018) 106–115.

[10]. Alonso A, Barrio PA, Muller P, Kocher S, Berger B, Martin P, Bodner M, Willuweit S, Parson W, Roewer L, Budowle B, Current state-of-art of STR sequencing in forensic genetics, Electrophoresis 39(21) (2018) 2655–2668. [PubMed: 29750373]

[11]. Young B, Faris T, Armogida L, A nomenclature for sequence-based forensic DNA analysis, Forensic Science International: Genetics 42 (2019) 14–20.

[12]. Just RS, Irwin JA, 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. [PubMed: 29525576]

[13]. Bleka O, Storvik G, Gill P, EuroForMix: An open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts, Forensic Sci Int Genet 21 (2016) 35–44. [PubMed: 26720812]

[14]. Urquhart A, Kimpton CP, Downes TJ, Gill P, 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.

[15]. Puers C, Hammond HA, Jin L, Caskey CT, Schumm JW, 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 (1993) 953–958. [PubMed: 8105685]

[16]. Bar W, Brinkmann B, Lincoln P, Mayr WR, Rossi U, 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.

[17]. Bar W, Brinkmann B, Budowle B, Carracedo A, Gill P, Lincoln P, Mayr WR, Olaisen B, DNA recommendations: Further report of the DNA Commission of the ISFH regarding the use of short tandem repeat systems, Int.J.Legal Med. 110(4) (1997) 175–176. [PubMed: 9274938]

[18]. Gill P, Brenner C, Brinkmann B, Budowle B, Carracedo A, Jobling MA, de Knijff P, Kayser M, Krawczak M, Mayr WR, Morling N, Olaisen B, Pascali V, Prinz M, Roewer L, Schneider PM, Sajantila A, Tyler-Smith C, DNA commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs, Int.J.Legal Med. 114(6) (2001) 305–309. [PubMed: 11508794]

[19]. Gusmao L, Butler JM, Carracedo A, Gill P, Kayser M, Mayr WR, Morling N, Prinz M, Roewer L, Tyler-Smith C, Schneider PM, 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 Legal Med (2005) 1–10.

[20]. King JL, Churchill JD, Novroski NMM, Zeng X, Warshauer DH, Seah LH, Budowle B, Increasing the discrimination power of ancestry- and identity-informative SNP loci within the ForenSeq DNA Signature Prep Kit, Forensic Sci Int Genet 36 (2018) 60–76. [PubMed: 29935396]

[21]. van der Gaag KJ, de Leeuw RH, Hoogenboom J, Patel J, Storts DR, Laros JF, de Knijff P, Massively parallel sequencing of short tandem repeats-Population data and mixture analysis results for the PowerSeq system, Forensic Sci Int Genet 24 (2016) 86–96. [PubMed: 27347657]

[22]. Wendt FR, Churchill JD, Novroski NM, King JL, Ng J, Oldt RF, McCulloh KL, Weise JA, Smith DG, Kanthaswamy S, Budowle B, Genetic analysis of the Yavapai Native Americans from West-Central Arizona using the Illumina MiSeq FGx forensic genomics system, Forensic Sci Int Genet 24 (2016) 18–23. [PubMed: 27243782]

[23]. Wendt FR, King JL, Novroski NM, Churchill JD, Ng J, Oldt RF, McCulloh KL, Weise JA, Smith DG, Kanthaswamy S, Budowle B, Flanking region variation of ForenSeq DNA Signature Prep Kit STR and SNP loci in Yavapai Native Americans, Forensic Sci Int Genet 28 (2017) 146–154.

[24]. Casals F, Anglada R, Bonet N, Rasal R, van der Gaag KJ, Hoogenboom J, Solé-Morata N, Comas D, Calafell F, Length and repeat-sequence variation in 58 STRs and 94 SNPs in two Spanish populations, Forensic Science International: Genetics 30 66–70. [PubMed: 28633070]

[25]. Silva D, Sawitzki FR, Scheible MKR, Bailey SF, Alho CS, Faith SA, Genetic analysis of Southern Brazil subjects using the PowerSeq AUTO/Y system for short tandem repeat sequencing, Forensic Sci Int Genet 33 (2018) 129–135. [PubMed: 29275088]

[26]. Borsuk L, Gettings KB, Steffen CR, Kiesler KM, Vallone PM, Sequence-based US population data for the SE33 locus Electrophoresis 0 (2018) 1–8.

[27]. Huszar TI, Jobling MA, Wetton JH, A phylogenetic framework facilitates Y-STR variant discovery and classification via massively parallel sequencing, Forensic Sci Int Genet 35 (2018) 97–106. [PubMed: 29679929]

[28]. Kim SY, Lee HC, Chung U, Ham SK, Lee HY, Park SJ, Roh YJ, Lee SH, Massive parallel sequencing of short tandem repeats in the Korean population, Electrophoresis 39(21) (2018) 2702–2707. [PubMed: 30084488]

[29]. Salvador JM, Apaga DLT, Delfin FC, Calacal GC, Dennis SE, De Ungria MCA, Filipino DNA variation at 12 X-chromosome short tandem repeat markers, Forensic Sci Int Genet 36 (2018) e8–e12. [PubMed: 29909139]

[30]. Hussing C, Bytyci R, Huber C, Morling N, Borsting C, The Danish STR sequence database: duplicate typing of 363 Danes with the ForenSeq DNA Signature Prep Kit, Int J Legal Med 133(2) (2019) 325–334. [PubMed: 29797283]

[31]. Hwa HL, Wu MY, Chung WC, Ko TM, Lin CP, Yin HI, Lee TT, Lee JC, Massively parallel sequencing analysis of nondegraded and degraded DNA mixtures using the ForenSeq system in combination with EuroForMix software, Int J Legal Med 133(1) (2019) 25–37. [PubMed: 30374565]

[32]. Wu J, Li JL, Wang ML, Li JP, Zhao ZC, Wang Q, Yang SD, Xiong X, Yang JL, Deng YJ, Evaluation of the MiSeq FGx system for use in forensic casework, Int J Legal Med 133(3) (2019) 689–697. [PubMed: 30604102]

[33]. Barrio PA, Martin P, Alonso A, Muller P, Bodner M, Berger B, Parson W, Budowle B, Consortium D, Massively parallel sequence data of 31 autosomal STR loci from 496 Spanish individuals revealed concordance with CE-STR technology and enhanced discrimination power, Forensic Sci Int Genet 42 (2019) 49–55. [PubMed: 31252251]

[34]. Hoogenboom J, van der Gaag KJ, de Leeuw RH, Sijen T, de Knijff P, Laros JF, FDSTools: A software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise, Forensic Sci Int Genet 27 (2017) 27–40. [PubMed: 27914278]

[35]. Lee JC, Tseng B, Chang LK, Linacre A, SEQ Mapper: A DNA sequence searching tool for massively parallel sequencing data, Forensic Sci Int Genet 26 (2017) 66–69. [PubMed: 27792894]

[36]. Woerner AE, King JL, Budowle B, Fast STR allele identification with STRait Razor 3.0, Forensic Science International: Genetics (2017).

[37]. Friis SL, Buchard A, Rockenbauer E, Borsting C, Morling N, Introduction of the Python script STRinNGS for analysis of STR regions in FASTQ or BAM files and expansion of the Danish STR sequence database to 11 STRs, Forensic Sci Int Genet 21 (2016) 68–75. [PubMed: 26722765]

[38]. Ganschow S, Silvery J, Kalinowski J, Tiemann C, toaSTR: A web application for forensic STR genotyping by massively parallel sequencing, Forensic Sci Int Genet 37 (2018) 21–28. [PubMed: 30071493]

```
CE11_TATC[8]TGTC[1]TATC[3]AATC[1]ATCT[3]
CE11_TATC[10]AATC[3]ATCT[3]
CE11_TATC[11]AATC[2]ATCT[3]
CE11_TATC[12]AATC[1]ATCT[3]
CE11_TATC[12]AATC[1]ATCT[3]_-24G>A
CE11_TATC[12]AATC[1]ATCT[3]_-25C>T
CE11_TATC[13]ATCT[3]
CE12_TATC[7]TATT[1]TATC[5]AATC[1]ATCT[3]
CE12_TATC[12]AATC[2]ATCT[3]
CE12_TATC[13]AATC[1]ATCT[3]
CE12_TATC[13]AATC[1]ATCT[3]_-24G>A
CE12_TATC[13]AATC[1]ATCT[3]_-25C>T
CE12_TATC[13]AATC[2]ATCT[2]
CE12_TATC[14]ATCT[3]
CE13_TATC[13]AATC[2]ATCT[3]
CE13_TATC[14]AATC[1]ATCT[3]
CE13_TATC[14]AATC[1]ATCT[3]_-24G>A
CE13_TATC[14]AATC[1]ATCT[3]_-25C>T
CE13_TATC[15]AATC[1]ATCT[3]_+9GTCT>-
CE13_TATC[15]ATCT[3]
```

Fig. 1.
Example of automated bracketing results for a collection of alleles a the D13S317 locus.

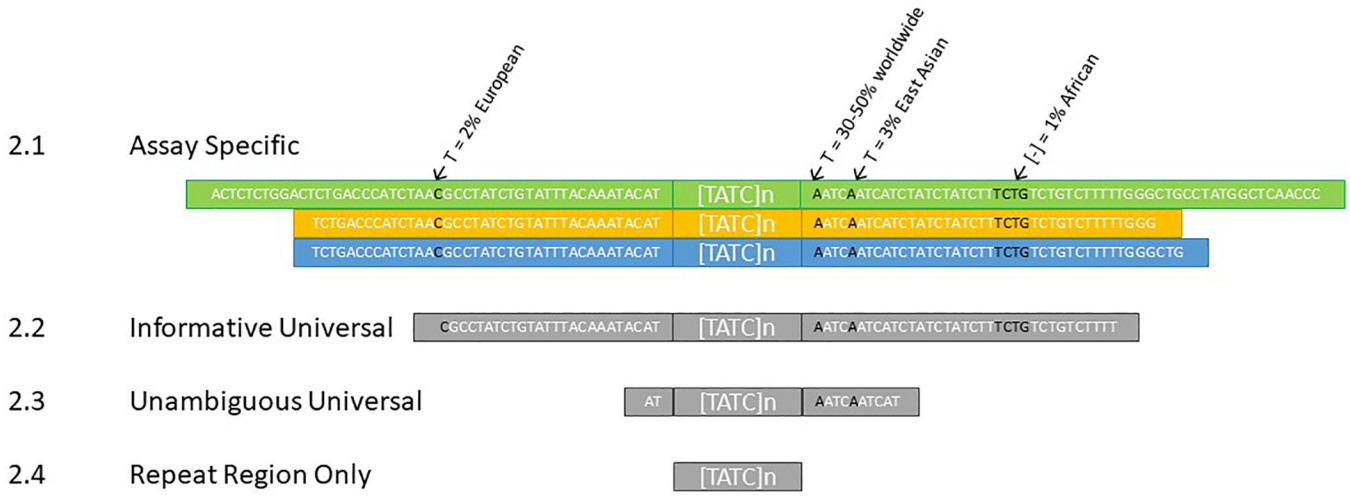

Fig. 2.

Four possible range definitions applied to the D13S317 locus. Flanking region polymorphisms > 1% frequency are shown, associated rs numbers are (left to right) rs73250432, rs9546005, rs202043589, rs561167308.

Table 1.

Publications containing STR sequence population data

Citation	Year	First Author	Total Number of Samples	Populations	Sequenced STR Loci	Additional data	Bioinformatic Method(s)
[6]	2016	Novroski	777	Caucasian	27 Autosomal STR	CE-STR	ForenSeq UAS
				Hispanic	24 Y-STR		STRait Razor v2.0
				African American	7 X-STR		
				East Asian			
[21]	2016	van der Gaag	297	Netherlands	17 Autosomal STR	CE-STR	TSSV (FDSTools)
				Nepal			
				Bhutan			
				Central African Pygmy			
[22, 23]	2016, 2017	Wendt	62	Yavapai	27 Autosomal STR	94 iiSNP	STRait Razor v2s
					24 Y-STR	56 aiSNP	
					7 X-STR	22 piSNP	
[24]	2017	Casals	231	Spanish Roma	27 Autosomal STR	94 iiSNP	ForenSeq UAS
				Catalans	24 Y-STR		
					7 X-STR		
[25]	2017	Silva	59	South Brazilian	22 Autosomal STR	CE-STR	Altius Cloud System
					23 Y-STR		
[26]	2018	Borsuk	1036	Caucasian	1 Autosomal STR (SE33)	CE-STR	STRait Razor v2.0
				African American			
				Hispanic			
				Asian			
[7]	2018	Devesse	400	White British British	27 Autosomal STR	CE-STR	ForenSeq UAS
				Chinese			
[9]	2018	Gettings	1036	Caucasian	27 Autosomal STR	CE-STR	ForenSeq UAS
				African American			STRait Razor v2.0
				Hispanic			
				Asian			
[27]	2018	Huszar	100	African	23 Y-STR	CE-STR	FDSTools vl.1.1
				European			
				Australian			
				Asian			
				Near and Middle Eastern			
				American			
[28]	2018	Kim	209	Korean	27 Autosomal STR	CE-STR	ForenSeq UAS
					24 Y-STR		

Citation	Year	First Author	Total Number of Samples	Populations	Sequenced STR Loci	Additional data	Bioinformatic Method(s)
					7 X-STR		
[8]	2018	Phillips	944	CEPH (51 populations)	27 Autosomal STR	CE-STR	ForenSeq UAS
					24 Y-STR		
					7 X-STR		
[29]	2018	Salvador	143	Filipino	7 X-STR	CE-STR	ForenSeq UAS STRait Razor v2s
[30]	2019	Hussing	363	Danish	26 Autosomal STR	CE-STR 94 iiSNP	STRinNGS 1.0
					24 Y-STR	56 aiSNP	ForenSeq UAS
					6 X-STR	22 piSNP	
[31]	2019	Hwa	119	Taiwanese	27 Autosomal STR	CE-STR 94 iiSNP	ForenSeq UAS
					24 Y-STR		
					7 X-STR		
[32]	2019	Wu	108	Han Chinese	27 Autosomal STR	CE-STR	ForenSeq UAS
					24 Y-STR		
					7 X-STR		
[33]	2019	Barrio	496	Spanish	31 Autosomal STR	CE-STR	Converge 2.0 STRait Razor v3.0

Table 2.

STR Sequence Analysis Software

Name	Availability
Agnostic, freeware	
FDSTools [34]	Python Package Index; www.fdstools.nl
Seqmapper [35]	http://forensic.mc.ntu.edu.tw:9000/SEQMapperWeb/Default.aspx
STRait Razor v2s [3]	https://www.unthsc.edu/graduate-school-of-biomedical-sciences/molecular-and-medical-genetics/laboratory-faculty-and-staff/strait-razor/
STRait Razor 3.0 [36]	Upon request from the University of Copenhagen https://www.toastr.de/
STRinNGS [37]	https://www.battelle.org/government-offerings/homeland-security-public-safety/security-law-enforcement/forensic-genomics/exactid
ToaSTR [38]	https://softgenetics.com/GeneMarkerHTS.php
Agnostic, for purchase	
ExactID	https://nichevision.com/mixtureace/
GeneMarkerHTS	https://www.thermofisher.com/order/catalog/product/A35131
Armed Expert Mixture Ace	https://verogen.com/products/
Assay specific, for purchase	
Converge	https://www.thermofisher.com/order/catalog/product/A35131
Universal Analysis Software	https://verogen.com/products/