King's Research Portal

DOI:

[10.1016/j.fsigen.2018.02.017](https://doi.org/10.1016/j.fsigen.2018.02.017)

Document Version

Peer reviewed version

[Link to publication record in King's Research Portal](#)

Citation for published version (APA):

Phillips, C., Gettings, K. B., King, J. L., Ballard, D., Bodner, M., Borsuk, L., & Parson, W. (2018). "The devil's in the detail": Release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide. *Forensic Science International: Genetics*, 34, 162-169. <https://doi.org/10.1016/j.fsigen.2018.02.017>

Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Accepted Manuscript

Title: “The devil’s in the detail”: Release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide

Authors: C. Phillips, K.B. Gettings, J.L. King, D. Ballard, M. Bodner, L. Borsuk, W. Parson

PII: S1872-4973(18)30111-X
DOI: <https://doi.org/10.1016/j.fsigen.2018.02.017>
Reference: FSIGEN 1858

To appear in: *Forensic Science International: Genetics*

Received date: 27-11-2017
Revised date: 11-2-2018
Accepted date: 18-2-2018

Please cite this article as: C.Phillips, K.B.Gettings, J.L.King, D.Ballard, M.Bodner, L.Borsuk, W.Parson, “The devil’s in the detail”: Release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide, *Forensic Science International: Genetics* <https://doi.org/10.1016/j.fsigen.2018.02.017>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

“The devil’s in the detail”: Release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide

C. Phillips^{a*}, K.B. Gettings^b, J.L. King^c, D. Ballard^d, M. Bodner^e, L. Borsuk^b, W. Parson^{e,f}

^a Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Spain

^b National Institute of Standards and Technology, Biomolecular Measurement Division, Gaithersburg, MD, USA

^c Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, Fort Worth, TX, USA

^d King’s Forensics, King’s College London, Franklin-Wilkins Building, London, UK

^e Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria

^f Forensic Science Program, The Pennsylvania State University, University Park, PA, USA

*Corresponding author, E-mail address: c.phillips@mac.com

Highlights

- The sequence template file originally published with 2016 ISFG STR sequence minimal nomenclature requirements has been comprehensively revised
- The number of forensic STRs listed has been expanded to include loci less commonly used but requiring detailed sequence descriptions
- A wide range of sequence variation features in and around forensic STRs have been identified and the revised sequence guide provides new annotations of these variants
- The sequence guide is now available as an FTP download that allows dynamic revisions and inclusion of a date-stamped change log to detail future changes

Abstract

The STR sequence template file published in 2016 as part of the considerations from the DNA Commission of the International Society for Forensic Genetics on minimal STR sequence nomenclature requirements, has been comprehensively revised and audited using the latest GRCh38 genome assembly. The list of forensic STRs characterized was expanded by including supplementary autosomal, X- and Y-chromosome microsatellites in less common use for routine DNA profiling, but some likely to be adopted in future massively parallel sequencing (MPS) STR panels. We outline several aspects of sequence alignment and annotation that required care and

attention to detail when comparing sequences to GRCh37 and GRCh38 assemblies, as well as the necessary matching of MPS-based allele descriptions to previously established repeat region structures described in initial sequencing studies of the less well known forensic STRs. The revised sequence guide is now available in a dynamically updated FTP format from the STRidER website with a date-stamped change log to allow users to explore their own MPS data with the most up-to-date forensic STR sequence information compiled in a simple guide.

Keywords: Massively parallel sequencing MPS; Short tandem repeat STR; Indels; SNPs; dbSNP; sequence alignment

1. Introduction

In 2016, an Excel-based STR sequence template file accompanied the set of considerations published by the DNA Commission of the International Society for Forensic Genetics (ISFG) on minimal STR sequence nomenclature requirements [1]. The publication of these considerations was designed to foster consensus in the forensic community about the optimum way to arrange sequence alignments, variant annotation and an eventual allele nomenclature framework necessary for mainstream use of massively parallel sequencing (MPS) to genotype forensic STRs. The first principal guideline was a directive requiring STR sequences to conform to the standardized system, applied to all human microsatellites, of alignment to the genome reference sequence: a haploid, single-strand nucleotide string arranged in a unified p-arm to q-arm direction per chromosome. The second principal guideline recommended that variant annotation: the systematic description of genome sequence differences between individuals, should use the locus identifiers and novel variant reporting methods applied in the 1000 Genomes and NCBI dbSNP databases. It was recognized at the time of publication that sequence variation within the repeat regions of microsatellites presents particular challenges when tracking sequence changes relative to the human reference sequence, which would require care and a period of time to allow early adopters of forensic MPS systems to compile sufficient sequencing data. The STR sequence template file embodied these guidelines by summarizing each STR's sequence alignment and variant/repeat region annotations. As well as mapping the relevant segments of the human reference sequence for each STR, all recorded flanking region variants with more than 10% polymorphism (in one or more population groups) were placed in the context of the STR's repeat region. Annotation was extended to less frequent variants that become important when differentiating repeat region nucleotides from those in flanking regions (e.g. SNPs creating an uncounted repeat unit motif immediately next to the first or last true repeat - see Fig. 1 of [1]). Therefore, defining each STR's repeat region start and end points became the keystone for defining the allelic structure of the marker and protecting its backward compatibility to capillary electrophoresis (CE) genotypes populating all national DNA databases.

With the need for precision and detail in mind, the STR sequence working group (this authorship) have used the original template file as the main data-exchange facility in order to easily update or add variant annotations, as well as explore additional STR loci and re-align many of the originally published repeat region bounds. Periodic releases of a sequence template file continuously revised in this way would lead to confusion, multiple versions in common use and conflicting sequence descriptions. The template file also needs to be regularly calibrated to the most up-to-date human genome sequence build. Therefore, a clear need exists for a dynamic version of the sequence template file that can be placed in an open-access online file transfer scheme. Such a

framework is best arranged in a dedicated FTP site with a date-stamped change-log updated in each file release. The STRidER database [2] has been set up to manage the compilation of forensic STR variation data in the MPS era - providing the obvious host site for a dynamic FTP version of the sequence template file.

This paper reports the release of a comprehensively revised sequence template file, herein the STR Sequence Guide, as an FTP file that can be downloaded periodically by forensic DNA practitioners to keep STR sequence information up-to-date. We briefly describe the surprisingly wide range of STR structure and sequence variation factors that the working group considered when revising the sequence data. Detailed descriptions of STR sequences are just as important for established CE genotyping regimes as MPS. The recent discovery of ambiguity in the genomic descriptions of forensic STRs when comparing CE and MPS data [3,4] highlights the need for greater care and accuracy when mapping microsatellites during their initial development for forensic adoption. For this reason, the listed STRs have been expanded to include an additional 45 autosomal, Y-chromosome and X-chromosome loci less commonly used in forensic DNA profiling.

2. Results and discussion

The revised STR Sequence Guide is available for download from the STRidER website at: <https://strider.online>

In addition to revisions of the STR sequence data and inclusion of less commonly used forensic STRs, a detailed change log has been added listing all changes made to the original sequence template file at the date the revision is checked, agreed (by the authorship) and compiled. An additional worksheet lists all STRs as simple FASTA sequence strings with their GRCh38 chromosome coordinates and individual GRCh38 nucleotides (i.e. one per Excel cell) within the stated 'bedfile' sequence segment bounds.

2.1. Extra forensic STRs added to the original sequence template file

Coverage of forensic STRs was expanded by adding: i. a significant number of autosomal STRs developed for supplemented forensic analyses (e.g. in complex kinship tests), comprising all the newly-adopted markers in supplementary CE kits that were compiled in the 2017 study of Phillips [5]; ii. a further five X-chromosome STRs of the panel of twelve analyzed by the Qiagen Argus X-12 CE kit; iii. additional rapidly-mutating Y-chromosome STRs (RM Y-STRs), yet to be adopted in

MPS panels but of interest and forensic utility. All previously listed and additional forensic STRs now included in the revised STR Sequence Guide are detailed in Table 1.

Although MPS analysis of some of the above STRs may not be possible at present, the technology and sequence alignment algorithms to analyze sequences continue to improve to the point where the complex sequence structure of SE33 may be amenable to genotyping by MPS in the near future. More importantly, many STRs are genotyped by CE in mainstream forensic analyses and interpretation of allelic patterns is improved when reference can be made to their repeat region sequence structures. Sequence data is now divided into four worksheets, adapting the original S1 designations: S1A, common use autosomal A-STRs (35 loci); S1B, common use XY-STRs (29 Y and 7 X loci); S1C, additional A-STRs (34 loci); and S1D, additional XY-STRs (6 RM Y-STRs and 5 X loci). The misidentified D5S2800 STR included in the Thermo Fisher Precision ID STR MPS panel has been placed in the common use A-STRs worksheet, alongside D5S2500 used in several CE kits.

2.2. Audit of GRCh38 reference genome builds released between 2013 and 2017

In the original template file, reference sequence was collected from the 1000 Genomes database and cross-checked against the chromosome coordinates of the two main genome assemblies of GRCh37 and GRCh38, using the *In Silico PCR* web-tool to map sequences in the first GRCh38 build (released 17-December-2013). The GRCh38 assembly has undergone periodic revisions that have identified a series of sequence inversions, segmental duplications and translocations with increasing precision. Therefore, a fresh review was made of the most recent stable GRCh38 genome build, GRCh38.p10 (released 1-June-2017), which is held in the Ensemble sequence repository [6]. A new build has since been published: GRCh38.p11, released 14-June-2017, but is not yet viewable in the Ensemble genome browser (each GRCh38 build is available to download at a dedicated NCBI site [7]).

The comparison of each GRCh38 genome build showed no nucleotide differences at any positions originally listed in the sequence template file.

Some additional XY-STRs showed differences in sequence arrangements between GRCh37 and GRCh38 assemblies, similar in character or more complex, than the differences listed for DYS437, DYS438 and DYS439 in the original template file. Simple STR sequence differences were found in DXS10134, showing two more [GAAA] repeats in GRCh38. The DXS10146 sequence is more complex, with GRCh37 an inversion of the GRCh38 sequence, and four sets of nucleotide differences, including an extra T nucleotide (nt) in the GRCh37 assembly at

X:149584331; and an extra [AAAG] repeat unit in GRCh38 at X:150403993-150403996). As a result, 27 repeats for DXS10146 in GRCh38, differs from the Qiagen Argus X-12 CE kit guide listing of 26, based on the original X-STR sequence studies of Edelmann [8]. To match CE and MPS repeat region annotations, an uncounted sequence tract was extended to include the GRCh38-only AAAG nucleotides described above - losing one repeat (uncounted sequence revised to GRCh38 X:150403993-150404002). With these adjusted allele numbers in place for GRCh38, there are also contrasting repeat motifs between each assembly at repeats 13 and 19.

2.3. Revised STR repeat region sequence structure summaries

In the original sequence template file, the summary descriptions of each STR's repeat region structure (in column B) summarized the patterns of repeated motifs seen in the reference sequence. We took the opportunity to revise these summary descriptions to better reflect common variation in forensic STR repeat region structures seen in everyday practice, which may not always be represented by the repeat region in the reference sequence. In some loci there was also some inconsistency in the way complex repeat region structures were formatted to describe the repeat motifs and any uncounted nucleotide tracts between them. Simple format rules were refined for forensic STR repeat region descriptions and applied to their revised data, which now occupy column D, while columns B and C in the STR Sequence Guide now list the bounding GRCh38 coordinates for the whole sequence shown and the repeat region, plus the length of each sequence.

STR repeat region formatting rules apply to: 1) repeating elements (usually termed motifs) and; 2) uncounted nucleotide tracts imbedded within repeat regions. Table 2 outlines these formatting rules in detail with relevant examples. Rule 1.1 applies the established regime of square brackets around each distinct repetitive motif followed by the repeat number applicable. The Guide's repeat region structure summaries now have the suffix 'n' for each bracketed motif, signifying a repeat number would normally be relevant. Rule 1.2 applies a single space between multiple elements (separating each motif and uncounted tracts), if present in the repeat region. Rule 1.3 indicates non-repetitive motifs by an absence of brackets. This rule allows for flexibility in future STR reporting - should such motifs be observed in multiple copies; they can be bracketed. Two examples of adaptation of rule 1.3 in the STR Sequence Guide are the formatting of the first [ATGT]_n motif in D6S1043 and the final [TCTA]_n motif in D21S11. Multiple copies were observed for the first D6S1043 [ATGT] motif in a 20-repeat allele (KBG), and in the last D21S11 [TCTA] motif in multiple examples of 32.2 and 33.2 alleles (CP). Rule 1.4 describes non-standard motifs that create intermediate alleles, generally from deleted nucleotides, as separate elements. The widely observed TH01 9.3 allele is therefore properly described as: [AATG]₆ ATG [AATG]₃.

Rule 2 applies lower case formatting to uncounted nucleotide tracts within repeat regions. Uncounted nucleotides lack within-tract spaces and numerical descriptions. For brevity, long uncounted tracts of ten or more nucleotides are simply described with a number in the STR Sequence Guide, e.g. DYS448 [AGAGAT]_n N42 [AGAGAT]_n. However, proper reporting of such tracts in individual STR sequences should provide the full sequence string, as shown in the example of DYS449 in Table 2.

It is important to note that eleven STR repeat region descriptions now differ from those given in the original sequence template file, as it has been necessary to emphasize the description of common forensic STR variation in the repeat region structure summaries, rather than certain patterns seen in the reference sequence. The details of these STRs are shown in Table 3. The bulk of these revised repeat region descriptions indicate common intermediate alleles from shorter motifs, now included in the appropriate format, but D21S11 is a notable example of the repeat region variation at the 3' endpoint of: AT [TCTA]_n; commonly seen in longer D21S11 alleles, but absent from the reference sequence.

2.4. Inverted multiple-allele Y-STRs

The most complex STR sequence patterns were seen in Y-STRs, in particular RM Y-STRs with multiple amplified sequences and therefore showing two, three or four alleles that may or may not have different sizes in CE. In all cases of multiple sequences, the most 'upstream' 5' sequence position was used to set the direction for each read and therefore the repeat region structure (e.g. DYS385b is 5' upstream of DYS385a and has [TTTC] repeats, so the DYS385a sequence is inverted to match). Additional XY-STRs with multiple inverted sequences were: DYF403S1 (4 fragments), DYF399S1 (4) and DYF404S1 (2). DYF399S1 and DYF404S1 were given provisional labels of "fragments" 1-4 and 1-2 respectively, based on their relative 5' positions, as it is not viable to match the very similar repeat numbers in each sequence to patterns observed in CE genotyping of these loci.

The four sequences of DYF403S1 were labeled according to the recommendations of Lee [9] as: DYF403S1a (312 nt in Lee's study); DYF403S1a (316 nt); DYF403S1b1 (341 nt); DYF403S1b2 (437 nt). Note that the two DYF403S1a fragments are identical in their flanking regions so are not be distinguishable as individual sequences. Although this is not a problem in routine genotyping, it shows that identical flanking sequences for two amplified fragments leads to an inability to identify them individually. During developmental studies of DYF403S1 in one contributing laboratory (DB), a fifth sequence fragment of 342 nt was detected using the primers of Lee [9].

Significantly, this fifth amplified sequence is only listed for GRCh38 with *In Silico PCR*, not for the GRCh37 genome assembly. Further investigation revealed the 342 nt fragment was actually the STR DYS627, where enough primer sequence homology has remained between each STR to allow some low level amplification (approximately 20%) of DYS627 from DYF403S1 primers (Supplementary File S1). Therefore, it can be concluded that a proportion of RM Y-STRs represent replicated sequences of previously established Y-chromosome microsatellites that have diverged sufficiently to become distinct STRs with locus-specific allelic variation and in certain cases differentiated repeat region structures.

Many multiple allele Y-STRs have relatively large distances between their sequence positions and are interspersed with single-sequence Y-STRs. Fig. 1 summarizes patterns of these Y-STR sequence positions in two 4-megabase sections of the Y-chromosome.

2.5. Incidental microsatellites

Two 'incidental' STRs were identified in the flanking region of the target STRs, making use of the comprehensive 2006 survey of forensic Y-chromosome STRs by Hansen and J Ballantyne [10]. First, STR DYS461 (GRCh38, Y:18888804-18888851) is separated by 104 nt on the 5' side of DYS460 and has been annotated in the same way as the target STR. DYS460 primers used for CE analysis bind between each STR, so DYS461 does not influence DYS460 fragment length estimations. However for analysis of DYS460 with the Illumina Forenseq DNA Signature kit, both STRs are amplified together but DYS461 variation is not reported. Second, STR DYS467 is separated by 50 nt on the 3' side of DYS389-II (GRCh38, Y:12500662-12500709). DYS467 is also circumvented by existing CE and MPS primers and may not be highly polymorphic, although it comprises 12/14 GATA repeats (two more repeats in GRCh37), suggesting a standard microsatellite locus. Although incidental STRs closely sited to the target STR may not always be amplified in forensic tests, we decided it is informative to track all polymorphisms found in flanking regions, not just SNPs and Indels.

The RM Y-STR DYS526 listed in additional XY-STRs, was reported in the 2010 study of K Ballantyne et al [11] as two loci: DYS526a and DYS526b. However, Hansen and J Ballantyne identified DYS526a as independent STR DYS505, separated by 93 nt on the 3' side of the DYS526b locus [10]. Both STRs are included in the sequence details of DYS526.

2.6. Mobility-shift SNPs

Studies of SE33 and DYS481 sequence variation have identified a mobility shift effect from the presence of flanking region SNP variant alleles creating altered DNA folding patterns [12,13]. Although denaturing CE protocols should reduce formation of secondary structures, the effect appears to be consistent in certain kits and explainable from modeling the stem and loop structures formed by the sequence change (e.g. where a SNP variant allele forms a new C-G triple bond). The three SE33 SNP variants comprise: rs549958510-A; rs189881506-T; rs538644460-T; and the DYS481 SNP variant is rs368663163-A. As it is important to match sequence-based repeat number data with genotyping from CE fragment length estimations with the knowledge of potential discordant genotypes, we have highlighted the presence of the above SNPs with simple orange labels.

Two additional mobility shift SNPs have recently been identified in the Penta E and D2S441 flanking regions and were added to the Sequence Guide. The Penta E SNP variant is rs188309642-G, creating a -1 nt mobility shift confined to 11 repeat alleles in this STR [14]. The D2S441 flanking region variant is a G>C substitution creating a -2 nt mobility shift which does not have an rs-number, located at: GRCh38, 2:68011921 (personal communication, Rita Weispfenning, Promega)

2.7. Compiling insertion-deletion polymorphisms sited in repeat and flanking regions

It is important to track all insertion-deletion polymorphisms (Indels) within repeat regions as they can create a high level of variability in the sequence; are often population specific; can be ambiguously positioned; and when combined with Indels in flanking regions may create isometric fragments that go undetected by CE. We found it difficult to match the 1000 Genomes or dbSNP annotations of Indels sited in repeat regions with the accumulating knowledge of these variants from forensic MPS studies.

The most common forensic STR repeat region Indel that forms the TH01 9.3 allele, provides a good illustration of the difficulties of identifying such Indel positions. All TH01 9.3 MPS sequences collected to date show an [A/-] deletion in the seventh repeat which can be annotated as: [AATG]6 ATG [AATG]3. Previous sequence studies of rarer TH01 alleles indicated 6.3 = [AATG]3 ATG [AATG]3 [15]; 8.3 = [AATG]5 ATG [AATG]3; and 10.3 = [AATG]6 ATG [AATG]4 [16]. The reference sequence consists of 7 AATG repeats (GRCh38 11:2171088-2171115, placing the deleted A nucleotide in 9.3 alleles at 11:2171112). However, dbSNP reports the 9.3 sequence change as the [-/AATGAATGATG] 11 nt insertion rs763206927 (GRCh38 11:2171103). The 1000 Genomes annotations of TH01 consist of the [-/GTGAA/GTGAATGAA] 5 nt and 9 nt insertion alleles in rs554658416 (GRCh38 11:2171084, 5' upstream of the repeat region), plus the [-

/ATGAATGAATG] 11 nt insertion rs375879846 (GRCh38 11:2171100). The rs375879846 variant in 1000 Genomes matches the 9.3 allele frequency in Europeans (0.2783) and produces the correct sequence, but treating the 9.3 variant as an insertion in order to adjust the 6 repeats of the reference sequence seems counter intuitive when TH01 MPS sequences align in all positions apart from the deleted nucleotide at GRCh38 11:2171112.

Having three alternative ways to annotate the TH01 9.3 allele (Fig. 2) highlights the difficulty of describing copy number variation of one type located within copy number variation of another type, so it should be no surprise that current Indel annotation of repeat region reference sequence is not always consistent and remains incomplete. Therefore, it is a sensible policy to avoid overly precise descriptions of Indels in repeat regions, whether common or rare. Since forensic MPS STR data will be handled as full sequence strings, the annotation of Indels that are likely to occupy different positions in a range of alleles is unnecessary. The existing system of describing repeat regions with bracketed repeat motifs also captures any Indels that occur in these sequence tracts.

Despite the difficulty of annotating Indels, we retained details of two well-characterized Indel loci close to the 3' repeat region endpoints of D18S51 (rs575219471) and D19S433 (rs147936416); because they are key to fixing the repeat region 3' bounds in each STR. Indels positioned in a polymeric sequence tract were placed as insertions or deletions starting at the most 5' nt, following 1000 Genomes and dbSNP conventions. A further two flanking region Indels have modified details from data given in the original template files: D19S433 has the 2 nt Indel rs745607776 moved to GRCh38 19:29926229-29926230; and Penta D now has an unnamed 13 nt [AAGAAAGAAAAA/-] Indel deletion forming 2.2 and 3.2 alleles; changed from a 3 nt deletion placed in the first repeat of the reference sequence in the original template file. As an illustration of how knowledge of forensic STR sequence variation can contribute to a growing database of such variation in dbSNP, the 13 nt Penta D Indel has been given the provisional identifier ss2137535200 and this can provide a "place-holding" link to the variant until it is assigned an rs-number by dbSNP and this is added to the STR Sequence Guide.

A 4 nt deletion has been characterized in the D13S317 flanking region since the original template file was published. This unnamed [ATCT/-] Indel on the 3' side of the D13S317 repeat region at GRCh38, 13:82148077-82148080, is an important factor influencing repeat allele size estimation and has been observed in multiple samples from a range of populations in two contributing laboratories (CP, KBG). Although the deleted nucleotide tract cannot be positioned exactly, we placed it at the start of the deletion at the most 5' nucleotide coordinate. This 4 nt Indel has been given the provisional identifier ss2137543798 by dbSNP.

Two 3' flanking region 4 nt deletions in SE33 that potentially influence repeat allele size estimation have been added at GRCh38 6:88277313-88277316 (provisional identifier ss2137535201) and GRCh38 6:88277355-88277358 (rs369314007). Two 5' flanking region Indels creating intermediate alleles in D7S280 were also added, comprising the [T/-] deletion rs754976988 (GRCh38, 7:84160203; X.3 alleles) and the [T/TA] insertion, provisionally ss2137543824 (GRCh38, 7:84160204; X.1 alleles). Lastly, in D9S1122, the 5' flanking region [TG/-] deletion rs754976988 was added (GRCh38, 9:77073816-77073817; creating X.2 alleles).

3. Concluding remarks: Considerations for moving towards an agreed STR allele nomenclature system in the future

The phrase “the devil’s in the detail” describes how a seemingly simple task can turn out to be more complicated than supposed, as individual details produce unforeseen problems. This has often been the case during the compilation of sequence data, thoroughly revised here from the original sequence template file, in order to strengthen the foundations for a forensic STR allele nomenclature system. A persistent challenge has been the need to match repeat region structures found in the reference sequence and in MPS data, with the repeat allele numbers suggested by early Sanger sequencing analyses of STRs genotyped by CE. We have often used historical precedence, when the first published sequences of an STR allowed a repeat structure to be proposed. However, a period of comparative studies will be increasingly necessary for the less commonly used STRs compiled here. We place importance in the inclusion of as many forensic STRs as possible, since it is likely that MPS multiplexes will continue to expand and the compilation of often little-used STRs provides a properly curated set of genomic details about their sequence characteristics alongside the core STRs. This is particularly important for RM Y-STRs that unsurprisingly, tend to be found in the more unstable regions of the Y-chromosome, which in turn may have influenced choice of these microsatellites in the first generation of forensic MPS STR panels. However, Y-STRs generally appear more prone to multiple sequences; Hansen and Ballantyne [10] observed ~13% of 417 forensic microsatellites on the Y-chromosome had two duplicated sequences (40 Y-STRs) or 3, 4, 5 and 9 duplications (11 Y-STRs). The close similarity in sequence between DYF403S1 and DYS627 we highlight in this report also suggests that replicated STR sequences eventually evolve into differentiated microsatellites. Such STRs can have distinct patterns of repeat variation, but may still retain enough sequence homology to cause problems in distinguishing their amplified fragments when they are combined in the same PCR.

The complexities revealed when GRCh37 and GRCh38 genome assemblies are compared, underlines the importance of a single stable reference sequence to act as the template on which all MPS sequence data can be reliably aligned. At all the forensic STR sequence tracts checked, GRCh38 has not changed in four years of re-assembly, but shows critical differences with GRCh37 at certain nucleotides. This issue is highlighted by the need to re-annotate the repeat region of DXS10146 because GRCh38 differs in sets of nucleotides in four separate positions, from the GRCh37 assembly originally used to map the repeats. We recommend exclusive use of the GRCh38 human genome sequence to align forensic MPS data, but retain the GRCh37 coordinates because publications still commonly map sequence variants to GRCh37 positions. In the case of DXS10146, the Argus X-12 kit’s ladder fragments, control genotypes and supporting

literature made use of the GRCh37 assembly to name the repeat alleles in the component X-STRs. As the 1000 Genomes project has now officially completed its work, the transition to GRCh38 coordinates for all variants in this large-scale catalog of human variation is in process, as we expected to happen [1]. However, the 1000 Genomes browser that lists GRCh37 coordinates, combined with dbSNP that has both, currently provides the best way to check variation found in STR flanking regions.

Assessments of the range of sequence variation in forensic STRs from collective efforts such as STRSeq [17] will accelerate the progress towards an agreed sequence allele nomenclature framework, but these initiatives will be greatly helped by contributions from the whole community. A dynamically revised STR Sequence Guide makes the submission of new sequence discoveries from any forensic MPS practitioner wishing to compare their own data, much more straightforward. Recent discussions and exchange of details in the STR sequence working group have been prompted by revisiting CE information as much as new data generated from MPS. Therefore, anyone with an interest in understanding forensic STR sequences are free to access, and via STRidER, contribute new variant annotations to, the revised STR Sequence Guide launched with this publication.

NIST funding sources and disclaimers

This work was funded in part by the National Institute of Justice (NIJ) interagency agreement 1609-602-18NIJ: "Forensic DNA Applications of Next Generation Sequencing". Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Departments of Commerce or Justice. Certain commercial equipment, instruments and materials 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 materials, instruments or equipment identified are necessarily the best available for the purpose.

UNT funding sources and disclaimers

This work was supported in part by award no. 2015-DN-BX- K067, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the U.S. Department of Justice.

Acknowledgements

The authors wish to thank Rita Weispfenning of Promega for sharing data on the D2S441 mobility shift variant.

References

- [1] W. Parson, D. Ballard, B. Budowle, J.M. Butler, K.B. Gettings, P. Gill, L. Gusmão, D.R. Hares, J.A. Irwin, J.L. King, P. de Knijff, N. Morling, M. Prinz, P.M. Schneider, C. Van Neste, S. Willuweit, C. Phillips, Massively parallel sequencing of forensic STRs: considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements, *Forensic Sci. Int. Genet.* 22 (2016) 54–63.
- [2] M. Bodner, I. Bastisch, J.M. Butler, R. Fimmers, P. Gill, L. Gusmão, N. Morling, C. Phillips, M. Prinz, P.M. Schneider, W. Parson, Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER), *Forensic Sci. Int. Genet.* 24 (2016) 97–102.
- [3] C. Phillips, W. Parson, J. Amigo, J.L. King, M.D. Coble, C.R. Steffen, P.M. Vallone, K.B. Gettings, J.M. Butler, B. Budowle, D5S2500 is an ambiguously characterized STR: identification and description of forensic microsatellites in the genomics age, *Forensic Sci. Int. Genet.* 23 (2016) 19–24.
- [4] M. Whittle, More on the genomic identification of forensic STRs, *Forensic Sci. Int. Genet.* 25 (2016) e1–e2.
- [5] C. Phillips, A genomic audit of newly-adopted autosomal STRs for forensic identification, *Forensic Sci. Int. Genet.* 29 (2017) 193–204.
- [6] The Ensembl GRCh38 genome browser at: <http://www.ensembl.org/index.html? 1000> Genomes GRCh37 variant catalogs, Data Slicer, etc. now hosted by Ensembl, at: <http://grch37.ensembl.org/index.html> Accessed November 2017.
- [7] The NCBI GRCh38 genome assembly/build downloads and details page at: <https://www.ncbi.nlm.nih.gov/assembly/?term=GRCh38&cmd=DetailsSearch> Accessed November 2017.
- [8] J. Edelmann, S. Hering, C. Augustin, R. Szibor, Characterisation of the STR markers DXS10146, DXS10134 and DXS10147 located within a 79.1 kb region at Xq28, *Forensic Sci. Int. Genet.* 2 (2008) 41–46.
- [9] E.Y. Lee, H.Y. Lee, S.Y. Kwon, Y.N. Oha, W.I. Yang, K.J. Shin, A multiplex PCR system for 13 RM Y-STRs with separate amplification of two different repeat motif structures in DYF403S1a, *Forensic Sci. Int. Genet.* 26 (2017) 85–90.
- [10] E.K. Hanson and J. Ballantyne, Comprehensive annotated STR physical map of the human Y chromosome: Forensic implications, *Leg. Med.* 8 (2006) 110–120.
- [11] K.N. Ballantyne, M. Goedbloed, R. Fang, O. Schaap, O. Lao, A. Wollstein, Y. Choi, K. van Duijn, M. Vermeulen, S. Brauer, R. Decorte, M. Poetsch, et al., Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. *Am. J. Hum. Genet.* 87 (2010) 341–353.

- [12] D.Y. Wang, R.L. Green, R.E. Lagacé, 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.* 24 (2012) e7-e8.
- [13] E.Y. Lee, H.Y. Lee, K.S. Shin, Off-ladder alleles due to a single nucleotide polymorphism in the flanking region at DYS481 detected by the PowerPlex® Y23 System, *Forensic Sci. Int. Genet.* 24 (2016) e7-e8.
- [14] D. Ballard, A. Viggars, J. Stickley, L. Devesse, D. Syndercombe Court, C. Phillips, A SNP-based mobility shift in the Penta E short tandem repeat, submitted to *Forensic Sci. Int. Genet.*
- [15] M. Klintschar, Z. Kozma, N. al Hammadi, F.A. Fatah, C. Nöhammer, A study on the short tandem repeat systems HumCD4, HumTH01 and HumFIBRA in population samples from Yemen and Egypt, *Int. J. Legal Med.* 111 (1998) 107-109.
- [16] B. Brinkmann, A. Sajantila, H.W. Goedde, H. Matsumoto, K. Nishi, P. Wiegand, Population genetic comparisons among eight populations using allele frequency and sequence data from three microsatellite loci, *Eur. J. Hum. Genet.* 4 (1996) 175-182.
- [17] K.B. Gettings, D. Ballard, L. Devesse, J.L. King, W. Parson, C. Phillips, P. Vallone, STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci, *Forensic Sci. Int. Genet.* 31 (2017) 111-117.

Figure legends

Fig. 1. Arrangement of inverted and replicated Y-STR sequences that are interspersed with single sequence loci in two 4-megabase (Mb) segments of the Y-chromosome. The repeat motif set by the reference sequence direction of the most 5' fragment is indicated for each multiple allele Y-STR.

Fig. 2. Three alternative annotations of the TH01 9.3 repeat allele placed in relation to the reference sequence. Note that dbSNP describes the rs763206927 insertion Indel as [-/AATGAATGATG] without reference to the shared 5' G nucleotide at 11:2171103, in contrast to the 1000 Genomes annotation system for Indel variants.

Table 1 Forensic STRs compiled in the revised STR Sequence Guide. Black and gray text distinguishes previously and newly listed X-STRs/rapidly mutating (RM) Y-STRs respectively. Gray highlights in the autosomal loci (A-STRs) show NIST Mini-STRs.

Previously listed common use autosomal STRs		Newly listed additional autosomal STRs	
DIS1677	D9S1122	D1GATA113	D10S2325
DIS1656	D9S2157	DIS1627	D11S2368
TPDX	D10S1248	D2S1360	D11S4463
D2S441	D12ATA63	D2S1772	D13S325
D2S1776	TH01	D3S3045	D14S608
D2S1338	VWA	D3S1744	D15S659
D3S1358	D12S391	D3S3053	D17S974
D3S4529	D13S317	D4S2366	D17S1290
D4S2408	D14S1434	D4S2364	D18S853
FGA	Penta E	D6S477	D18S535
D5S818	D16S539	D6S1017	D18S1364
D5S2500*	D17S1301	D7S3048	D19S253
D5S2800*	D18S51	D7S1517	D20S470
CSF1PO	D19S433	D8S1115	D20S1082
SE33	D20S482	D8S1132	D21S1270
D6S1043	D21S11	D9S925	D21S2055
D6S474	Penta D	D10S1435	D22GATA198
D7S820	D22S1045		
D8S1179			

* D5S2500 and D5S2800 now added to the common use A-STRs

X-STRs	RM Y-STRs	Previously listed Y-STRs (DYS458 newly listed)	
DXS10074	DYF387S1	DYS19	DYS456
DXS10103	DYS449	DYS385	DYS460 (DYS461)'
DXS10135	DYS518	DYS389-I /-II	DYS481
DXS7132	DYS570	DYS390	DYS505
DXS7423	DYS576	DYS391	DYS522
DXS8378	DYS612	DYS392	DYS533
HPRTB	DYS627	DYS393	DYS549
DXS10079	DYF399S1	DYS437	DYS635
DXS10101	DYF403S1	DYS438	DYS643
DXS10134	DYF404S1	DYS439	Y-GATA-H4
DXS10146	DYS526b [#]	DYS448	DYS458
DXS10148	DYS547		
	DYS626		

[#] DYS526a=DYS505

' DYS461=Incidental STR close to DYS460

Table 2 Rules for the regularized description of repeat region sequence structures, summarized with examples of the formatting conventions recommended for reporting forensic loci listed in the STR Sequence Guide. Rule 1 items refer to the description of the *counted* elements or motifs in an STR's repeat region. Rule 2 items refer to *uncounted* nucleotides within a repeat region. It is important to note that motifs can have brackets in the Guide's STR repeat region structure summaries, but may not be present in any one individual sequence, or can be observed as a single copy of the motif, so are not bracketed in the STR description.

STR	Length designation	Examples of properly formatted repeat region sequence descriptions
1.1 Brackets and numerical designations are used to shorten the description of repetitive elements/motifs ('n' in the Guide)		
TPOX	6	[AATG]6
SE33	9	[AAAGA]9
1.2 A single space is used between multiple motifs		
D2S1338	18	[GGAA]11 [GGCA]7
D8S1179	15	[TCTA]2 [TCTG]2 [TCTA]11
1.3 Non-repetitive motifs are not bracketed or numbered (but may rarely be present in multiple copies)		
D3S1358	15	TCTA TCTG [TCTA]13
D3S1359	15	TCTA [TCTG]2 [TCTA]12
1.4 Nucleotide tracts creating intermediate length alleles (X.1, X.2, etc.) are formatted as separate elements		
TH01	9.3	[AATG]6 ATG [AATG]3
D1S1656	17.3	CCTA [TCTA]12 TCA TCTG [TCTA]3
FGA	17.2	[GGAA]2 GGAG [AAAG]10 AA AAAA [GAAA]3
D12S391	17.1	AGAT T [AGAT]9 [AGAC]6 AGAT
2. Nucleotide tracts within the repeat region which are not counted are given in lower case and lack spaces		
D19S433	14	[CCTT]12 ccta CCTT cttt CCTT
D19S434	14	[CCTT]9 TCTT [CCTT]2 ccta CCTT cttt CCTT
DXS101048	22	[GGAA]4 [AAGA]12 [AAAG]4 aaggaaag [AAGG]2
DYS449	33	[TTCT]15 ctctctctctctcttcttcc [TTCT]3 ttcctcttcc [TTCT]15
Note 1: In the STR Sequence Guide, uncounted nucleotide tracts longer than 10 nt are given as numbers for brevity		
DYS449	33	[TTCT]15 N22 [TTCT]3 N12 [TTCT]15
Note 2: Motifs that have been observed to repeat are bracketed in the STR Sequence Guide, but individual sequences may not contain all the motifs of a repeat region, and/or these may not be repetitive (so should not be bracketed).		
D8S1179	Summary structure	[TCTA]n [TCTG]n [TCTA]n
D8S1179	11	TCTA TCTG [TCTA]9
Note 3: No font colours, bold type, italics, sub- or super-scripted characters are used in the repeat region descriptions		

Table 3 Eleven STRs where the repeat region sequence structure summaries given in the STR Sequence Guide do not describe the human reference sequence patterns shown.

STR	STR Sequence Guide repeat region sequence structure summary	Reference Sequence repeat region sequence structure summary	Notes
D1S1656	CCTA [TCTA] _n TCA [TCTA] _n	CCTA [TCTA] _n	TCA motif creates X.3 alleles
D2S441	[TCTA] _n TCA [TCTA] _n	[TCTA] _n	TCA motif creates X.3 alleles
SE33	[CTTT] _n TT CT [CTTT] _n	[CTTT] _n TT [CTTT] _n	CT motif in several SE33 alleles
D6S1043	[ATCT] _n [ATGT] _n [ATCT] _n ATGT [ATCT] _n	[ATCT] _n	[ATGT] motifs common in longer alleles
D6S474	[AGAT] _n [GATA] _n [GGTA] _n [GACA] _n	[AGAT] _n [GATA] _n	
D9S1122	TAGA [TCTG] _n [TAGA] _n	[TAGA] _n	
TH01	[AATG] _n ATG [AATG] _n	[AATG] _n	ATG motif creates X.3 alleles
D12S391	[AGAT] _n GA T [AGAT] _n [AGAC] _n AGAT	[AGAT] _n [AGAC] _n AGAT	GAT / T motifs create X.3 / X.1 alleles
D18S51	[AGAA] _n AG	[AGAA] _n	AG motif creates X.2 alleles
D21S11	[TCTA] _n [TCTG] _n [TCTA] _n ta [TCTA] _n tca [TCTA] _n tccata [TCTA] _n TA [TCTA] _n	[TCTA] _n [TCTG] _n [TCTA] _n ta [TCTA] _n tca [TCTA] _n tccata [TCTA] _n	Final TA [TCTA] _n motifs in X.2 alleles
DXS10074	[AAGA] _n [AAGG] _n [AAGA] _n	[AAGA] _n	