Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements

Walther Parson[a,b,*], David Ballard[c], Bruce Budowle[d,e], John M. Butler[f], Katherine B. Gettings[f], Peter Gill[g,h], Leonor Gusmão[i,j,k], Douglas R. Hares[l], Jodi A. Irwin[l], Jonathan L. King[d], Peter de Knijff[m], Niels Morling[n], Mechthild Prinz[o], Peter M. Schneider[p], Christophe Van Neste[q], Sascha Willuweit[r], Christopher Phillips[s]

[a] *Institute of Legal Medicine, Innsbruck Medical University, Innsbruck, Austria*
[b] *Forensic Science Program, The Pennsylvania State University, University Park, PA, USA*
[c] *Faculty of Life Sciences, King's College, London, UK*
[d] *Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, Fort Worth, TX, USA*
[e] *Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia*
[f] *National Institute of Standards and Technology, Gaithersburg, MD, USA*
[g] *Norwegian Institute of Public Health, Department of Forensic Biology, Oslo, Norway*
[h] *Department of Forensic Medicine, University of Oslo, Oslo, Norway*
[i] *DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Brazil*
[j] *IPATIMUP, Institute of Molecular Pathology and Immunology of the University of Porto, Portugal*
[k] *Instituto de Investigação e Inovação em Saúde, University of Porto, Portugal*
[l] *FBI Laboratory, Quantico, VA, USA*
[m] *Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands*
[n] *Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark*
[o] *Department of Sciences, John Jay College for Criminal Justice, New York, NY, USA*
[p] *Institute of Legal Medicine, Medical Faculty, University of Cologne, Cologne, Germany*
[q] *Laboratory of Pharmaceutical Biotechnology, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium*
[r] *Institute of Legal Medicine, Humboldt University, Berlin, Germany*
[s] *Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Galicia, Spain*

ARTICLE INFO

Article history:
Received 14 January 2016
Accepted 16 January 2016
Available online 21 January 2016

Keywords:
Massively parallel sequencing
MPS
Next generation sequencing
NGS
Short tandem repeats
STRs
Nomenclature

ABSTRACT

The DNA Commission of the International Society for Forensic Genetics (ISFG) is reviewing factors that need to be considered ahead of the adoption by the forensic community of short tandem repeat (STR) genotyping by massively parallel sequencing (MPS) technologies. MPS produces sequence data that provide a precise description of the repeat allele structure of a STR marker and variants that may reside in the flanking areas of the repeat region. When a STR contains a complex arrangement of repeat motifs, the level of genetic polymorphism revealed by the sequence data can increase substantially. As repeat structures can be complex and include substitutions, insertions, deletions, variable tandem repeat arrangements of multiple nucleotide motifs, and flanking region SNPs, established capillary electrophoresis (CE) allele descriptions must be supplemented by a new system of STR allele nomenclature, which retains backward compatibility with the CE data that currently populate national DNA databases and that will continue to be produced for the coming years. Thus, there is a pressing need to produce a standardized framework for describing complex sequences that enable comparison with currently used repeat allele nomenclature derived from conventional CE systems. It is important to discern three levels of information in hierarchical order (i) the sequence, (ii) the alignment, and (iii) the nomenclature of STR sequence data. We propose a sequence (text) string format the minimal requirement of data storage that laboratories should follow when adopting MPS of STRs. We further discuss the variant annotation and sequence comparison framework necessary to maintain compatibility among established and future data. This system must be easy to use and interpret by the DNA specialist,

[*] Corresponding author at: Medical University of Innsbruck, Muellerstr. 44, Innsbruck 6020, Austria.
E-mail address: walther.parson@i-med.ac.at (W. Parson).

http://dx.doi.org/10.1016/j.fsigen.2016.01.009

based on a universally accessible genome assembly, and in place before the uptake of MPS by the general forensic community starts to generate sequence data on a large scale. While the established nomenclature for CE-based STR analysis will remain unchanged in the future, the nomenclature of sequence-based STR genotypes will need to follow updated rules and be generated by expert systems that translate MPS sequences to match CE conventions in order to guarantee compatibility between the different generations of STR data.

1. Introduction

Short tandem repeats (STRs) were introduced as polymorphic DNA loci in the forensic field in the early 1990s [1,2] and have become the primary workhorse for individual identification in criminal casework, paternity analyses, and identification of missing persons [3,4]. The STR loci used in forensic DNA analysis were selected using stringent criteria (e.g. [5]). Later, core loci were defined with broad overlap among international legislations [6]. Allele categories have been identified by PCR-based amplicon sizing methods and gel or capillary electrophoretic (CE) systems [3] following simple nomenclature convention [7–9]. Size categories were operationally called relative to sequenced alleles that made up the allelic ladders, with integer values indicating the number of complete repeat motifs and additional nucleotides (i.e. incomplete repeats) separated by a decimal point (e.g. TH01 9.3 [7]). This convention was based on the observed variation generated by CE systems; however, it does not account for sequence differences between alleles that may be caused by transversions, transitions, insertions, deletions, and inversions of one or more nucleotides, including repetitive motifs. Nevertheless, this nomenclature is quite robust, having been adopted universally. In addition, the discrimination power of size-based alleles has proved to be sufficiently high to give useful information for forensic genetic purposes, and even more so with the introduction of large multiplexes [10,11].

Massively parallel sequencing (MPS) is adding a new dimension to the field of forensic genetics, providing distinct advantages over CE systems in terms of captured information, multiplex sizes, and analyzing highly degraded samples [12–14]. In recent years, MPS has been applied to the generation of STR sequence data [15–19] with the general outcome that STRs can be successfully typed producing genotypes compatible with those of CE analyses, even from compromised forensic samples [20]. Furthermore, MPS derived STR genotypes provide additional information to that generated by CE separation by capturing the full nucleotide sequence underlying the repeat units and nearby flanking regions. It was demonstrated by earlier studies using mass spectrometric (MS) systems that the discrimination power of STR typing could be increased by differentiating the nucleotide sequences of alleles with identical size [21–23]. With MPS, forensic tests will further discern STR variants that cannot be distinguished by MS, e.g. repeat motifs that are shifted relative to each other in the repeat region [22]. Early assessments of MPS STR typing show it will be highly beneficial to routine casework by increasing the discrimination power, improving resolution of mixtures, and enhancing the identification of stutter peaks and artifacts [12,18].

However, MPS STR analysis poses challenges to the forensic practitioner. The new technology will affect how the data are analyzed and reported, as well as how they should be stored and searched in databases. This is on top of the necessity to store raw MPS data at the laboratory level. Sequence-based STR variants are more complex and the previously defined nomenclature guidelines do not accommodate the additional variation. While the field is still learning about the sequence variation observed to date and has begun to develop strategies to harmonize nomenclature [24] some laboratories are starting to develop their own large-scale population studies to provide a basis for the introduction of MPS into forensic practice.

For the above reasons, the executive board of the ISFG decided to introduce a DNA commission to evaluate initial considerations regarding STR nomenclature. The primary goal is to define minimum criteria for data analyses and database storage. Ultimately, this should facilitate compatibility between MPS STR data generated currently and the data that will inevitably follow with wider adoption, while ensuring backward and parallel compatibility to the millions of profiles derived from CE-based STR typing in national DNA databases as well as published population data. At present, it can be expected that both CE- and MPS-based STR typing methods will continue to coexist. Their application to casework will depend on laboratory-specific considerations, such as resources, ease of use, speed of analysis, the value of the increased resolution power, and each technique's relevance to complex and challenging cases.

This paper discusses the scientific issues concerning the use of MPS technology for STR typing in forensics and highlights relevant points that should be considered to maintain compatibility of data between technological generations and within and among countries. The adoption of sequenced STR alleles in practical forensic work requires considerations at three hierarchical levels: the full sequence, i.e. the sequence string (Section 2), alignment of sequences relative to a reference sequence (Section 3), and annotation of alleles (Section 4).

2. MPS STR typing and sequence strings

With the application of MPS, the molecular genetic analysis of forensically relevant STR loci results in full nucleotide sequences that harbor the maximum discrimination power possible with DNA-based analyses. The most comprehensive representation of such data is the entire text string of sequenced nucleotides capturing all the information—the sequence string. This string is often referred to as the 'FASTA format', which derives from a more comprehensive and complex 'FASTQ format' that is produced from the raw data of MPS analysis software. It has already been demonstrated that the sequence string is the most convenient and reliable system for storing mitochondrial DNA sequences in database format, as both storage and search tasks become disentangled from alignment and notation (see [25] for mitochondrial DNA sequence strings held in EMPOP [26]). The established analysis regimes for mitochondrial DNA data demonstrate that sequences are not missed in searches performed with an alignment-free format [25], a feature that is particularly desirable and relevant in the forensic field. However, the format of sequence strings is unwieldy when reporting mitochondrial or STR variation in expert reports and cannot be communicated and compared easily without dedicated software.

Consideration 1. MPS analysis should be performed with software that allows STR sequences to be exported and stored in databases as sequence (text) strings to capture the maximum consensus sequence information.

3. Alignment of STR sequences

The forensic community is currently discussing diverse approaches to designate new MPS-based STR data in a suitably compact format. The proposed systems for defining STR sequence variation vary with respect to their complexity and information content. They share the common requirement that they must all be compatible with the existing CE-based STR data (backward compatibility) that populate current forensic databases worldwide. These approaches involve comparison to a reference sequence, a feature that is common practice in the field of mitochondrial DNA sequencing.

3.1. Reference sequences

3.1.1. Lessons learned from mitochondrial DNA

In a discussion about the use of reference sequences to report STR variability, the experience gained with other markers historically reported with respect to a reference sequence is worth revisiting. In the 1990s, the forensic community successfully adopted the concept of using a reference sequence to communicate and report mitochondrial DNA haplotypes [27,28]. The decision to use the first human mitochondrial sequence produced in 1981 [29] as the reference was practically based and was compatible with other fields of research. Every newly generated (partial) mitochondrial DNA sequence was reported relative to this first mitochondrial sequence, known as the Cambridge Reference Sequence (CRS). Eighteen years later, the same source DNA was re-sequenced with improved sequencing technology and alignment software, which resulted in the publication of the revised Cambridge Reference Sequence (rCRS, [30]). The rCRS contains corrections at eleven positions, ten of which were base substitutions at positions 3423T, 4985A, 9559C, 11335C, 13702C, 14199T, 14272C, 14365C, 14368C, and 14766C relative to the CRS. One additional difference was observed at positions 3106 and 3107, where two Cs were recorded in the CRS but only one C was determined in the rCRS. Practically, this means that the rCRS is shorter than the CRS by one nucleotide (16,568 vs. 16,569 total nucleotides). Instead of adjusting all positions downstream of 3107 (or 3106) in their numbering, this position is indicated in the rCRS as a gap [30]. This pragmatic decision allows the numbering system employed for the CRS and by the body of earlier established data to continue to be used unadjusted with the rCRS and subsequent studies.

More recently, the switch to a new mitochondrial DNA reference sequence was proposed. In contrast to the phylogenetically modern rCRS, the proposed sequence represents the deepest root in the known human mtDNA phylogeny (Reconstructed Sapiens Reference Sequence; RSRS [31]). Despite some appealing features of the RSRS, especially with respect to the interpretation of ancient and derived mutations, the forensic community has not adopted it for a number of reasons [32]. Most importantly, lack of adoption eliminates the risk of introducing error as a consequence of the translation between different versions of the mitochondrial reference sequence, especially when comparisons are performed manually. However, the decision was also based on the potential lack of stability of the RSRS that could produce unforeseen consequences for the forensic field [33].

The lessons learned in the field of mitochondrial DNA demonstrate that an established nomenclature system can remain stable and be employed by the forensic community even though (length) changes in the reference sequence were detected (in the shift from CRS to rCRS). As more laboratories begin to use MPS, numerous new STR variants will be discovered. Therefore, it is important to stress that an adapted STR allele nomenclature framework needs to be both flexible and stable in the forensic field. This functionality is easiest to achieve if the nomenclature is ‘natural’, i.e. is derived from the sequence of the allele.

3.1.2. Choice of a reference framework to define STR sequence variation

For any future STR nomenclature scheme, it is necessary to define which of the two DNA strands is reported and to harmonize this criterion so that a universal approach is applied to sequence alignment and comparisons. In contrast to earlier STR nomenclature guidelines that gave general preference to reporting of the coding region strand [7], we propose standardized use of one strand direction. This approach can be framed in a straightforward way by reference to the current standardized genome assembly (the term ‘build’ also is used for a full genome sequence construction, but builds can be short-lived and create multiple numbers within one assembly). A genome assembly assigns each nucleotide a unique chromosome coordinate that positions it precisely in the sequence and follows the system universally applied to locating genomic features such as Single Nucleotide Polymorphisms (SNPs) and Insertions/Deletions (InDels). Genomic coordinates are coded by integers denoting chromosome:position and in the human genome run from the start of the chromosome 1 p-arm to the end of the chromosome 22 q-arm (i.e. 1:1 to 1:248956422 through to 22:1 to 22:50818468 in the autosomal sequences of the most recent genome assembly GRCh38) with equivalent values for the X and Y chromosomes. These genomic coordinates dictate that the strand direction be reported for the human genome as 5′ to 3′—often referred to as “forward” or “positive”. Although strand selection is sometimes arbitrary for other species (i.e. the coordinates can start at the q-arm and go towards the p-arm), in human genome mapping there is a single universal sequence direction dictated by chromosome arm length.

Use of an agreed standard human reference sequence (the reference assembly) for the nuclear portion of the genome provides the key framework from which to generate nucleotide difference-coded genotypes and to designate variants in the sequence string. At the time of writing, the current published genome assembly will be the best framework, as it represents the most accurate sequence curation, i.e. taking into account the precise mapping of complex sequence segments such as duplications and inversions. During the last three to four years, the human genetics community has worked with two human genome assemblies termed GRCh37 and GRCh38. Both GRCh37 and GRCh38 are referenced in the three main human genome databases (NCBI Genome Browser: http://www.ncbi.nlm.nih.gov; UCSC Genome Browser: http://genome.ucsc.edu; and 1000 Genomes Browser: http://browser.1000genomes.org/Homo_sapiens/Info/Index) with data consisting of both sets of coordinates. Although the 1000 Genomes data are still aligned to the GRCh37 assembly [34], at the time of writing, all sequence data from this project are undergoing the transition to map the full human sequence and its variant positions onto the GRCh38 assembly. Therefore, the GRCh38 genome assembly currently is recommended to be the reference sequence adopted by the forensic community and the nucleotide coordinates of this assembly used to map each sequence feature when describing STR variants, whether they are differences in sequence motif, SNPs, or InDels.

Of relevance here is the fact that each MPS platform has analysis software that generates sequence alignments of forensic loci from a standardized assembly. Therefore, agreement between the forensic community and MPS system suppliers about the appropriate assembly used for sequence alignments and annotation becomes a key objective for the DNA Commission on forensic STR sequence nomenclature.

Since the translation of one set of integer values to another is relatively straightforward, it is feasible to have in place an agreed genome assembly for all forensic markers, and retain references to the coordinates of previous assemblies. This compatibility need is important as the entire catalog of SNPs, InDels and microsatellite variants currently accessible from the 1000 Genomes variant database is positioned according to GRCh37 genomic coordinates. When the current GRCh38 assembly is eventually replaced with a new one, the (potentially) necessary transition in coordinate data can be organized within the forensic community while retaining the previous GRCh37 and GRCh38 nucleotide position data. Although genotypes based on previous assemblies could, in principle, be re-coded, the reference assembly difference between any two genotypes could instead be handled bioinformatically when necessary—e.g. at the time of a comparison between two samples. Human genome assembly changes became less frequent in recent years: GRCh38 (hg38) was introduced in December 2013; GRCh37 (hg19) February 2009; NCBI36 (hg18) March 2006; NCBI35 (hg17) May 2004; NCBI34 (hg16) February 2003. Nevertheless, the data processing infrastructure organized for forensic analysis should be prepared to accommodate inevitable changes. Future developments in genome assemblies will be monitored by the Commission and the decision whether or not to adapt the reference sequence to a new assembly will be subject to later discussion.

Consideration 2. The forward strand direction assigned in the human genome has been constant for all assemblies published since the first draft in 2001 and can be used to align STR sequences.

Consideration 3. The choice of reference sequence is crucial for standardizing STR nomenclature systems. At the time of writing, GRCh38 is the most up-to-date sequence assembly and is recommended as the framework with which to define repeat region structure for sequence alignment and for the mapping of sequence features such as SNPs. Software will be required to handle comparisons between multiple reference sequences, particularly in the short term, where sequence variants listed by 1000 Genomes currently retain GRCh37 coordinates. Continued discussions are necessary to decide whether or not to adapt to novel genome assemblies

3.2. Findings from early research on alignment

Having one agreed-upon and up-to-date genome assembly with a unified strand direction presents a logical format as the coordinate integers are ascending values that can be tracked by all forensic scientists using online access to public domain genomic databases. However, this approach is not without complications, as demonstrated by the following examples indicating that more research is required.

Out of 58 STR loci for which MPS designs have become available at the time of this writing (listed in Tables 2–4 of [35]), 23 have been designated historically on the reverse strand. In 17 of these loci, the change to the forward strand for repeat region designation results in a potential shift of the reading frame (Table 1). This shift of reading frame would be consistent with the earlier ISFG

Table 1
Twenty-three STR loci previously aligned relative to the reverse strand (past repeat region sequence) with coordinates and sequences from the current human genome reference GRCh38 [34]. Bolded nucleotides are not counted for the repeat number designation. Seventeen loci for which a potential frameshift exists when converting to forward strand are denoted with "*". The repeat region sequence based on the reference sequence direction (future repeat region sequence) maintains the same location on the reference assembly and is recommended to facilitate comparison to existing sequence data and to length-based STR types. DYS385a/b and DYF387S1a/b: when reporting the forward strand, one allele will contain the reverse complement motif of the other allele, reflecting the occurrence of inversions in each STR.

STR	Chr.	Human reference genome assembly GRCh38					Potential frameshift exists
		Location of repeat region start	Location of repeat region stop	Repeat no.	Past repeat region sequence summary	Future repeat region sequence summary	
D1S1656	1	230769616	230769683	17	[TAGA]16 [TAGG] **[TG]5**	**[CA]5** [CCTA] [TCTA]16	*
D2S1338	2	218014859	218014950	23	[TGCC]7 [TTCC]13 [GTCC] [TTCC]2	[GGAA]2 [GGAC] [GGAA]13 [GGCA]7	
FGA	4	154587736	154587823	22	[TTTC]3 [TTTT] [TTCT] [CTTT]14 [CTCC] [TTCC]2	[GGAA]2 [GGAG] [AAAG]14[AGAA] [AAAA] [GAAA]3	*
D5S818	5	123775556	123775599	11	[AGAT]11	[ATCT]11	*
CSF1PO	5	150076324	150076375	13	[AGAT]13	[ATCT]13	*
D6S1043	6	91740225	91740272	12	[AGAT]12	[ATCT]12	*
D7S820	7	84160226	84160277	13	[GATA]13	[TATC]13	
VWA	12	5983977	5984044	17	[TCTA] [TCTG]5 [TCTA]11 **TCCA TCTA**	**TAGA TGGA** [TAGA]11 [CAGA]5 [TAGA]	*
Penta E	15	96831015	96831039	5	[AAAGA]5	[TCTTT]5	*
D19S433	19	29926235	29926298	16	[AAGG] **AAAG** [AAGG] **TAGG** [AAGG]12	[CCTT]12 **CCTA** [CCTT] **CTTT** [CCTT]	*
DYS19	Y	9684380	9684443	15	[TAGA]3 **TAGG** [TAGA]12	[TCTA]12 **CCTA** [TCTA]3	*
DYS635	Y	12258860	12258951	23	[TCTA]4 [TGTA]2 [TCTA]2 [TGTA]2 [TCTA]2 [TGTA]2 [TCTA]9	[TAGA]9 [TACA]2 [TAGA]2 [TACA]2 [TAGA]2 [TACA]2 [TAGA]4	*
DYS389I	Y	12500448	12500495	12	[TCTG]3 [TCTA]9	[TAGA]9 [CAGA]3	*
DYS389II	Y	12500448	12500611	29	[TCTG]5 [TCTA]12 **48 nt.** [TCTG]3 [TCTA] 9	[TAGA]9 [CAGA]3 **48 nt.** [TAGA]12 [CAGA]5	*
DYS390	Y	15163067	15163162	24	**[TCTA]2** [TCTG]8 [TCTA]11 TCTG [TCTA] 4	[TAGA]4CAGA [TAGA]11 [CAGA]8 **[TAGA]2**	*
Y-GATA-H4	Y	16631673	16631720	12	[TAGA]12	[TCTA]12	
DYS385ab	Y	18639713	18639756	11	[GAAA]11	[TTTC]11	*
		18680632	18680687	14	[GAAA]14	[GAAA]14	
DYS460	Y	18888810	18888849	10	[GATA]10	[TATC]10	*
DYS392	Y	20471987	20472025	13	[TAT]13	[ATA]13	*
DYF387S1ab	Y	23785361	23785500	35	[AAAG]3 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]9 [AAAG]13	[AAAG]3 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]9 [AAAG]13	
		25884581	25884724	36	[AAAG]3 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]10 [AAAG]13	[CTTT]13 [CTTC]10 [CTTT]2CTTC [CTTT] 2 [CTTC]4CTAC [CTTT]3	*
DXS8378	X	9402262	9402301	10	[CTAT]10	[ATAG]10	
HPRTB	X	134481506	134481561	13	[TAGA]14	[TCTA]14	
DXS7423	X	150542522	150542589	15	[TCCA]3 **TCTGTCCT** [TCCA]12	[TGGA]12 **AGGACAGA** [TGGA]3	

recommendations [7] that the repeat region begins with the first possible repeat motif. This change can cause a shift in the position of features within the motif and/or an increase in the number of apparent repeats. For example, the D19S433 locus historically has been reported on the reverse strand as an AAGG repeat interspersed with one AAAG and one TAGG that are uncounted (see first example sequence below, underlined bases are counted while bolded bases are not counted). The reverse complement consists of a CCTT repeat interspersed with one CCTA and one CTTT that are uncounted (second example sequence below). However, under earlier recommendations, the first possible repeat motif of TCCT would be reported (one nucleotide shift to the left, third example sequence below), and the interspersed feature becomes ACCT TCTT. This change could complicate comparisons to existing sequence data.

1. TGTTG AAGG **AAAG** AAGG **TAGG** AAGG AAGG AAGG AAGG AAGG AAGG AGAGA

2. TCTCT CCTT CCTT CCTT CCTT CCTT CCTT **CCTA** CCTT **CTTT** CCTT CAACA

3. TCTC TCCT TCCT TCCT TCCT TCCT TCCT TCCT **ACCT TCTT** TCCT TCAACA

At the DYS389I/II loci, the potential exists for a two nucleotide shift, which would result in the appearance of one extra repeat in the larger allele. The first two bracketed sequences below show the change from reverse to forward strand maintaining identical repeat region positions on GRCh38, while the third bracketed sequence shows the change of strand with a shifted motif, yielding an extra repeat at the 3′ end. If sequence based analysis counted this repeat while traditional CE assays did not, the results would appear discordant by one repeat unit.

Previously reported reverse strand:	$[TCTG]_5$ $[TCTA]_{12}$ 48 nt. $[TCTG]_3$ $[TCTA]_9$
Forward strand, no frame shift:	$[TAGA]_9$ $[CAGA]_3$ 48 nt. $[TAGA]_{12}$ $[CAGA]_5$
Forward strand, frame shift:	$[GATA]_9$ $[GACA]_3$ 48 nt. $[GATA]_{12}$ $[GACA]_6$

Lastly, the DYS385 a/b marker has two repeat regions located in the most recent human reference sequence at Y:18639713-18639756 and Y:18680632-18680687 (Table 1). On the forward strand the first fragment has TTTC motifs while the second one comprises an inversion of the same sequence presenting GAAA motifs. In this case, using the forward strand, it is not possible to summarize DYS385 a/b repeats by a uniform motif description as was reported in the past. In addition, it is expected that some individuals will exhibit a larger first fragment and a smaller second fragment, resulting in a genotype of, e.g. 14, 11.

These examples aptly demonstrate potential complications arising from conversion of STR loci to the forward strand. It is clearly indicated that this conversion needs to be performed by designed software once MPS has reached routine application, and not manually, as the risk of introducing error would be too high. Also, it is imperative that repeat region start and end locations be strictly defined for all STR loci employed in MPS. This work is underway in various laboratories and updates will be made available to the forensic community.

As a simple guide to the human genome reference sequence, Supplementary file S1 outlines the reference strings of the repeat regions plus 50 nucleotides of each flanking sequence of STRs that will form the next generation of MPS multiplexes or have already become established for this type of forensic DNA analysis. Supplementary file S1A details 35 autosomal STRs (12 ESS, 20 CODIS markers) in common use, and Supplementary file S1B details 29 Y-STRs plus 7 X-STRs. The SNPs and InDels currently recorded by 1000 Genomes are identified in the flanking sequences, and the most polymorphic of these flanking region variants (>10% minor allele frequencies) are summarized with pie charts.

Although the human genome assembly coordinates of GRCh37 and GRCh38 can be translated in a straightforward way, three common STRs have nucleotide differences in the repeat region sequences reported by each assembly. These are for the loci DYS437 (GRCh38 one less repeat), DYS438 (two more repeats), and DYS439 (one less repeat), each reference sequence is summarized in Supplementary file S2. These nucleotide differences illustrate the challenges that must be addressed when future human genome assemblies are published and used for STR sequence alignments of MPS data.

Lastly, during detailed examination of the human genome assembly sequences at each STR, it emerged that the forensic marker named D5S2500 is represented by two different microsatellites that each form separate components in commercial CE multiplexes (e.g. Qiagen's HD-plex (Hilden, Germany) and AGCU ScienTech's 21-plex (Wuxi, China)). Investigations of both sites reveal that D5S2500 in Qiagen's HD-plex is the correctly assigned STR name. The microsatellite targeted in AGCU ScienTech's 21-plex is not a named microsatellite at the time of writing, being positioned 1688 nucleotides further upstream. The microsatellite in the AGCU kit was originally developed as a miniSTR, incorrectly named D5S2500 and reported by Hill et al. [36]. To avoid confusion while including sequence details of each of these important forensic STRs, the locus used in Qiagen's HD-plex is labeled with its NCBI accession number D5S2500.G08468, while the locus used in AGCU ScienTech's 21-plex is coded as D5S2500.AC008791(Supplementary file S1C). Details of both D5S2500 markers are summarized in the same way as the other STRs but placed in a separate Supplementary File S1C. More thorough characterization of these two microsatellites is the subject of a separate paper in preparation.

Consideration 4. Further work is needed to translate the nomenclature of STR loci thus far coded relative to the reverse strand and repeat region start and end points. There is a need to strictly define these and other anchor points to specify the repeat regions.

4. Annotation of STR alleles—nomenclature systems

Established conventions for the nomenclature of forensic CE-based STR genotypes will remain unchanged. Updated and extended nomenclature systems that can be performed by expert systems will be required for STR sequences that can be performed by specifically designed software. It is crucial that this software allow for translation of MPS-derived genotypes to the CE-based nomenclature convention to stay compatible with established STR databases and future CE-based STR results. We note that it is too early to set strict guidelines for new nomenclature formats for MPS. The following exemplar systems are presented here to explore different ways to call MPS-based STR results and can serve as the basis for further discussion and development.

4.1. Comprehensive (high level) STR nomenclature systems

Comprehensive STR nomenclature systems capture the majority, preferably all, of the information present in the STR sequence string and can be delineated from the recommendations of the human genome variation society (http://www.hgvs.org). A comprehensive format includes the STR locus information, the size-based allele category, which provides backward compatibility to existing STR databases, and an unambiguous description of the sequence variation of each allele. An example of a minimum nomenclature format that could be used in the case of the D13S317 locus is shown in Textbox 1. When a particular genome assembly is used as the reference for the sequence alignment, the assembly version should be stated. Information must be also compiled on the chromosome number and coordinates relating to the whole STR amplicon to compare alleles generated with different primer pairs and the repeat region to differentiate identical repeat and flanking sequence motifs, from which the allele designation was made. Finally, the repeat motif should be fully described with the relevant nucleotide 'blocks' and repeat numbers in brackets as well as SNPs and/or InDels described by genome coordinates or rs-numbers. Common SNP and InDel variants, including those in repeat regions, typically have been identified already and have rs-numbers. Novel variants not yet catalogued tend to keep their chromosome coordinates as identifiers until an rs-number is assigned. This process of rs-number assignment is becoming an increasingly difficult process to complete as a large proportion of SNP variation is unique to an individual [34].

Comprehensive STR nomenclature systems are informative and can be translated to lower level nomenclature systems at any time to maintain backward compatibility with existing databases. However, they cannot easily be applied for communication among forensic analysts and stakeholders as is currently practiced with simple repeat number notation. To facilitate communication and maintain backwards compatibility, any nomenclature system will need to take into account the number of repeats presented in the human reference sequence.

4.2. Simple (low level) STR nomenclature systems

Low-level STR nomenclature systems are based on the translation of sequence strings or comprehensive STR nomenclature systems and typically represent easy-to-read unique identifiers. They typically consist of the STR locus name and the operationally-defined repeat-based allele designation derived from CE. This approach makes the data directly compatible with those of existing STR databases. In order to capture the additional sequence information, accompanying letters have been proposed or numbers and letters in alternating order could be applied, a system that is currently used to display the phylogenetic relationship between linearly inherited markers [37,38]. Simple STR nomenclature systems are easy to communicate and therefore

Textbox 1. An example of a possible sequence nomenclature regime using the example STR D13S317 allele 12 ([CE12]) compared to the reference allele 11 (Ref [11]). Sequence descriptions include the following bolded components: (1) the reference genome assembly sequence (includes allele 11); (2) locus name and CE allele number; (3) chromosome number and reference genome assembly used; (4) repeat region coordinates of the reference allele (start-end nucleotide positions, but eventually to also include the reported region start-end coordinates); (5) description of the repeat motifs; and (6) location of flanking region variants. See D13S317 in Supplementary file S1A for more details of the reference sequence.

```
D13S317 Ref (11)  TCTAACGCCT ATCTGTATTT ACAAATACAT TATC TATC TATC TATC
D13S317 [CE12]    ......A... .......... .......... .... .... .... ....

D13S317 Ref (11)  TATC TATC TATC TATC TATC TATC TATC ++++ AATCAATCAT
D13S317 [CE12]    .... .... .... .... .... .... .... TATC T.........

D13S317 Ref (11)  CTATCTATCT TTCTGTCTGT
D13S317 [CE12]    .......... ..........
```

G Known polymorphic sites

++++ Additional nucleotides compared to reference sequence

1. Bold segment = the reference genome assembly sequence description
D13S317 Ref (11) -Chr13-GRCh38 82148025-82148068 $[TATC]_{11}$
D13S317[CE12]-Chr13-GRCh38 82148025-82148068 $[TATC]_{12}$ 82148001-A; 82148069-T

2. Locus name and capillary electrophoresis allele name
D13S317[CE12]-Chr13-GRCh38 82148025-82148068 $[TATC]_{12}$ 82148001-A; 82148069-T

3. Chromosome and human genome assembly version
D13S317[CE12]-**Chr13-GRCh38** 82148025-82148068 $[TATC]_{12}$ 82148001-A; 82148069-T

4. STR repeat region co-ordinates (start-end) for reference allele
D13S317[CE12]-Chr13-GRCh38 **82148025-82148068** $[TATC]_{12}$ 82148001-A; 82148069-T

5. Description of STR motifs
D13S317[CE12]-Chr13-GRCh38 82148025-82148068 **$[TATC]_{12}$** 82148001-A; 82148069-T

6. Location of flanking region variants
D13S317[CE12]-Chr13-GRCh38 82148025-82148068 $[TATC]_{12}$ **82148001-A; 82148069-T**

preferred for routine exchange of STR data between analysts and stakeholders and may be easier to apply to existing software packages that perform various population genetic and statistical analyses. However, the translation process will have to be managed by a centralized nomenclature commission to avoid ambiguous or imprecise allele names being adopted, or assigning different names to identical alleles. It has been suggested that an online system could be used that is curated by a nomenclature commission, which would be responsible for new allele designations upon validation of the observed sequence variation. Criteria for the validation of sequence variation and its comparison with existing variants need to be defined in more detail. Numerous new variants will be discovered; hence, it is necessary to automate the process as much as possible. If a 'natural' nomenclature is adopted, then cataloguing of variants can be accommodated by an open source algorithm, which should be a key aim of the community.

Fig. 1 illustrates examples of potential difficulties that can arise from the more detailed characterization of STR sequences that MPS provides. There can be unforeseen challenges when aligning the sequence generated by MPS to the established repeat motif description of any STR. Each of the three STRs is described by its respective human reference sequences, which include the repeat regions plus the short segments of the flanking regions.

The D18S51 reference sequence comprises 18 AGAA repeat motifs (ten nucleotides of flanking region also displayed). Two repeat region InDels create intermediate repeats: x.3 (rs572637907); x.2 (rs575219471); or x.1 (presence of both deletions or another unmapped deletion). Furthermore, the flanking A/G SNP rs535823682 potentially complicates the alignment of the repeat sequence.

The D13S317 reference sequence comprises 11 TATC repeat motifs (extended flanking regions displayed). The two 3′ flanking region A/T SNPs, rs9546005 and rs202043589, create TATC tetranucleotides matching the repeat motifs, but these are not counted when deriving the total repeat number. The rs561167308 TCTG deletion potentially creates a four-nucleotide fragment size disparity with CE-based allele descriptions depending on the position of the 3′ primer-binding site. The 5′ SNP rs146621667 is the site of the '82148001-A' variant described in Textbox 1.

The D19S433 reference sequence comprises 14CCTT repeat motifs, which contain two 'punctuated' stable repeat motifs, CCTA and CTTT, that should be counted, but in the initial development of forensic CE kits for D19S433 were not. The D19S433 STRbase (http://www.cstl.nist.gov/strbase/) fact sheet therefore provides a cautionary note to highlight that current allelic ladders retain the numbering system first used that did not count the above two non-standard motifs in combination with the CCTT motifs. The 16 nucleotide 5′ flanking sequence also shows permutations on the CCTT motif that have no sequence variants but can present alignment challenges for analysis of MPS sequence data.

The above examples illustrate that when characterization of repeat regions does not follow previously agreed nomenclature rules [7] it potentially creates discrepancies between CE-based repeat counts and MPS sequence analyses made from the same amplified fragments. In this case, a nomenclature commission can preempt potential issues by harmonizing CE numbering systems and repeat region sequence descriptions. However, since STR types based on CE already populate national DNA databases, the existing nomenclature rules must be applied to MPS sequence data to prevent data mismatches, even though they may not follow common logic.

Consideration 5. Although simple STR nomenclature systems may be required at some point in the future to facilitate communication and data exchange, comprehensive STR nomenclature systems are preferred for early adopters of STR MPS analysis in order to ensure compatibility with MPS data generated in the future. Backward compatibility to the

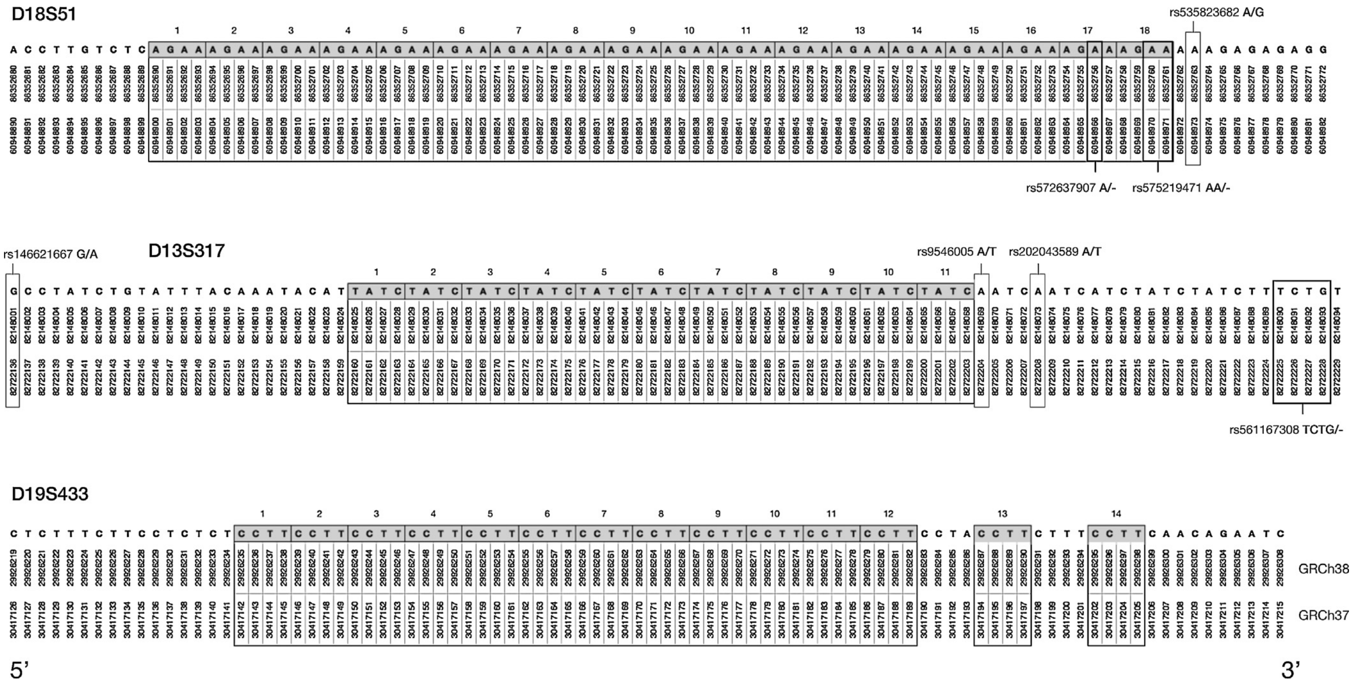

Fig. 1. Three examples of STR repeat regions plus the short segments of their 5′ and 3′ flanking sequences that illustrate potential difficulties with repeat motif description. All sequences are taken from the current human reference genome assembly and coordinates are given for both GRCh37 and GRCh38. Repeat regions are denoted by thin black boxes, InDels by thick black boxes, and SNPs by grey boxes. For a more detailed description of each STR sequence see [17]. D18S51 reference sequence of 18 AGAA repeat motifs and ten nucleotides of flanking region. D13S317 reference sequence of 11 TATC repeat motifs with extended flanking regions. In both STRs InDel polymorphisms and/or SNPs in the 3′ flanking region create intermediate alleles but these sequence changes can mimic repeat motifs not included in the CE-based nomenclature. D19S433 reference sequence of 14CCTT repeat motifs and flanking regions. In this STR not all tandemly-arranged tetra-nucleotide motifs are counted in the description of the repeat region.

repeat-based nomenclature derived from CE needs to be maintained to preserve the universal applicability of established national STR databases

4.3. Flanking regions

The inclusion of flanking region sequence variants (between primer binding sites and the repeat region) in compiled MPS data is important for several reasons. First, it provides additional informative polymorphisms with which to differentiate alleles that have identical repeat region sequences. Second, the mapping of InDel variants informs the assignment of size-based allele designations from CE analyses, where the total fragment size is altered by the presence of the variant. One example is the occurrence of a four-nucleotide deletion (rs561167308) close to the repeat region of the D18S51 locus that changes the repeat length but is not a detected repeat itself [18]. This is also the case with the DXS10148 locus, which has a variable motif of eight bases adjacent to the core tetra-nucleotide repeat region [39]. Third, it is likely that a small but regular proportion of novel rare variants will be discovered in full STR sequence segments that potentially provide additional ways to differentiate STR alleles amongst related individuals, but which have no previously defined frequency data. In these instances, it is important to compare the novel variants with a database of established flanking region variants including sample population sizes to provide allele frequencies. As flanking region variants and repeat region sequence variants are present on one DNA fragment, the database must compile all variation in the sequence string from any one sample. Novel variants can be described by their genome coordinates, while recognized variants that already are catalogued will have rs-numbers. To ensure compatibility between/among different primer sets used for library preparation and sequencing, it is mandatory to provide genome coordinates of the sequence read start and end points similar to current practices with difference-coded variants describing mtDNA haplotypes [28]. This procedure should cover annotation of InDels, as it is possible that some MPS primer sets will be positioned inside those used for CE analysis such that InDel sites may escape detection by sequencing and create discordant fragment sizes. Such checks have been made successfully, e.g. the concordance studies of MiniFiler systems, where modified primer positions did influence the observed repeat numbers [40].

Supplementary file S1 illustrates seven common flanking region SNPs within 50 nucleotides flanking region of the listed autosomal STRs. The SNPs are shown with population frequency data from 1000 Genomes samples and represent the most informative levels of flanking region variation, defined here as having minor allele frequencies of 10% or more in most populations (average heterozygosities of 18% or higher). These SNPs are: rs4847015 in the D1S1656 locus; rs6736691 in the D2S1338 locus; rs25768 in the D5S818 locus; rs16887642 in the D7S820 locus; rs75219269 in the VWA locus; rs9546005 in the D13S317 locus, and rs11642858 in the D16539 locus. However, their detection is dependent on the amplified fragment sizes of each locus (i.e. the position of the primers). For example, certain SNPs within 50 nucleotides of the repeat region will not be genotyped when much shorter STR fragment lengths are generated by MPS primer sets.

Consideration 6. To account for relevant genetic variation outside common repeat regions, STR sequences stored as sequence strings should include flanking sequences as well as the genome coordinates of the sequence read start and end points.

5. Updated allele frequencies

Current allele frequency tables are not sufficient to quantify any new variation gained by sequencing of STRs. Preliminary studies indicate that the number of rare STR alleles will increase substantially with MPS [18,41,42]. Thus, comprehensive MPS databasing will be required to characterize the extent of STR sequence variation for use in STR frequency estimates. Therefore, there is a particular need to promptly harmonize nomenclature frameworks, since a coordinated effort is required to collate the sequence variation found by early adopters, before this process reaches the wider community of forensic laboratories.

From data published so far [18,41,42] and from previous assessments of sequence variation with ICEMS technology [22,23,43] it is certain that many common STRs (e.g. D12S391, D21S11) will require large-scale efforts to compile representative samples of their variation, while other STRs such as FGA appear to have largely unchanged levels of polymorphism. In addition, flanking sequence variation will show a proportion of 'private' variants at <1% frequencies that have not been previously described [34]. Thus, the community must adopt a nomenclature framework that captures variation within the repeats and a framework for flanking SNPs lacking rs-numbers. Prompt standardization of nomenclature will facilitate the development of large-scale sequence databases and expedite the collection of rare variant allele frequencies, much of which may be population-specific.

Consideration 7. Updated allele frequency databases will be necessary to take full advantage of the increased power of discrimination offered by MPS generated STR data. A unified nomenclature system is needed to ensure compatibility of worldwide population databases.

6. Selection of STR loci

While the choice of the first forensic STR loci was previously driven by individual research groups (e.g. [44]) and later commercially produced (e.g. [45]), the addition of new forensically-relevant STR loci was led by world-wide forensic societies and working groups (e.g. [5,6,10]). This emphasis on localized needs was important for laboratories to meet legal requirements defined in their respective countries, with particular regard to database search strategies. It is desirable to continue dialogues between forensic groups and commercial suppliers to ensure provision of appropriate loci, chemistry, and software.

The variation of new STR loci should be tested with studies of populations from the main continental groups with particular emphasis on discrimination power, heterozygosity levels, sequence variation in the flanking regions, and inter- and intra-population variation. Given the complexities of STR sequence alignments and the current limitation of MPS read length, SE33 [46] is unlikely to be part of initial forensic MPS multiplexes. In its place many miniSTRs, newer to mainstream use, could be suitable alternatives and are certain to be incorporated into future MPS marker sets [36]. These STRs will require full characterization, including crucial information about possible linkage to the already well established STR markers [47]. so that frequency data and knowledge of sequence characteristics can be added to the extensive data in place for the commonly used loci.

At present, the key factors that must be considered in the application of sequencing technologies to STRs center on standardized representation of sequence variation. Until an appropriate, agreed upon framework for simplified STR nomenclature is established, STR sequence data should reflect the most detailed and inclusive level of information for any given allele, while still retaining compatibility with current CE-defined variants. The likely near-term development of reference population data should serve to test the utility and robustness of the considerations presented here, and also provides the necessary data framework for refinement and establishment of a practical and durable simplified nomenclature scheme.

At a future point in time when MPS-based databases have grown in size, algorithms could be used to determine frequency databases without the need to annotate alleles. A strength-of-evidence calculation would follow without any reference to nomenclature. However, this approach would require a broad application of MPS-based STR typing by the forensic community.

> **Consideration 8.** Future forensic MPS multiplexes would benefit from retention of past markers for backward compatibility and a marker selection process based on population data, molecular biology, sequencing chemistry, and a continued dialogue between the forensic community and commercial suppliers.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

Concepts and ideas stated in this manuscript were stimulated by a panel discussion on STR nomenclature at the 26th Conference of the International Society for Forensic Genetics, 2 September 2015, Krakow, Poland. The authors are indebted to all those, who contributed to the discussions both during and after the panel's review of STR sequence nomenclature issues. The authors would like to thank Bettina Zimmermann (Innsbruck, Austria) for technical help and Chris Tyler Smith, Sanger Institute (Hinxton, UK) for valuable guidance on current 1000 Genomes data policies. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice or the National Institute of Standards and Technology (US). 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. This is FBI Laboratory publication number 16-11.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2016.01.009.

References

[1] C. Puers, H.A. Hammond, L. Jin, C.T. Caskey, J.W. 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 (1993) 953–958.

[2] P. Gill, C. Kimpton, E. D'Aloja, J.F. Andersen, W. Bar, B. Brinkmann, et al., Report of the European DNA profiling group (EDNAP)—towards standardisation of short tandem repeat (STR) loci, Forensic Sci. Int. 65 (1994) 51–59.

[3] P. Gill, R. Sparkes, C. Kimpton, Development of guidelines to designate alleles using an STR multiplex system, Forensic Sci. Int. 89 (1997) 185–197.

[4] B. Budowle, T.R. Moretti, A.L. Baumstark, D.A. Defenbaugh, K.M. Keys, Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S., Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians, J. Forensic Sci. 44 (1999) 1277–1286.

[5] P. Gill, E. d'Aloja, B. Dupuy, B. Eriksen, M. Jangblad, V. Johnsson, et al., Report of the European DNA profiling group (EDNAP)—an investigation of the hypervariable STR loci ACTBP2, APOAI1 and D11S554 and the compound loci D12S391 and D1S1656, Forensic Sci. Int. 98 (1998) 193–200.

[6] 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 (2012) 819–826.

[7] W. Bär, 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. International Society for Forensic Haemogenetics, Int. J. Legal Med. 110 (1997) 175–176.

[8] J.M. Butler, Genetics and genomics of core short tandem repeat loci used in human identity testing, J. Forensic Sci. 51 (2006) 253–265.

[9] P.M. Schneider, Scientific standards for studies in forensic genetics, Forensic Sci. Int. 165 (2007) 238–243.

[10] D.R. Hares, Selection and implementation of expanded CODIS core loci in the United States, Forensic Sci. Int. Genet. 17 (2015) 33–34.

[11] P.M. Schneider, Expansion of the European Standard Set of DNA database loci—the current situation, Profiles DNA (2009) 6–7.

[12] C. Borsting, N. Morling, Next generation sequencing and its applications in forensic genetics, Forensic Sci. Int. Genet. 18 (2015) 78–89.

[13] W. Parson, G. Huber, L. Moreno, M.B. Madel, M.D. Brandhagen, S. Nagl, et al., Massively parallel sequencing of complete mitochondrial genomes from hair shaft samples, Forensic Sci. Int. Genet. 15 (2015) 8–15.

[14] M. Eduardoff, C. Santos, M. de la Puente, T.E. Gross, M. Fondevila, C. Strobl, et al., Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using the Ion PGM, Forensic Sci. Int. Genet. 17 (2015) 110–121.

[15] S.L. Fordyce, M.C. Avila-Arcos, E. Rockenbauer, C. Borsting, R. Frank-Hansen, F.T. Petersen, et al., High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform, BioTechniques 51 (2011) 127–133.

[16] D.H. Warshauer, D. Lin, K. Hari, R. Jain, C. Davis, B. Larue, et al., STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data, Forensic Sci. Int. Genet. 7 (2013) 409–417.

[17] K.B. Gettings, K.M. Kiesler, S.A. Faith, E. Montano, C.H. Baker, B.A. Young, R.A. Guerrieri, P.M. Vallone, Sequence variation of 22 autosomal STR loci detected by next generation sequencing, Forensic Sci. Int. Genet. 21 (2016) 15–21.

[18] K.B. Gettings, R.A. Aponte, P.M. Vallone, J.M. Butler, STR allele sequence variation: current knowledge and future issues, Forensic Sci. Int. Genet. 18 (2015) 118–130.

[19] C. Van Neste, F. Van Nieuwerburgh, D. Van Hoofstat, D. Deforce, Forensic STR analysis using massive parallel sequencing, Forensic Sci. Int. Genet. 6 (2012) 810–818.

[20] M. Scheible, O. Loreille, R. Just, J. Irwin, Short tandem repeat typing on the 454 platform: strategies and considerations for targeted sequencing of common forensic markers, Forensic Sci. Int. Genet. 12 (2014) 107–119.

[21] J.M. Butler, J. Li, T.A. Shaler, J.A. Monforte, C.H. Becker, Reliable genotyping of short tandem repeat loci without an allelic ladder using time-of-flight mass spectrometry, Int. J. Legal Med. 112 (1999) 45–49.

[22] F. Pitterl, H. Niederstatter, G. Huber, B. Zimmermann, H. Oberacher, W. Parson, The next generation of DNA profiling–STR typing by multiplexed PCR–ion-pair RP LC-ESI time-of-flight MS, Electrophoresis 29 (2008) 4739–4750.

[23] F. Pitterl, K. Schmidt, G. Huber, B. Zimmermann, R. Delport, S. Amory, et al., Increasing the discrimination power of forensic STR testing by employing high-performance mass spectrometry, as illustrated in indigenous South African and Central Asian populations, Int. J. Legal Med. 124 (2010) 551–558.

[24] K. van der Gaag, P. de Knijff, Forensic nomenclature for short tandem repeats updated for sequencing, Forensic Sci. Int. Genet. Suppl. Ser. 5 (2015) e542–e544.

[25] A. Röck, J. Irwin, A. Dur, T. Parsons, W. Parson, SAM: string-based sequence search algorithm for mitochondrial DNA database queries, Forensic Sci. Int. Genet. 5 (2011) 126–132.

[26] W. Parson, A. Dür, EMPOP–a forensic mtDNA database, Forensic Sci. Int. Genet. 1 (2007) 88–92.

[27] A. Carracedo, W. Bär, P. Lincoln, W. Mayr, N. Morling, B. Olaisen, et al., DNA commission of the international society for forensic genetics: guidelines for mitochondrial DNA typing, Forensic Sci. Int. 110 (2000) 79–85.

[28] W. Parson, L. Gusmao, D.R. Hares, J.A. Irwin, W.R. Mayr, N. Morling, et al., DNA Commission of the International Society for forensic genetics: revised and extended guidelines for mitochondrial DNA typing, Forensic Sci. Int. Genet. 13 (2014) 134–142.

[29] S. Anderson, A.T. Bankier, B.G. Barrell, M.H. de Bruijn, A.R. Coulson, J. Drouin, et al., Sequence and organization of the human mitochondrial genome, Nature 290 (1981) 457–465.

[30] R.M. Andrews, I. Kubacka, P.F. Chinnery, R.N. Lightowlers, D.M. Turnbull, N. Howell, Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA, Nat. Genet. 23 (1999) 147.

[31] D.M. Behar, M. van Oven, S. Rosset, M. Metspalu, E.L. Loogvali, N.M. Silva, et al., A Copernican reassessment of the human mitochondrial DNA tree from its root, Am. J. Hum. Genet. 90 (2012) 675–684.

[32] A. Salas, M. Coble, S. Desmyter, T. Grzybowski, L. Gusmao, C. Hohoff, et al., A cautionary note on switching mitochondrial DNA reference sequences in forensic genetics, Forensic Sci. Int. Genet. 6 (2012) e182–e184.

[33] H.J. Bandelt, A. Kloss-Brandstätter, M.B. Richards, Y.G. Yao, I. Logan, The case for the continuing use of the revised Cambridge Reference Sequence (rCRS) and

the standardization of notation in human mitochondrial DNA studies, J. Hum. Genet. 59 (2014) 66–77.

[34] C. Genomes Project, A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, et al., A global reference for human genetic variation, Nature 526 (2015) 68–74.

[35] ForenSeq DNA Signature Prep Reference Guide, Document No. 15049528 v.01, (2015), Illumina.

[36] C.R. Hill, J.M. Butler, P.M. Vallone, A 26plex autosomal STR assay to aid human identity testing, J. Forensic Sci. 54 (2009) 1008–1015.

[37] S.K. Lim, Y. Xue, E.J. Parkin, C. Tyler-Smith, Variation of 52 new Y-STR loci in the Y chromosome consortium worldwide panel of 76 diverse individuals, Int. J. Legal Med. 121 (2007) 124–127.

[38] A. Torroni, A. Achilli, V. Macaulay, M. Richards, H.J. Bandelt, Harvesting the fruit of the human mtDNA tree, Trends Genet. 22 (2006) 339–345.

[39] I. Gomes, A. Brehm, L. Gusmao, P.M. Schneider, New sequence variants detected at DXS10148, DXS10074 and DXS10134 loci, Forensic Sci. Int. Genet. 20 (2016) 112–116.

[40] 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 (2007) 870–873.

[41] E. Rockenbauer, S. Hansen, M. Mikkelsen, C. Borsting, N. Morling, Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing, Forensic Sci. Int. Genet. 8 (2014) 68–72.

[42] C. Gelardi, E. Rockenbauer, S. Dalsgaard, C. Borsting, N. Morling, Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles, Forensic Sci. Int. Genet. 12 (2014) 38–41.

[43] J.V. Planz, K.A. Sannes-Lowery, D.D. Duncan, S. Manalili, B. Budowle, R. Chakraborty, et al., Automated analysis of sequence polymorphism in STR alleles by PCR and direct electrospray ionization mass spectrometry, Forensic Sci. Int. Genet. 6 (2012) 594–606.

[44] J.E. Lygo, P.E. Johnson, D.J. Holdaway, S. Woodroffe, J.P. Whitaker, T.M. Clayton, et al., The validation of short tandem repeat (STR) loci for use in forensic casework, Int. J. Legal Med. 107 (1994) 77–89.

[45] B. Budowle, C.J. Sprecher, Concordance study on population database samples using the PowerPlex 16 kit and AmpFlSTR Profiler Plus kit and AmpFlSTR COfiler kit, J. Forensic Sci. 46 (2001) 637–641.

[46] D. Warne, C. Watkins, P. Bodfish, K. Nyberg, N.K. Spurr, Tetranucleotide repeat polymorphism at the human beta-actin related pseudogene 2 (ACTBP2) detected using the polymerase chain reaction, Nucleic Acids Res. 19 (1991) 6980.

[47] C. Phillips, D. Ballard, P. Gill, D.S. Court, A. Carracedo, M.V. Lareu, The recombination landscape around forensic STRs: accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data, Forensic Sci. Int. Genet. 6 (2012) 354–365.