Forensic Science International: Genetics 22 (2016) 54–63

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
1872-4973/

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.