Development of forensic genomics research toolkits by the use of Massively Parallel Sequencing

Gaag, K.J. van der

Chapter 3

Forensic Nomenclature for Short Tandem Repeats updated for sequencing

Forensic Science International: Genetics Supplement Series 5 (2015) e542–e544
Published online 26 September 2015

Kristiaan J. van der Gaag
Peter de Knijff

Abstract

The introduction of Massive Parallel Sequencing (MPS) techniques enables sequencing of Short Tandem Repeats (STR) as a new tool for forensic research. In addition to variation in fragment-length, MPS also reveals allelic sequence-variation in STR-fragments. This additional variation demands a new way of describing allelic variants. Here we propose a nomenclature of MPS-derived STR alleles for use in forensic research.

Introduction

For over two decades, the analysis of Short Tandem Repeats (STRs) in forensics was routinely performed using Capillary Electrophoresis (CE). With CE, the length of a DNA fragment containing an STR is determined. STR alleles are identified by comparing unknown fragment lengths with a reference allelic ladder containing fragments with known repeat-lengths. The use of a simple number, representing the number of repeats was sufficient as nomenclature for STR allele variation. Recent developments in Massive Parallel Sequencing (MPS) technologies enable high-throughput sequencing of STRs, revealing additional sequence-variation in many of the STRs [1, 2]. A uniform nomenclature for MPS-STR alleles describing this additional variation still needs to be developed. Here, we propose a universal way of describing STR allele variation, specifically designed for use in forensic casework.

Material and Methods

Previously suggested ways of describing STR- and other genome-variation were compared. Based on published variation [2] and in-house available data for 22 STRs (van der Gaag et al., manuscript in preparation) Human genome coordinates were identified for the genomic regions containing STR-variation, and rules were developed to describe sequence allele-variation within STRs and in repeat-flanking regions.

Chapter 3

Results

In several studies, MPS of STRs revealed substantial sequence variation in addition to that already described in STRbase [2, 4]. For comparison of published data and, more importantly, for comparison of profiles among databases, it is important that a uniform and straight-forward nomenclature is used. For this nomenclature several aspects should be taken into consideration:

- Different MPS assays will be used to analyse the same markers, introducing variation in the genome-coordinates of the analysed fragment containing an STR (fragments of different assays will not completely overlap).
- Allele-nomenclature should be compact and readable. However, allele-description should also contain all relevant information to reconstruct the original sequence.
- A direct comparison between CE- and MPS-results should be possible.

For CE-nomenclature, ranges have been determined in the past, defining the genomic coordinates of the STRs. For some STRs (like the example of D13S317 discussed below), additional repeating elements adjacent to the defined STR turned out to vary in length resulting in a difference between the total number of variable repeat-units for sequencing and the CE allele-count. Here, we determined genomic coordinates for the region in which STR-variation has been observed for 22 commonly used autosomal STRs (Figure 1a) in the following way:

The start-position of the STR-motif represents the first possible position while retaining maximum length for the longest repeated element. Any repeated elements directly adjacent to the STR of at least three repeats long was included as part of the STR-region. If a complex repeat consists of multiple blocks, interruptions were divided into blocks of the same length where possible (D2S1338, D21S11, FGA and vWA in figure 1a).

For studies of genome variation, HGVS-nomenclature rules [5] describe almost any possible type of genomic variation including STRs. Based on the STR-variation analysed in this study we propose general rules for a straightforward forensic nomenclature that describes sequence-variation in STRs and flanking regions. We mostly follow HGVS-guidelines but some specific rules are optimised for the intended use in forensics.

Figure 1b shows an example of a hypothetical allele for marker D13S317.

Figure 1, determined genome-coordinates for STR-variation and example of allele-description for an STR sequence-variant of D13S317 according to nomenclature rules

Locus	Range STR-structure in Reference-sequence (GRCh37/hg19)	CE-length	STR description refseq:
D1S1656	chr1:g.230905351-230905426	16	AC[6]CTAT[16]
TPOX	chr2:g.1493423-1493454	8	TGAA[8]
D2S1338	chr2:g.218879582-218879673	23	GGAA[2]GGAC[1]GGAA[13]GGCA[7]
D2S441	chr2:g.68239079-68239126	12	TCTA[12]
D3S1358	chr3:g.45582231-45582294	16	TCTA[1]TCTG[1]TCTA[14]
FGA	chr4:g.155508886-155508973	22	AAGG[3]AGAA[14]AGAG[1]AAAA[1]AAGA[3]
D5S818	chr5:g.123111246-123111293	11	CTCT[1]ATCT[11]
CSF1P0	chr5:g.149455885-149455936	13	CTAT[13]
D7S820	chr7:g.83789540-83789591	13	TCTA[13]
D8S1179	chr8:g.125907107-125907158	13	TCTA[1]TCTG[1]TCTA[11]
D10S1248	chr10:g.131092508-131092559	13	GGAA[13]
TH01	chr11:g.2192316-2192343	7	TGAA[7]
D12S391	chr12:g.12449953-12450028	18	TAGA[12]CAGA[7]
vWA	chr12:g.6093125-6093208	17	GATG[2]GATA[1]GATG[1]GATA[11]GACA[5]GATA[1]
D13S317	chr13:g.82722160-82722223	11	TATC[11]AATC[2]ATCT[3]
PentaE	chr15:g.97374242-97374266	5	TTTTC[5]
D16S539	chr16:g.86386308-86386351	11	GATA[11]
D18S51	chr18:g.60948900-60948981	18	AGAA[18]AA[1]AG[5]
D19S433	chr19:g.30417141-30417204	14	TCCT[13]ACCT[1]TCTT[1]TCCT[1]
D21S11	chr21:g.20554291-20554419	29	TCTA[4]TCTG[6]TCTA[3]TATC[4]A[1]TCTA[2]TCCA[1]TATC[12]
PentaD	chr21:g.45056086-45056155	13	AAAGA[13]AAAAA[1]
D22S1045	chr22:g.37536327-37536377	17	ATT[14]ACT[1]ATT[2]

a) Genomic coordinates and variant-description of the reference for 22 STRs

b) Example of the reference and variant-sequence with description of the allele

Reference-sequence (GRCh37:g.82.722.130-82.711.225)

tctaacgcctatctgtatttacaaatacat tatc tatc tatc tatc tatc tatc tatc tatc tatc tatc tatc aatc aatc atct atct atct tt

Variant-sequence:

TCTAACACCTATCTGTATTTACAAATACAT TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC TATC AATC ATCT ATCT ATCT TT

Allele-description following suggested nomenclature:

D13S317-CE12–chr13-GRCh37-g.82.722.160:82.722.223–TATC[13]AATC[1]ATCT[3]–x.136G>A

Chapter 3

From left to right, allele-description contains the following 4 elements:

1. Locus-name followed by "CE" and the allele-length (described as in current CE-nomenclature).
2. Chromosome-coordinates of the STR-motif in the reference sequence (including reference genome-version), representing the position of the first, and the last base of the STR-motif in the reference genome. If coordinates of the total analysed fragment (range between the used primers) are known, this information should always be provided (as a separate table).
3. STR-motif, described in the same orientation as the reference genome. For example: TATC[13]AATC[1]ATCT[3] (TATC repeated 13 times followed by AATC repeated 1 time and ATCT repeated 3 times).
4. Variation outside the STR-region (but within the analysed fragment) is described relative to the reference by genome position. Variants are described in the following order: Genome coordinate, reference > variant. The long number of the genome coordinate can be shortened by a 'x.' followed by the last three numbers (since the total coordinates of the STR-motif are already described before).

- A SNP of G>A on position 82.722.136 can be described as x.136G>A
- For a deletion of GC on position 82.722.136-82.722.137 we only write the starting-position of the deletion followed by the variation, for example: x.136GC>del.
- An insertion of AT after the same G is described as x.136.1–>insAT ('–' before the '>' is used since this position is absent in the reference).

Although we understand the suggested use of rs-nr.'s by Gelardi et al [1] to maintain a stable allele-name over different genome-versions, this will still result in different ways of describing variants within the same table because there can always be SNPs that are not listed in dbSNP. Rs-nr.s do not provide all the information that is needed to directly translate an allele-name back to the original sequence since the exact position of the SNP will need to be retrieved from dbSNP. Comparison of results from different assays (using different primers) is complicated if it is not directly visible from the name whether a SNP is within the range of both assays or not. Thereby, we prefer the use of genome-coordinates over rs-nr's. However, it is essential that the version of the used reference genome is described.

In addition to the CE-fragment allele-name, our rules have some small deviations from the HGVS nomenclature. As in HGVS nomenclature, only the positions that differ from the reference are displayed, but they are described in the fixed order of ancestral > derived sequence. This is different from HVGS since deletions and insertions of one nt are described as delA (in our rules A>delA) and insA (in our rules x.1–>A).

This fixed order keeps the description directly translatable and is feasible for the use of computer-algorithms to automatically compare variants. To avoid accumulation of numbers, larger insertions and deletions are described in a more mtDNA-like fashion [3] displaying only the start-position of the variant followed by the ancestral > derived sequence. To combine all parts of the allele-name, an en-dash (long dash) was chosen as delimiter between the separate parts to leave an open space between the different parts and increase readability. Although we provide a method that can be used to describe variants without prior knowledge of the exact positions of the primers, we recognise that this is a suboptimal situation which limits possibilities in comparison of STR sequencing-results between different assays.

Conclusion

Recommendations have been made for nomenclature of STRs in such a way to provide maximum information in the allele-name and help direct comparison of data from different assays and between CE- and sequencing-data.

Acknowledgements

We thank Rick de Leeuw and Jerry Hoogenboom for their contribution in the discussion of the proposed nomenclature. This study was supported by a grant from the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands.

Conflict of Interest Statement

The Authors declares that there is no conflict of interest.

Reference List

1. Gelardi C, Rockenbauer E, Dalsgaard S, Borsting C, Morling N. Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles. Forensic Sci.Int.Genet. 2014; 12:38-41
2. Gettings KB, Aponte RA, Vallone PM, Butler JM. STR allele sequence variation: Current knowledge and future issues. Forensic Sci.Int.Genet. 2015; 18:118-30
3. Parson W, Gusmao L, Hares DR, Irwin JA, Mayr WR, Morling N, Pokorak E, Prinz M, Salas A, Schneider PM, Parsons TJ. DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing. Forensic Sci. Int.Genet. 2014; 13:134-142
4. Ruitberg CM, Reeder DJ, Butler JM. STRBase: a short tandem repeat DNA database for the human identity testing community. Nucleic Acids Res. 2001; 29:320-322
5. Taschner PE, den Dunnen JT. Describing structural changes by extending HGVS sequence variation nomenclature. Hum.Mutat. 2011; 32:507-511

Chapter 3