Forensic Science International: Genetics Supplement Series 8 (2022) 23–25

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

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

Identification of STR/Y-STR alleles with tolerance for variants and stutter detection using GeneMarker®HTS software

Lidong Luo, Yiqiong Wu, James Todd, James Ruth, Eric Podlaszewski, Sarah Copeland, Teresa Snyder-Leiby *, Changsheng Liu

SoftGenetics, LLC, State College, PA, USA

ARTICLE INFO

Keywords:
Next generation sequencing
Massively parallel sequencing
STR
Iso-alleles

ABSTRACT

GeneMarker®HTS is a rapid, user-friendly software for forensic mtDNA and STR analysis with high throughput sequencing (HTS)/ next generation sequencing (NGS)/ massively parallel sequencing (MPS) data. Compared to the traditional capillary electrophoresis (CE) allele separation, HTS data provides a precise description of the repeat allele structure for each STR locus including the variants in the flanking areas for iso-allele genotype reporting and for future use in probabilistic genotyping analyses. The variants in the repeat region or the flanking area may in some cases cause failure for identification of STR/Y-STR alleles. To increase the accuracy, an iterated sequence alignment method is used to detect alleles. First, we align sequences to GRCh38 to determine the repeat sequence. Then, using the repeat sequence, we perform a second alignment that minimizes alignment errors and improves accuracy for repeat number. Stutters are also identified by stutter filters. To validate the method, concordance study is made and the concordance with the CE allele calls for 22 Autosomal STR Loci, 22 Chr Y STR Loci and Amelogenin is 100 % for NIST-2391d samples and 99.93 % for a 651 sample NIST-Promega dataset. The identification error due to sequence variants coupled with STR repeats is solved.

1. Short tandem repeats (STRs) analysis

GeneMarker®HTS analyzes high throughput sequencing data for forensic applications by selecting a built-in panel or by loading a custom panel. There are four built-in panels which correspond to Promega®PowerSeq® 46GY/56GMY/CRM/Mito System. The panels include analysis information for Amelogenin, Autosomal STR, ChrY STR and Mitochondrial DNA. Both built-in and custom panels require primer sequences which are used to sort and trim the input reads to different markers. For paired-end data, primer sequences are also used to merge the input reads with overlapping sequences to help correct any sequencing errors.

1.1. Identification of STR/Y-STR alleles with tolerance for variants

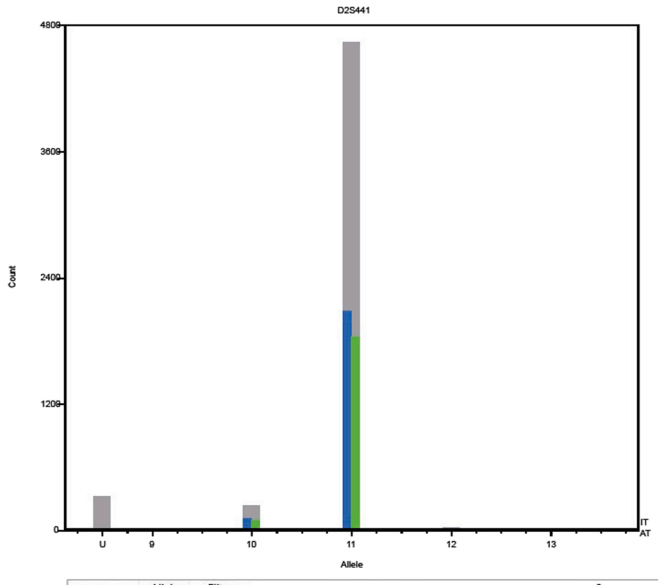
The STR or specific sequence (Amelogenin) is identified using an iterated sequence alignment method and regular expressions (regex strings). The iterated sequence alignment method implemented local sequence alignment [1–3] twice. Firstly, the reads are aligned to reference segments from GRCh38, GRCh37, or any standard reference segment that the STR belongs to. With the result from the first alignment, a new reference with similar repeat number is generated, and another alignment is made to name the allele sequences. The names of the allele sequences are compatible with the existing CE-based STR data. Additionally, we use the bracketed repeat [4,5] as the format for STR sequences based on the reference sequence direction. The bracketed repeat format accommodates sequence variation in the left/right flanking area outside the repeat region by means of variant calls, where variations 5′ or 3′ of the repeat region have negative or positive position numbers, respectively. The bracketed repeat format with variants in the flanking areas provides iso-allele genotype reporting.

For example, the bracketed repeat of the read:

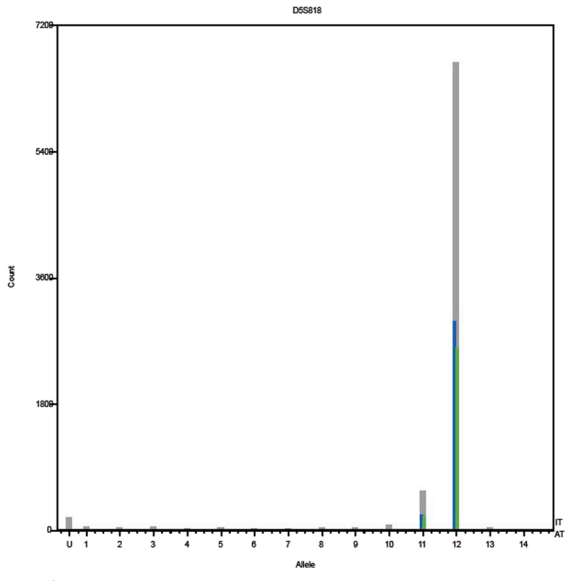
“aacatttgtatctttatctgtatccttatttatac**a**tctATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTtcaaaatattac**g**taaggataccaaagaggaaaatcacccttgtcacatacttgctattaaaatatactttttattagtaca” of Marker D5S818 from NIST SRM 2391d (NIST Certificate of Analysis Standard Reference Material 2391d PCR-Based DNA Profiling Standard https://www-s.nist.gov/srmors/certificates/2391d.pdf) Component B is “ATCT[12]

* Corresponding author.
E-mail address: teresa@softgenetics.com (T. Snyder-Leiby).

https://doi.org/10.1016/j.fsigss.2022.09.009
Received 19 September 2022; Accepted 23 September 2022
Available online 24 September 2022



Marker	Allele Name	Filter Status	Filter Reason	SW Call	Bracket Sequence	Sequence Total Count
D2S441	10	Passed		-1 Stutter	TCTA[10]	110
D2S441	10	Flagged	AT% < x < IT%	-1 Stutter	TCTA[8]TCTG[1]TCTA[1]	89
D2S441	11	Passed		Called	TCTA[9]TCTG[1]TCTA[1]	2080
D2S441	11	Passed		Called	TCTA[11]	1837



Marker	Allele Name	Filter Status	Filter Reason	SW Call	Bracket Sequence	Sequence Total Count
D5S818	11	Passed		-1 Stutter	ATCT[11]_+12A>G	218
D5S818	11	Passed		-1 Stutter	ATCT[11]_+16A>G	193
D5S818	12	Passed		Called	ATCT[12]_-4C>A_+13A>G	2975
D5S818	12	Passed		Called	ATCT[12]_+13A>G	2595

Fig. 1. Example of D2S441 with alleles: TCTA[9]TCTG[1]TCTA[1] and TCTA[11] and example of D5S818 with alleles: ATCT[12]_-4 C>A+ 13 A>G and ATCT [12]_+ 13A>G.

_-4C>A_+13A>G".“−4C>A” indicates that an A nucleotide is encountered 4 bases 5′ of the repeat region, whereas the reference sequence has a C in that position; and + 13 A>G indicates that a G nucleotide is at 13 bases 3′ of the repeat region. Though there is a variation which makes part of the left flank region the same as the repeat region (variant A makes CTCT in the left flank to ATCT), the start of the repeat region is recognized accurately with our iterated sequence alignment method.

1.2. Stutter detection

After the data is processed, the results can be viewed in an initial view which has all the reads merged with the same sequence. Low-frequency sequences caused by sequencing errors are also kept and calculated for manual review. With filter settings applied, low-frequency sequences can be filtered, and stutters can be labeled. The STR filters settings are based on analytical/interpretive thresholds and the STR stutter settings can be set for each marker with type and ratio.

1.3. CE results concordant with HTS results

NIST SRM 2391d dataset has pair-end fastq data for three samples: Component A/B/C. GeneMarker®HTS software with built-in panel “Promega®PowerSeq®46GY” outputs the results of 22 Autosomal STR Loci, 22 Chr Y STR Loci and Amelogenin. The Genotypes of the resulting calls are 100 % concordant with the certified genotypes/haplotypes.

National Institute of Standards and Technology (NIST), in conjunction with Promega corporation, generously supplied fastq sequence files of pair-end data for 673 samples, in which 651 samples have the corresponding CE allele calls amplified with the PowerSeq™Auto/Y System and analyzed on an Illumina®MiSeq. GeneMarker®HTS software results are 99.93 % concordant with the CE allele calls of 29,295 sampled loci. Discordant allele names were caused by variants on the boundary of the repeat and flanking region, and mainly on DYS385a/b, DYS389II, Penta E, and Penta D loci.

1.4. Additional sequence information

High throughput sequencing data can reveal additional information that is not available from the traditional CE data. Iso-alleles are loci that appear homozygous in length-based measurements (such as CE) but are heterozygous by sequence. High throughput sequencing data reports the percentage of sequences for a given allele and sequence variants. This depth of information has applications in identification of individuals and relatives in single source samples and the potential for improved assignment to contributors during analysis of mixtures. In GeneMarker®HTS, the STR Analysis screen provides histograms with blue/green color to highlight iso-alleles and bracket sequences to summarize the differences of sequences. Fig. 1 shows an example of D2S441 with alleles: TCTA[9]TCTG[1]TCTA[1] and TCTA[11], and also shows an example of D5S818 with alleles: ATCT[12]_-4C>A+13A>G and ATCT [12]_+13A>G.

2. Conclusion

Chemistries for mtDNA and STR amplification for HTS platforms enable the laboratory to have the benefits of both mtDNA and STR analysis at the same time. GeneMarker®HTS software provides a streamlined workflow for forensic mitochondrial and STR DNA data analysis from all major high throughput sequencing (HTS) systems and chemistries.

An iterated sequence alignment method is introduced to detect the STR/Y-STR alleles correctly with variants in the repeat region or the flanking area for GeneMarker®HTS. The results are 100 % concordant with CE allele calls for NIST SRM 2391d samples and 99.93 % concordant with CE allele calls of 29,295 sampled loci, compared to the 2019 concordance study reporting 99.74 % concordance with CE allele calls [6]. High throughput sequencing data can reveal additional information that is not available from the traditional CE data. The additional sequence information can be beneficial in forensic casework applications. Strengths of this data include both its resolving power for excluding an individual and the ability to determine potential

relationships between evidence and suspects due to Mendelian inheritance of nuclear DNA.

Acknowledgments

We would like to thank Dr. Peter Vallone at National Institute of Standards and Technology (NIST) for generously supplying data to complete the concordance study between the CE results and HTS STR/Y-STR results. We would also like to thank Promega Corporation, Madison, WI, USA for providing Autosomal and Y-STR data, and Drs. Mitchell Holland, Jennifer McElhoe, and John McGuigan at Penn State University for their comments/suggestions during the mitochondrial/STR analysis development.

References

[1] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197.

[2] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[3] M. Zhao, W.P. Lee, E.P. Garrison, G.T. Marth, SSW library: an SIMD Smith-Waterman C/C++ library for use in genomic applications, PLOS One 8 (2013), e82138.

[4] K. Gettings, D. Ballard, M. Bodner, L. Borsuk, J. King, W. Parson, C. Phillips, Report from the STRAND working group on the 2019 STR sequence nomenclature meeting, Forensic Sci. Int.: Genet. (2019), 102165.

[5] 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. Knijff, N. Morling, M. Prinz, P.M. Schneider, C. V. 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.

[6] C.S. Liu, L. Luo, J. McGuigan, J. Wu, J. Todd, C. Prosser, S. Copeland, T. Snyder-Leiby, High throughput sequencing data analysis workflow: mtDNA variant detection and identification of STR/Y-STR alleles and iso-alleles, Forensic Sci. Int.: Genet. Suppl. Ser. 7 (2019) 639–640.