Forensic Science International: Genetics Supplement Series 7 (2019) 115–116

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

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

Understanding the behavior of stutter through the sequencing of STR alleles

Sarah Riman[a], Hariharan Iyer[b], Lisa A. Borsuk[a], Peter M. Vallone[a,*]

[a] U.S. National Institute of Standards and Technology, Applied Genetics Group, 100 Bureau Drive, Gaithersburg, MD, 20899-8314, USA
[b] U.S. National Institute of Standards and Technology, Statistical Design, Analysis, and Modelling Group, 100 Bureau Drive, Gaithersburg, MD, 20899-8980, USA

ARTICLE INFO

Keywords:
Sequencing
DNA
STR
Stutter

ABSTRACT

This work explores the influence of several variables on stutter formation across sequenced autosomal STR loci (simple, compound, and complex motifs) and different alleles within each locus. The variables are sequence variations within the repeating motifs and flanking region [1,2]; longest uninterrupted stretch (LUS) [3]; parental allele length [3]; and base pair content and length value of each repeating motif from which the stutter has generated [3,4]. Over six hundred unrelated individuals from different populations were amplified with the prototype PowerSeq 46GY System and sequenced on the Illumina MiSeq platform. Raw FASTQ files were analyzed with STRait Razor v3 [5]. Stutter ratio was calculated for motifs that exhibited stutter using the ratio of the observed coverage of the stutter sequence at (N-1) position to the observed coverage of the allelic sequence. Understanding the behavior (abundance, reproducibility, sequence context) of non-allelic artifacts will help in establishing probabilistic models for the prediction of stutter rate and interpretation of sequence-based STR profiles.

1. Introduction

PCR amplification is a standard step for enriching and targeting STR loci for high-throughput sequencing, thus stutter products and additional artifacts are generated and observed with sequencing data. Artifacts are in the form of (1) substitutions (sequences with the same length of the parent allele and/or stutter but with ≥ 1 base substitution); (2) insertions; and (3) deletions [6]. Stutter and artifacts can cause a challenge in accurately genotyping alleles and assigning the number of contributors when interpreting low-level mixtures. Thus, understanding the behavior of artifacts and characterizing the different stutter variants generated from the various motifs of the same parent allele will assist in generating models for interpreting sequence-based STR profiles. Studying artifacts can also help in establishing analytical thresholds to discern between allelic and non-allelic sequences. In this work, we focus on studying those sequences observed for 22 autosomal STR loci present in the prototype PowerSeq 46GY System (Promega, Madison, WI) and sequenced on the Illumina MiSeq FGx platform (Illumina, San Diego, CA).

2. Methods

2.1. Sample sequencing

672 NIST population samples were sequenced using the prototype PowerSeq 46 GY System (Promega, Madison, WI) following manufacturers' protocol using 1 ng of DNA template. The resulting library products were sequenced on a MiSeq FGx instrument (Illumina, San Diego, CA) using the MiSeq Reagent Kit v3 600-cycle (Illumina). R1 and R2 reads were collected. The average coverage per sample was greater than 190,000 X (for all 46 loci).

2.2. Data analysis

The FASTQ files were analyzed using STRait Razor 3.0 [5]. Data were further parsed using custom tools in R. The 22 autosomal STR loci in the kit were examined. Sequences observed at greater than one percent of the total locus coverage were binned and further examined. This analysis was performed solely for homozygous loci only to avoid sequence contributions from a sister allele that might confound the initial characterization.

* Corresponding author.
E-mail address: peter.vallone@nist.gov (P.M. Vallone).

https://doi.org/10.1016/j.fsigss.2019.09.045
Received 5 September 2019; Accepted 23 September 2019
Available online 30 September 2019

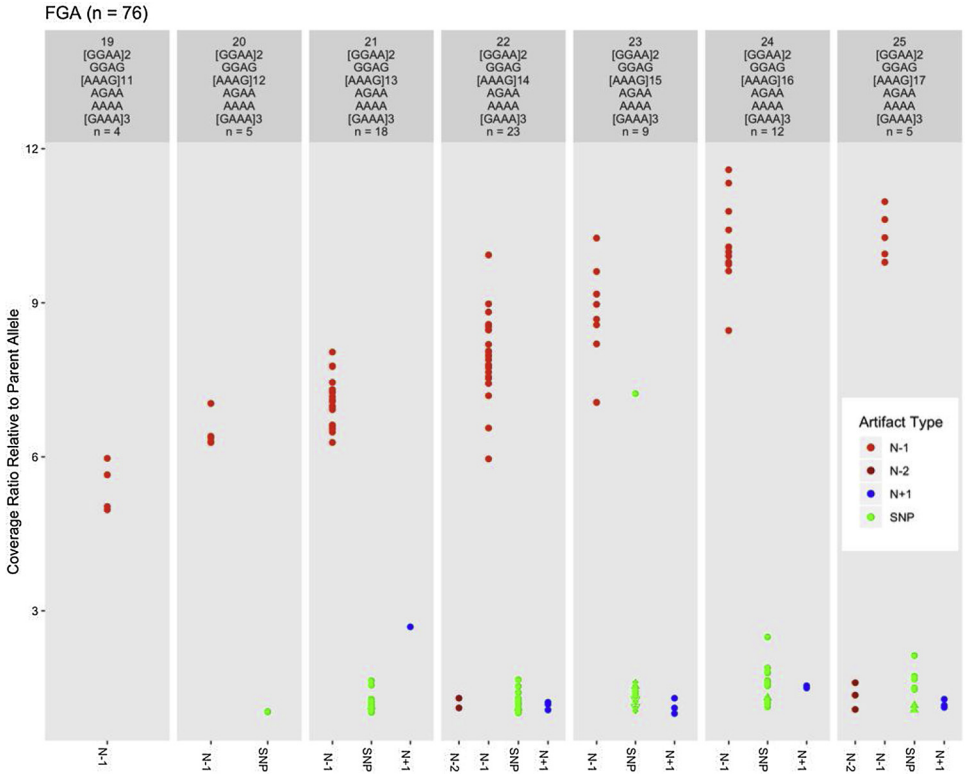

Fig. 1. Sequencing artifacts observed for FGA.

3. Results and discussion

3.1. Sequencing artifacts and stutter observed for FGA

Fig. 1 illustrates the results from examining the autosomal STR locus FGA with the compound repeat structure: [GGAA]n GGAG [AAAG]n AGAA AAAA [GAAA]n. The artifacts for seven allelic sequences observed in our data are shown. A total of 76 homozygous instances for FGA were present in the dataset. The y-axis is the coverage ratio (in percent) relative to the parent allele per sample. The bracketed motif for each allele is shown at the top of the plot. The red circles represent N-1 stutter that ranged from 4.9 to 11.6%. The trend of the N-1 stutter was relatively linear ($r^2 = 0.986$) for FGA and followed expectations of increasing with the LUS. For alleles with a greater number of observations (alleles: 21, 22, 23, 24) the range of N-1 stutter within the allele varied up to 4%. Darker red circles represent N-2 stutter products. Observations of N + 1 stutter (blue circles) were present at longer alleles (21–25). The green circles and triangles represent single base substitutions observed in the parent allele. These may have originated from polymerase misincorporation or sequencing bias and warrant further investigation.

4. Conclusions

The targeted sequencing of STR markers will further elucidate not only variations in stutter motifs but also additional reproducible artifacts that may originate from polymerase error, sequence platform bias, etc. Understanding the behavior (abundance, reproducibility, sequence context) of sequences other than the ‘pure’ parent allele will help in establishing probabilistic models for the prediction of stutter rate and interpretation of sequence-based STR profiles.

Role of funding

This work was funded by the NIST Special Programs Office and the FBI Biometric Center of Excellence Unit: DNA as a Biometric.

Declaration of Competing Interest

None.

Acknowledgments

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 Commerce. 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 NIST, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose. This work was approved by the NIST Human Subjects Protection Office. We wish to thank Doug Storts and Promega for access to the PowerSeq 46GY chemistry, Becky Steffen and Kevin Kiesler of the Applied Genetics Group (NIST) for performing the supporting sequencing experiments.

References

[1] A.E. Woerner, et al., Flanking variation influences rates of stutter in simple repeats, Genes (Basel) 8 (2017) 1–20.
[2] A.E. Woerner, et al., Compound stutter in D2S1338 and D12S391, Forensic Sci. Int. Genet. 39 (2019) 50–56.
[3] C. Brookes, et al., Characterizing stutter in forensic STR multiplexes, Forensic Sci. Int. Genet. 6 (2012) 58–63.
[4] S.B. Vilsen, et al., Stutter analysis of complex STR MPS data, Forensic Sci. Int. Genet. 35 (2018) 107–112.
[5] A.E. Woerner, et al., Fast STR allele identification with STRait razor 3.0, Forensic Sci. Int. Genet. 30 (2017) 18–23.
[6] G. Park, et al., Characterization of background noise in capture-based targeted sequencing data, Genome Biol. 18 (2017) 136.