Sequence-based Analysis of Stutter at STR Loci: Characterization and Utility

Rachel A. Aponte[a], Katherine B. Gettings[b], David L. Duewer[b],
Michael D. Coble[b], and Peter M. Vallone[b]

[a]Department of Forensic Sciences, The George Washington University, Washington, DC 20007-1150, USA

[b]U.S. National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA

Corresponding Author: Katherine Butler Gettings, +01-301-975-6401 (tel), katherine.gettings@nist.gov

Abstract: The development of next generation sequencing (NGS) technologies creates the potential for changing the method by which the forensic science community genotypes short tandem repeat (STR) loci. While the capabilities of NGS are promising, moving from current capillary electrophoresis (CE) methods would require new guidelines to be established and a new understanding of artifacts that may arise with the use of NGS. Stutter has been well characterized for CE technologies; however, NGS workflows may use different polymerases and amplification approaches, which could alter the appearance of this artifact. Stutter is most commonly seen in the n-4 position in CE data, but may be observed more rarely in the n+4 and n-8 positions. NGS data frequently contains detectable sequences consistent with stutter at the n+4 and n-8 positions, and may even contain stutter at the n-12 position for some loci. It is possible that these alternate types of stutter events occur at similar levels in CE workflows and go undetected due to the analytical threshold employed or because the artifacts do not exceed the background noise. Comparing stutter events in NGS data to what has been observed by CE will improve our understanding of the effects of library preparation and sequencing. Characterizing stutter events by sequence will contribute to the development of guidelines and facilitate implementation of NGS technology. Further, determining stutter ratios for each isoallele would allow for individual sequence thresholds to be set, which could then be used to improve mixture interpretation models.

Keywords: stutter, next generation sequencing (NGS), short tandem repeat (STR)

1. Introduction

Stutter is a commonly encountered artifact in forensic DNA typing of short tandem repeat (STR) loci that can interfere with data interpretation. It is caused by slippage of the DNA polymerase during the extension phase of the polymerase chain reaction (PCR), and results in the deletion or addition of one repeat unit in the nascent DNA strand produced [1]. While stutter is straightforward for simple repeat loci, compound and complex loci require more attention to understand this artifact. The longest uninterrupted stretch (LUS) is the longest consecutive portion of the same repeat unit within a compound allele [2]. The LUS may be more predictive of expected stutter percentage than the total number of repeats within an allele [1,2]; however, stutter may occur in any repeated portion of the motif. Discerning the origin of stutter within the sequence, and how different sequence motifs affect stutter, will allow for more specific stutter thresholds to be set and may improve mixture profile interpretation.

2. Methods

A collection of two 96-well plates including population samples with individuals from three population groups – Caucasian, African American, and Hispanic – were analyzed for this study.

2.1 Capillary Electrophoresis

Capillary electrophoresis (CE) data was generated using Promega PowerPlex® Fusion STR multiplex amplification kit (Promega Corporation, Madison WI, USA), as described previously [3]. For this study, profiles were analyzed using GeneMapper® *ID-X* software version 1.3, with the peak detection threshold set to 10 relative fluorescent units (RFU) and no stutter filter applied. All peaks were verified by two analysts and stutter peaks indistinguishable from noise, affected by bleed-through, or with poor morphology were discarded. Stutter ratios were calculated by dividing the RFU value of the stutter peak height by the RFU value of the allele peak height.

2.2 Next-Generation Sequencing

Generation of the NGS data analyzed herein using a beta version of the Promega PowerSeq® Auto kit has been previously described [4]. Bioinformatic analysis specific to this study is as follows: FASTQ files for each sample were parsed into separate loci using STRait Razor version 1.5 [5], then the output files were filtered through an excel program specifically designed to filter out noise (sequencing errors) from true alleles and stutter peaks. Stutter ratios were calculated by dividing the sequence coverage value of stutter by the sequence coverage value of the allele.

3. Results and Discussion

Results from CE and NGS data were analyzed and compared. Two compound repeat loci were included in the final analysis: D2S441 and D8S1179. 186 samples produced a profile by CE, while 79 samples were analyzed from the NGS data set. After excluding overlapping or adjacent alleles from analysis, at D8S1179, 137 alleles were analyzed by NGS and 115 alleles by CE, and at D2S441, 123 alleles were analyzed by NGS and 168 alleles by CE. The range of n-4 stutter peak heights fell between 10 and 87 RFU for D8S1179 and between 15 and 183 RFU for D2S441. The range of n-4 stutter sequence coverage extended from 16X to1084X for D8S1179 and from 10X to 651X for D2S441.

Figure 1 depicts n-4 stutter percentages for CE and NGS data by allele and LUS, for both D8S1179 and D2S441. Several rare motifs and microvariants were excluded from the graphs and calculations.

At D8S1179, the upper two graphs representing stutter by allele show three distinct trends corresponding to the different sequence motifs for NGS and CE. The graphs representing stutter by LUS, tends to show a more uniform average for the different sequence motifs by both NGS and CE. The average n-4 stutter percentages observed from the NGS data were approximately 3% higher than those of CE, indicating a generally higher stutter rate in NGS than in CE. The average range of stutter percentage observed per allele was also more widespread in NGS data (5%) compared to that of CE (2.8%).

At D2S441, the upper two graphs representing stutter by allele again display three distinct trends, corresponding to the different sequence motifs by NGS and CE. The lower graphs representing stutter by LUS; however, do not align the stutter percentages for different sequence motifs as was the case for D8S1179. At D2S441, it appears that the LUS is not the only factor contributing to the stutter percentage and the context of the sequence is also important. Moreover, the compound motif [TCTA]n[TTTA][TCTA]2 appears to result in a reduced incidence of stutter compared to the simple motif [TCTA]n. The n-4 stutter percentages observed from the NGS data were more closely aligned to the CE data (average 0.7% higher by NGS) for D2S441 compared to D8S1179. The average range of stutter percentage per allele is again greater for NGS (5.3%) than for CE (3.6%).

D2S441 contained frequent n+4 stutter by NGS in the 2-4% range. The CE data for D2S441 contained occasional n+4 stutter peaks with an average of 2%. D8S1179 contained frequent n+4 stutter by NGS, sporadic by CE, both detectable in the 1-2% range. For both loci, n-8 stutter was observed more rarely by NGS in the 0.5-1% range, and only once by CE. The rarity of n+4 and n-8 stutter by CE may be attributable to the generally low peak heights within the CE data set.

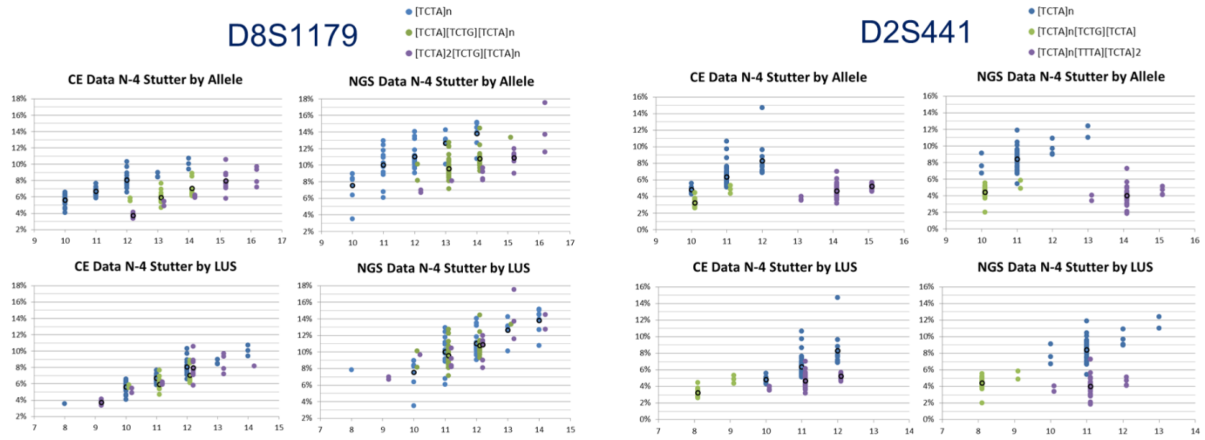

Figure 1. n-4 stutter percentages for three repeat motifs at D8S1179 and D2S441 from CE and NGS data, plotted by allele and LUS. Averages are indicated by black circles when ≥5 measurements were present. All alleles graphed are whole alleles; offsets in data points are for visualization purposes only. Sequence motifs shown in CE data are inferred from the NGS data.

4. Conclusions

Variation in level of stutter is attributable to the sequence motif, as demonstrated by both NGS and CE data sets. Differences in the surrounding sequence may have an effect on stutter percentages, regardless of LUS, for some loci. Observing additional compound/complex loci with various sequence motifs and comparing by allele and LUS will aid in understanding this phenomenon. Stutter percentages appear generally higher for NGS data than CE, but interlocus variation is anticipated. A CE data set with higher RFU would allow for better CE/NGS comparisons of stutter in the 1-2% range.

Future research includes extending this study to include all 22 loci compared between NGS and CE platforms, and further expanding to additional assays, in order to better understand how sequence motif affects stutter. Characterization of stutter by NGS will help establish future guidelines. Allele and sequence-based stutter thresholds will allow better differentiation of artifact from minor contributors compared to global thresholds currently applied to CE data. This is expected to offer improvements in mixture profile interpretation.

5. Acknowledgements: PowerPlex® Fusion CE data was provided by Becky (Hill) Steffen.

6. Conflict of Interest: none

7. Funding Source: This work was funded in part by the Federal Bureau of Investigation (FBI) interagency agreement DJF-13-0100-PR-0000080: "DNA as a Biometric".

8. References

[1] P.S. Walsh, N.J. Fildes, R. Reynolds, Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA, Nucleic Acids Research, 24 (1996) 2807-2812.

[2] J.A. Bright, K.E. Stevenson, M.D. Coble, et al., Characterising the STR locus D6S1043 and examination of its effect on stutter rates, Forensic Sci Int Genet. 8 (2014) 20-23.

[3] K. Oostdik, K. Lenz, J. Nye, et al. Developmental validation of the PowerPlex® Fusion System for analysis of casework and reference samples: A 24-locus multiplex for new database standards, Forensic Sci Int Genet. 12 (2014) 69-76.

[4] K.B. Gettings, K.M. Kiesler, S.A. Faith, et al., Sequence variation of 22 autosomal STR loci detected by next generation sequencing, Manuscript submitted (2015).

[5] D.H. Warshauer, D. Lin, K. Hari, 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.