SEQUENCE-BASED ANALYSIS OF STUTTER AT STR LOCI: CHARACTERIZATION AND UTILITY

National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce

Rachel A. Aponte[1], Katherine B. Gettings[2], David L. Duewer[2], Michael D. Coble[2] and Peter M. Vallone[2]

Email:
katherine.gettings@nist.gov

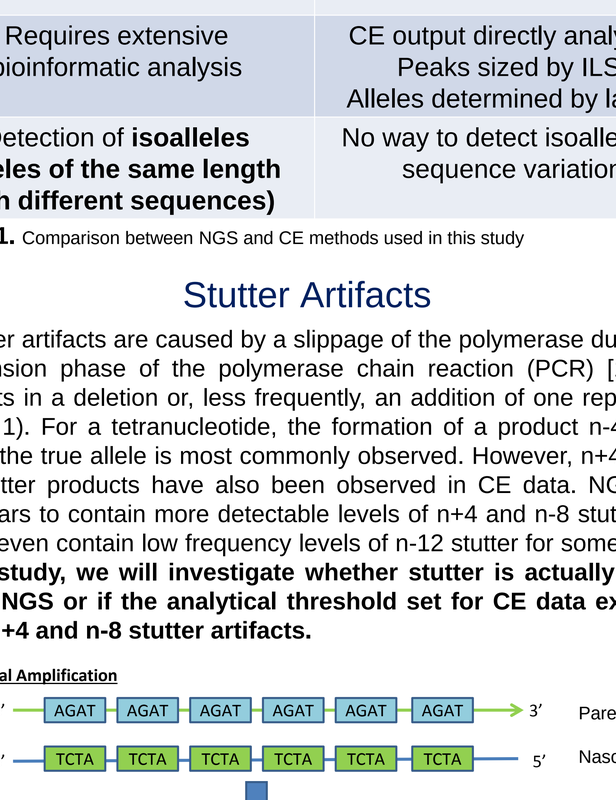
[1] Department of Forensic Sciences, The George Washington University, Washington, DC 20007-1150, USA
[2]U.S. National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA

P-145

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.

Next Generation Sequencing (NGS) vs Capillary Electrophoresis (CE)

Workflows used to generate this data set:



NGS	CE
Two amplifications	One amplification
Alleles sequenced in distinct clusters	Alleles separated based on size
Provides length (number of repeat units) and nucleotide sequence	Provides length (number of repeat units)
Signal measured as coverage (count of each sequence)	Signal measured as relative fluorescence units (RFU)
Requires extensive bioinformatic analysis	CE output directly analyzed: Peaks sized by ILS, Alleles determined by ladder
Detection of **isoalleles (alleles of the same length with different sequences)**	No way to detect isoalleles or sequence variation

Table 1. Comparison between NGS and CE methods used in this study

Stutter Artifacts

Stutter artifacts are caused by a slippage of the polymerase during the extension phase of the polymerase chain reaction (PCR) [1]. This results in a deletion or, less frequently, an addition of one repeat unit (Fig. 1). For a tetranucleotide, the formation of a product n-4 bases from the true allele is most commonly observed. However, n+4 and n-8 stutter products have also been observed in CE data. NGS data appears to contain more detectable levels of n+4 and n-8 stutter, and may even contain low frequency levels of n-12 stutter for some loci. **In this study, we will investigate whether stutter is actually higher with NGS or if the analytical threshold set for CE data excludes the n+4 and n-8 stutter artifacts.**

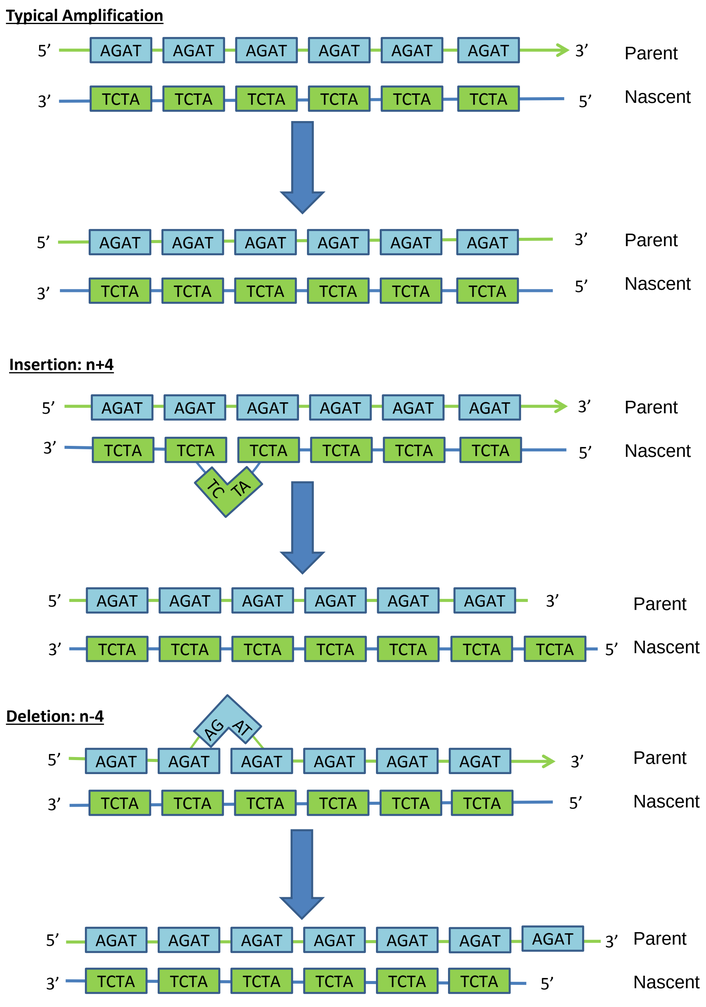


Figure 1. Model showing slippage events during amplification that will lead to stutter

Acknowledgements and Disclaimers: PowerPlex® Fusion CE data was provided by Becky (Hill) Steffen. This work was funded in part by the Federal Bureau of Investigation (FBI) interagency agreement DJF-13-0100-PR-0000080: "DNA as a Biometric". 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. Departments of Commerce or Justice. 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 the National Institute of Standards and Technology nor does it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

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, C.R. Hill, J.M. Curran, J.S. Buckleton, Characterising the STR locus D6S1043 and examination of its effect on stutter rates, Forensic Sci Int Genet. 8 (2014) 20-23.
[3] K.B. Gettings, R.A. Aponte, P.M. Vallone, J.M. Butler, STR allele sequence variation: Current knowledge and future issues, Forensic Sci Int, 18 (2015) 118-130.

Longest Uninterrupted Stretch (LUS)

Within the sequence, the LUS is the longest consecutive portion of the same repeat within an allele [2]. The LUS may be more predictive of expected stutter percentage than the total number of repeats. Therefore, setting analytical thresholds based on the LUS opposed to the total repeat number may provide a more accurate filter for NGS data analysis. In this study, this possibility is explored for D8S1179 and D2S441.

Allele	D8S1149 Repeat Structure	LUS
12	[TCTA]**12**	**12**
12	[TCTA] [TCTG] [TCTA]**10**	**10**
12	[TCTA]2 [TCTG] [TCTA]**9**	**9**

Table 2 (above). Different LUS values are found in isoalleles at D8S1179 [3].

Allele	D2S441 Repeat Structure	LUS
12	[TCTA]**12**	**12**
12	[TCTA]**10** [TCTG] [TCTA]	**10**
12	[TCTA]**9** [TTTA] [TCTA]2	**9**

Table 3 (above). Different LUS values are found in isoalleles at D2S441 [3].

Breakdown of Stutter by Sequence

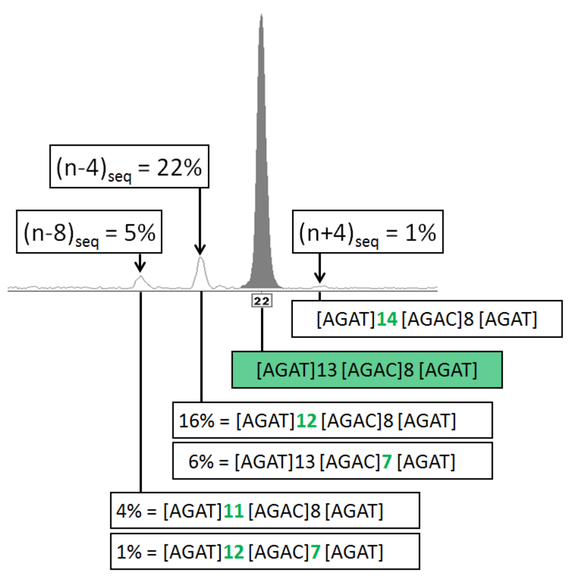


Figure 2. While stutter is more likely to happen in the LUS, it may occur from any part of the repeat pattern. NGS allows the analyst to observe exactly where stutter occurs within the sequence. This example of a 22 allele at D12S391 shows the breakdown of stutter observed with NGS at the n+4, n-4, and n-8 positions.

Experimental Design

Samples:

CE and NGS data was generated from two 96-well plates of population samples including Caucasian, African American, and Hispanic individuals.

Instruments:

ABI 3130xl Genetic Analyzer

Illumina MiSeq

Assays:

CE: Promega PowerPlex® Fusion (Beta version)
NGS: Promega PowerSeq® Auto (Beta version)

Loci Amplified by both CE and NGS:

AMEL, D1S1656, D2S1338, **D2S441**, D3S1358, D5S818, D7S820, **D8S1179**, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, CSF1PO, FGA, Penta D, Penta E, TH01, TPOX, vWA, DYS391

Analysis Parameters

CE: GeneMapper® ID-X software
- Detection threshold set to **10 RFU**
- No stutter filters
- Peaks called when distinguished from noise by two scientists

NGS: FASTQ File → Strait Razor → NGS stutter filter
- Analyzed at **10X minimum coverage**

Sample Analysis:

Two compound repeat loci were included in this analysis: D2S441 and D8S1179. 186 samples produced a profile for analysis by CE, while 79 samples were analyzed from the NGS data set. The number of alleles analyzed per locus is shown in the table, right.

	CE Alleles	NGS Alleles
D8S1179	115	137
D2S441	168	124

Results

NGS range of coverage:
D8S1179 n-4 stutter sequence coverage ranged from 16X-2084X
D2S441 n-4 stutter sequence coverage ranged from 10X-651X

CE range of peak heights:
D8S1179 n-4 stutter peak heights ranged from 10 – 87 RFU
D2S441 n-4 stutter peak heights ranged from 15-183 RFU

The average CE allele peak height was 940 RFU for D2S441 and 537 RFU for D8S1179; therefore, stutter peaks <2% were unlikely to be detected.

No correlation was detected between signal and n-4 stutter for the NGS or CE data.

Calculation of % Stutter Ratios:

$$\text{CE: } \frac{\text{RFU value of stutter peak height}}{\text{RFU value of allele peak height}} \times 100$$

$$\text{NGS: } \frac{\text{Sequence coverage value of stutter}}{\text{Sequence coverage value of allele}} \times 100$$

D8S1179 and D2S441 were analyzed and compared at the n+4, n-4, and n-8 stutter position between NGS and CE data. D2S441 contained frequent n+4 stutter by NGS in the 2-4% range. The CE data for D2S441 showed 19 instances of n+4 stutter peaks with an average of 2%. D8S1179 contained frequent n+4 stutter by NGS, and 9 instances by CE, 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 (at D8S1179) by CE.

The graphs below show n-4 stutter percentages for CE and NGS data by allele and LUS. All alleles graphed are whole alleles (offsets in data points are for visualization purposes only). There were several rare motifs and microvariants which were excluded from the graphs and following calculations.

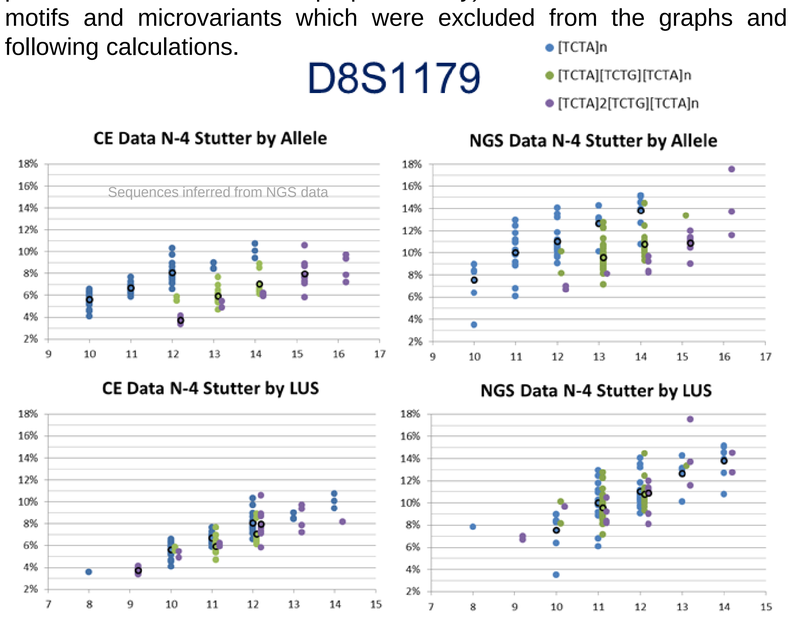


Figure 3 (above). At D8S1179, the upper graphs representing stutter by allele show three distinct trends corresponding to the three different sequence motifs for both NGS and CE. The graphs representing stutter by LUS, tends to show a more uniform average (averages indicated by black circles when ≥5 measurements were present) 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 range of stutter percentage observed per allele was also more widespread in NGS data (5.0% on average) compared to CE (2.8%).

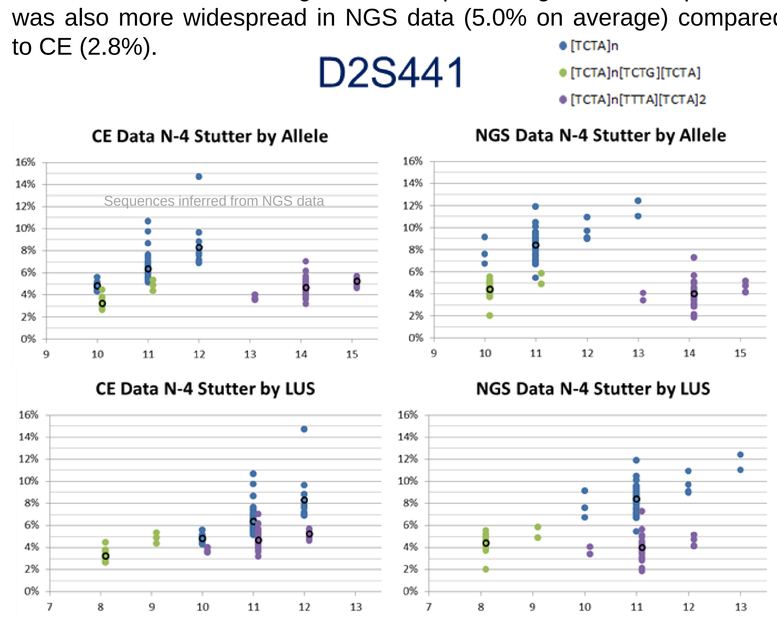


Figure 4 (above). At D2S441, the upper two graphs showing stutter by allele display three distinct trends, corresponding to the three different sequence motifs by both NGS and CE (similar to the results at D8S1179). The lower graphs representing stutter by LUS; however, do not align the stutter percentages 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 range of stutter percentage per allele is again greater for NGS (5.3% on average) than for CE (3.6%).

Conclusions and Future Directions

Variation in level of stutter artifact is attributable to the sequence motif, as demonstrated by both the NGS and CE data sets having distinct trends for the different sequence motifs observed. While D8S1179 shows similar averages between the motifs when grouped by the LUS, D2S441 averages do not normalize by LUS. This indicates that 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.

In this data set, there is no apparent correlation between coverage or peak height and percent n-4 stutter. Stutter percentages appear to generally average higher for NGS data than CE, but interlocus variation is anticipated. A CE data set with generally higher RFU would allow for better CE/NGS comparisons of stutter in the 1-2% range.

In the future, extending this study to include all 22 loci compared between NGS and CE platforms, and further expanding to additional assays, will aid in the understanding of 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.