Filtering Artifacts from Forensic PCR-MPS Data for Mixture Analysis

Brian Young, Tom Faris, Jeff Smith, Luigi Armogida
NicheVision Forensics brian@nichevision.com

Model of PCR-MPS Allele and Artifact Generation

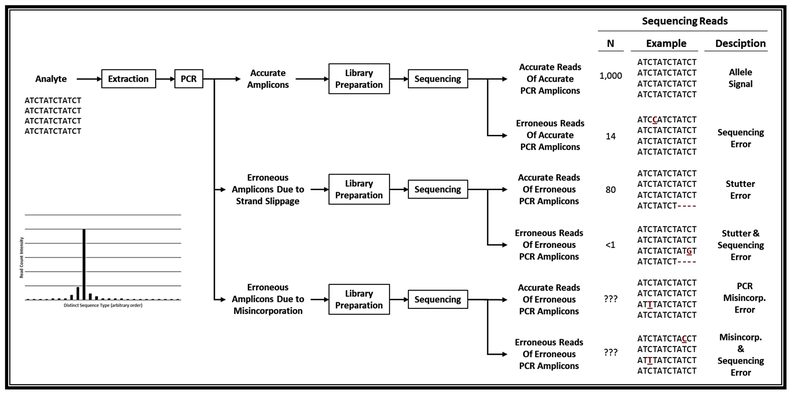

Figure 1. PCR-MPS is a hyphenated method involving repeated measurement of analytes. Allele signal arises from error-free analysis, while other distinct sequence types (DST) arise due to error.

Stutter Artifacts

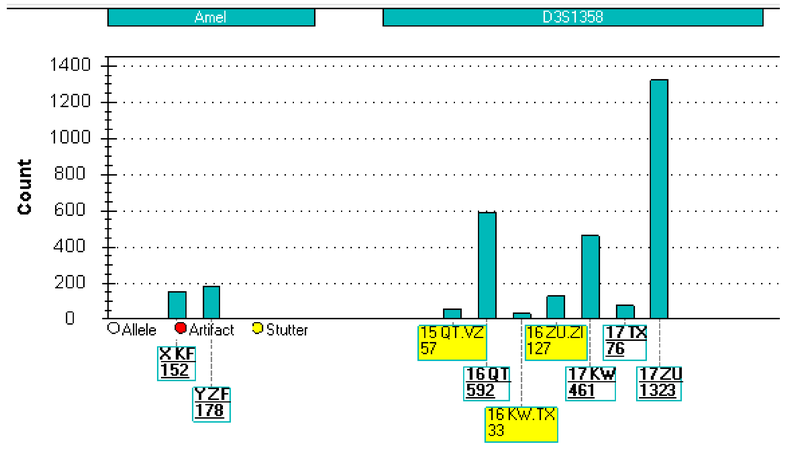

Figure 2. MixtureAce (NicheVision Forensics) display of a 1:3:6 3-person mixture from NIST PDR (1). Calling stutter at the motif level is necessary to attribute individual stutter to individual alleles. Stutter attribution is expressed by SID (2) labels in a "parent-dot-child" format.

Non-Stutter Artifacts

Table 1. Example of an allele, non-stutter artifact, and stutter artifact in reads targeting D2S1338. Artifacts can be attributed to parent alleles using SID labels in a 'parent-tic-child' format.

Category	Label	Bracketed Sequence of Haplotype Allele
Allele	25 BJ	GAG [GGAA]2 GGAC [GGAA]15 [GGCA]7 AGGCCAAGCCATTT
Non-Stutter	25 BJ`AS	GAG [GGAA]2 GGAC [GGAA]15 [GGCA]7 AGGCCAAGCCATGT
Stutter	24 BJ.SS	GAG [GGAA]2 GGAC [GGAA]14 [GGCA]7 AGGCCAAGCCATTT

PCR-MPS Read Sequences Are 'Haplotype Alleles'

Table 2. Trimmed read sequences of PCR-MPS amplicons are 'haplotype alleles' possibly containing multiple STR, SNP, and DIP markers. The entire sequence string is required to describe the allele, or to discriminate alleles from artifacts. Three isometric alleles targeting D16S539 are shown. Raw string and hash digests uniquely define haplotype alleles. SID labels using two letters of the hash digest generate pronounceable labels for labeling distinct sequence types within a locus profile.

Raw Sequence	SID Hash Digest	SID Label	Bracket Notation
TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTGGATAGATAGATAGATAGATAGATAGATAGATAGATAGATATCATTGAAAGACAAAC CAGAGATGGATGATAGATAC	QBCNTZTPIQYLRLCQCCIBFFRWAQZSBBVXQQGGXXTTFEWIJZKXYGMXPLC	10 QB	[GATA]10 rs11642858
TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTGGATAGATAGATAGATAGATAGATAGATAGATAGATAGATATCATTGAAAGACAAAA CAGAGATGGATGATAGATAC	XIVKFZSKIMSNTGFYEINGIJOGBCEIOLTLUEEGEHOWJDBNKLZYJFMYQHE	10 XI	[GATA]10
TCCTCTTCCCTAGATCAATACAGACAGACAGACAGGTGGATAGATAGATAGATAGATAGACAGATAGATAGATAGATATCATTGAAAGACAAAA CAGAGATGGATGATAGATAC	ERZHTOJAVBFTVGHWKTWALRKAMDZPEKJSVYQQOVVYKGBYPWVVNOZLZJD	10 ER	[GATA]5 GACA [GATA]4

Stutter Artifact Discrimination

Allele-Specific Stutter

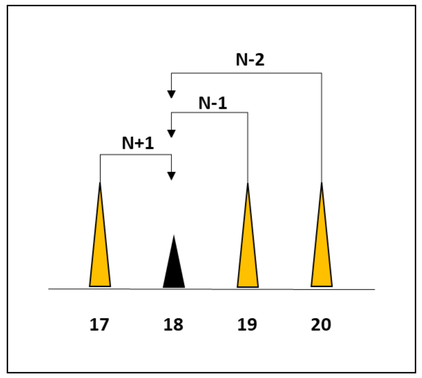

Figure 3. Illustration of stutter in a CE mixture scenario. Expected RFU intensity of stutter is dependent upon the alleles present in the profile that could plausibly contribute to stutter; and on a model of RFU intensity.

Motif-Specific Stutter

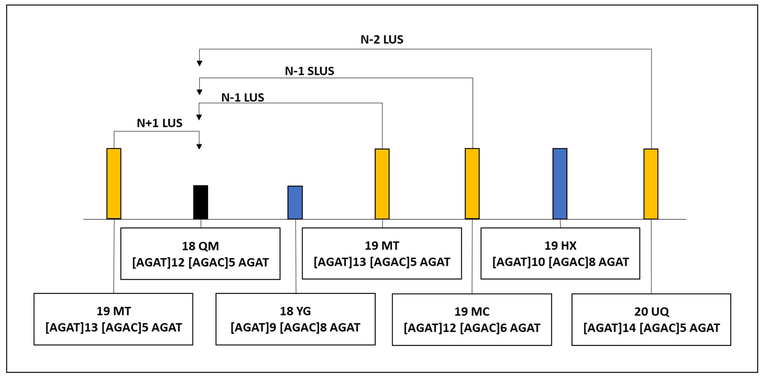

Figure 4. Illustration of stutter in a MPS mixture scenario. Expected read count intensity (RCI) of stutter is dependent upon the alleles present in the profile that could plausibly contribute to stutter (gold); and on a model of RCI. Major alleles present in the profile may not plausibly contribute to a candidate stutter (e.g., 19 HX). Minor alleles may be discerned from stutter when they are implausible stutter products (e.g., 18 YG). Minor alleles may also stack with stutter. Bracketed STR sequences are shown, but flanking sequence is omitted.

Non-Stutter Artifact Detection

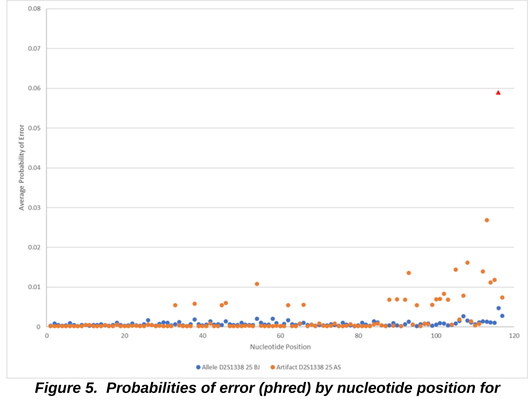

Figure 5. Probabilities of error (phred) by nucleotide position for sequence types at the D2S1338 locus. Data points represent the average probability of error over all sample reads in the haplotype allele 25 BJ (blue) or all sample reads of the artifact 25 SS (orange). An erroneous base call exists at the second to last position in the 25 SS reads (red triangle).

Read Pair	Base Calls and Q Scores									
Original Forward Read	G	A	T	C	A	C	A	G	G	T
	30	32	28	30	30	34	40	38	30	30
Original Reverse Read	G	A	T	C	T	C	A	G	G	T
	28	30	30	28	20	32	40	36	32	34
Forward Read	G	A	T	C	A	C	A	G	G	T
	30	32	28	30	30	34	40	38	30	30
Corrected Reverse Read	G	A	T	C	A	C	A	G	G	T
	28	30	30	28	30	32	40	36	32	34

Figure 6. Illustration of overlapping paired-end reads that differ in base call at a position (highlighted rectangle). Both forward and reverse reads are shown in forward orientation. When overlapping paired-end reads differ by sequence, they cannot both be correct (top panel). Quality scores can be used to infer which is the correct sequence. Optionally, the inferred incorrect base call can be corrected (bottom panel).

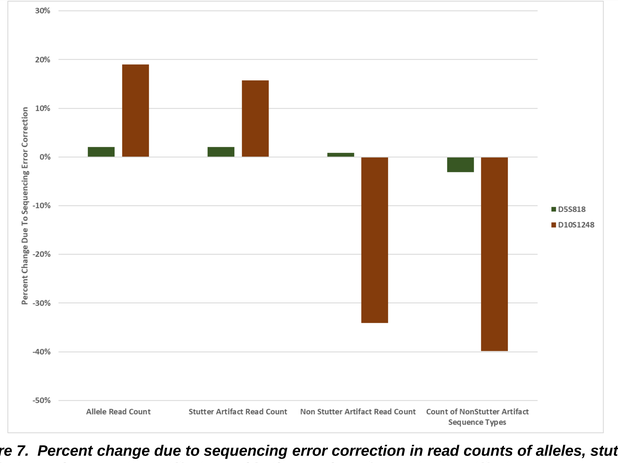

Figure 7. Percent change due to sequencing error correction in read counts of alleles, stutter artifacts, and non-stutter artifacts; and in the number of non-stutter artifact sequence types. Correction is more impactful for longer alleles. Percentages represent the average of three samples for two alleles: D5S818 (avg. allele length 65.3 nt) and D10S1248 (avg. allele length 109.3 nt).

(1) Sample H2_3P-C_10-30-60_1_R702-A508_S17. https://data.nist.gov/od/id/mds2-2157
(2) Young B, Faris T, Armogida L. A nomenclature for sequence-based forensic DNA analysis. Forensic Sci. Intl. Genet. 42(2019)14-20.

Funding partly from Army SBIR W911NF-22-C-0002

NicheVision Forensics, LLC