Correcting Sequencing Error in Forensic PCR-MPS Data

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

Introduction

Like forensic PCR-CE analysis, PCR-MPS analysis of STR markers generates both allele and artifact signals. PCR-MPS alleles are more complex as a result of being defined by the DNA sequence feature of the PCR amplicons as compared to the length feature used in PCR-CE analysis. PCR-MPS artifacts are also more complex. Multiple stutter artifacts can be generated from a single PCR-MPS allele. This is particularly true in compound and complex loci where stutter artifacts are commonly observed from both the longest unterupted stretch (LUS) of a repeat motif and the second longest uninterrupted stretch (SLUS). However, PCR-MPS stutter artifacts can be modeled using the same techniques that have long been used for PCR-CE methods.

Non-stutter artifacts in PCR-MPS are of a completely different nature than non-stutter artifacts in PCR-CE. All non-stutter artifacts arise from incorrect PCR amplicons, and none arise from other sources as does PCR-CE pullup and dye blobs. As alternate amplicon sequences, non-stutter artifacts may be confused with plausible alleles from minor contributors. Successful deconvolution of mixed PCR-MPS samples requires highly effective filters for non-stutter artifacts. Recovery of alleles from minor contributors can be improved by lowering the limit of detection (analytical threshold). However, increasing numbers of non-stutter artifacts are encountered as thresholds are lowered. Here, we use analytical thresholds equivalent to 0.75% of the per-locus read count. Non-stutter artifacts are thought to arise from library preparation, PCR misincorporation, and sequencing error (Figure 1).

The number of non-stutter artifacts observed in a profile is greatly affected by the techniques used to filter artifacts as well as the choice of sequencing platform and DNA preparation kit. Here we illustrate the impact of sequencing error correction, and machine learning algorithms for filtering non-stutter artifacts. All data were generated using the MiSeq sequencer, and ForenSeq™ DPMA (Verogen) and OmniSTR™ (NimaGen) kits.

Model of PCR-MPS Allele and Artifact Generation

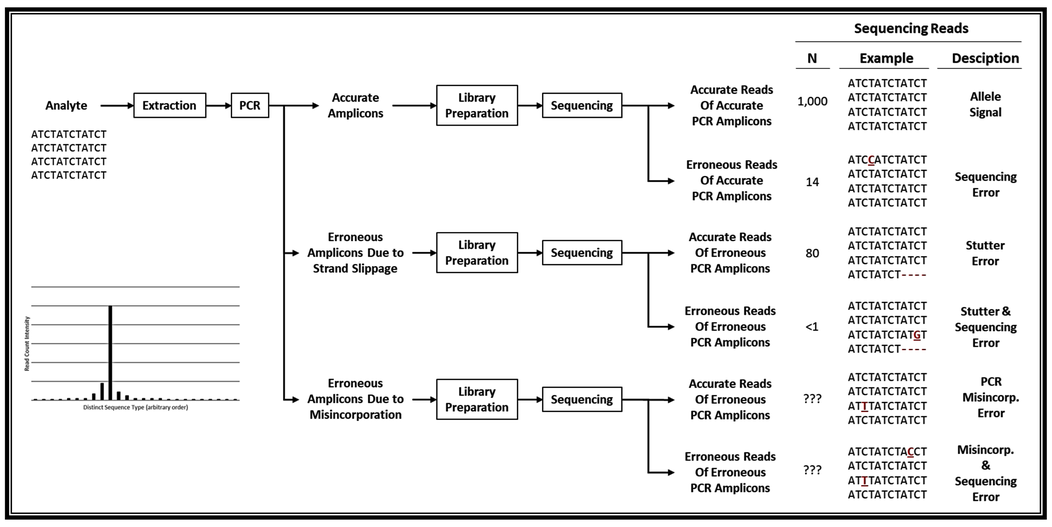

Figure 1. PCR-MPS is a hyphenated method involving repeated measurement of analytes. Allele signal arises from error-free analysis, while other distinct sequence types (DST) arise due to error.

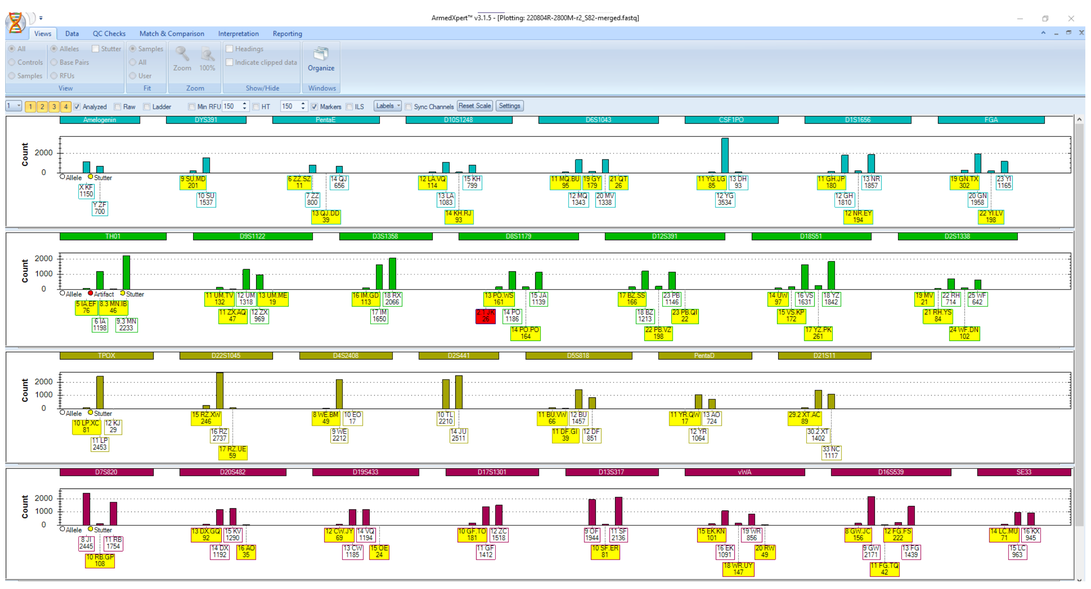

Figure 2. MixtureAce™ software (NicheVision) display of alleles and artifacts in arbitrarily assigned "color channels". Loci are arranged in increasing order of length. Alleles are identified by SID nomenclature (1), and stutter artifacts are attributed to parent alleles using "dot" notation (e.g., SE33 14 LC.MU). Data were generated on a MiSeq™ (Verogen) using the OmniSTR™ kit (NimaGen). Analytical threshold = 0.75%.

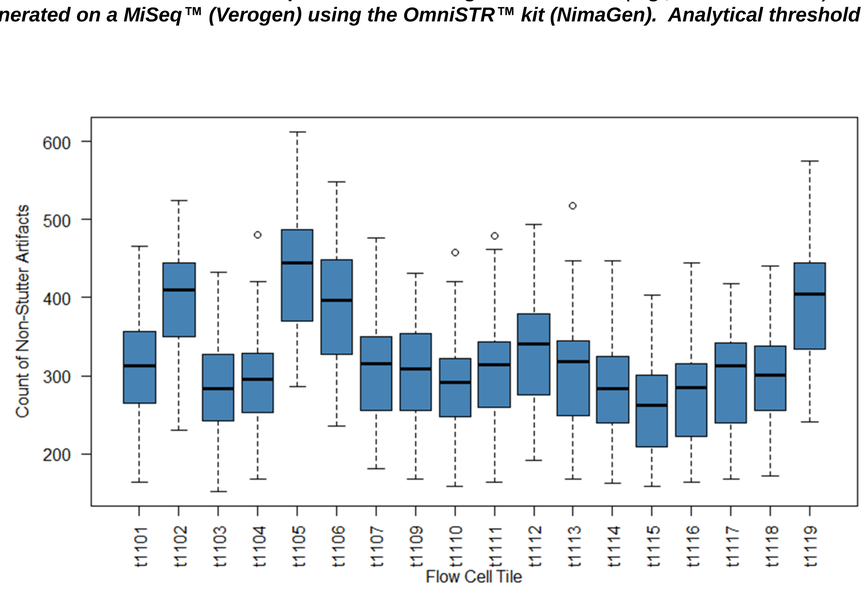

Figure 3. The count of non-stutter artifacts, including sequencing error can differ by flow-cell tile. Downselecting noisy tiles may improve signal-to-noise. Data represent non-stutter artifact counts across 47 single source samples sequenced using the ForenSeq™ DPMA kit (Verogen).

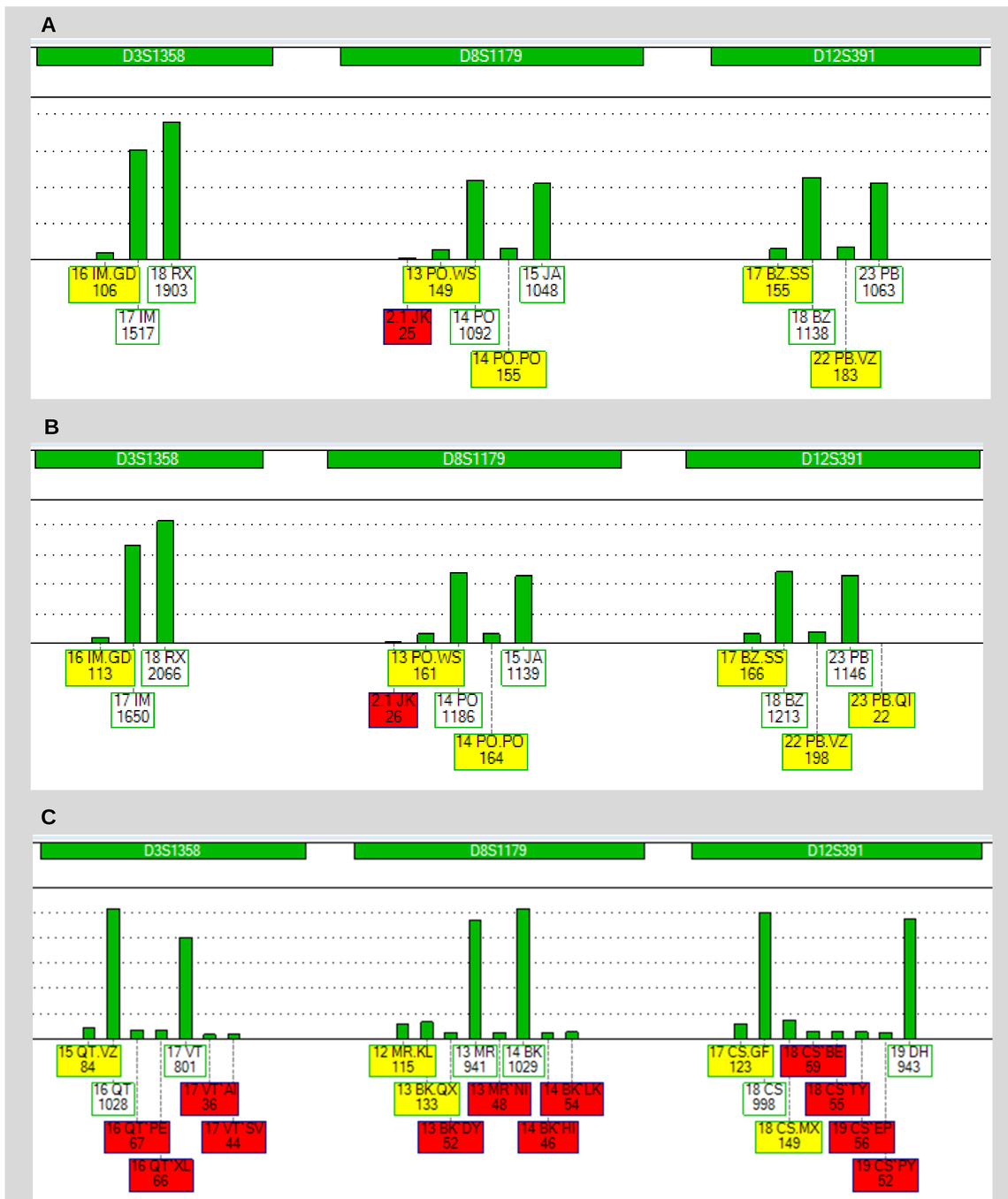

Figure 4. MixtureAce™ software displays of three loci from single source samples. All samples were analyzed with an analytical threshold of 0.75%. Both sequencing error correction and choice of kit impacts the level and type of non-stutter artifacts observed. A) NimaGen OmniSTR™ analysis of 2800M showing one non-stutter artifact arising from a large deletion in the amplicon. B) NimaGen OmniSTR™ analysis of 2800M after sequencing error correction. Allele read count intensities were increased about 8% due to error correction. C) Verogen ForenSeq™ DPMA kit analysis of NA17129 showing presence of non-stutter artifacts above threshold. The threshold was set at 0.75%, which is below the minimum Verogen recommendation, but used here for illustration purposes. Analysis using UAS software (Verogen) suppresses these artifacts.

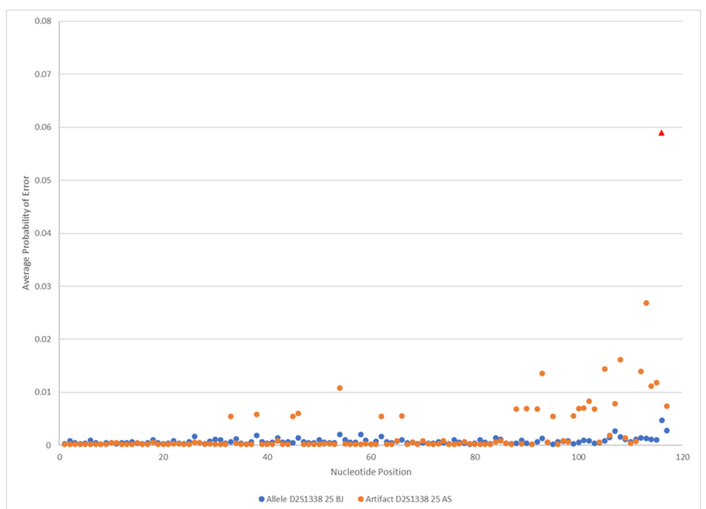

Figure 5. Probabilities of error (phred) by nucleotide position for sequence types at the D2S1338 locus in one sample. Data points represent the average probability of error over all sample reads in the haplotype allele 25 BJ (blue) or all sample reads of the artifact 25 SS (orange). An erroneous base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

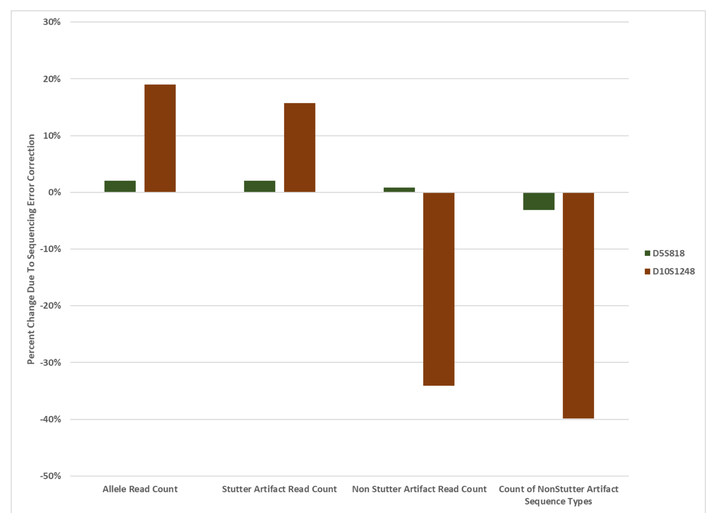

Figure 6. Percent change due to sequencing error correction in read counts of alleles, stutter artifacts, and non-stutter artifacts; and in the count of non-stutter artifact sequence types. Correction is more impactful for longer haplotype alleles. Percentages represent the average of three samples for two alleles: D5S818 (avg. allele length 65.3 nt) and D10S1248 (avg. allele length 109.3 nt).

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 7. 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).

(1) 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