Stutter Models for Sequence-Based Alleles in MixtureAce™ Software

Pete Delaune[1], David Diekema[1], Damani Johnson[2], Tim Kalafut[2], Brian Young[1,3,4]
[1]NicheVision Forensics, [2]Sam Houston State University, [3]Syracuse University, [4]Florida International University
Corresponding Author: brian@nichevision.com

Abstract

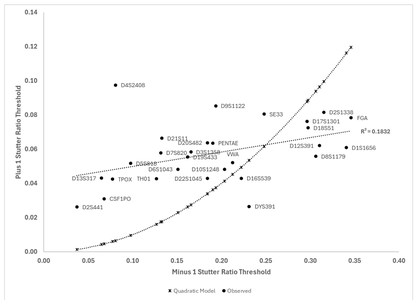
Stutter artifacts were characterized by i) stutter direction (minus and plus), ii) motif step change (one or two), iii) repeat structure (simple vs complex), iv) ordinal rank of the repeat motif (LUS versus SLUS; second longest uninterrupted stretch), and v) relative contribution of the LUS to the total number of repeat motifs in an STR allele. Each of these features was shown to improve ordinary least square (OLS) models for some loci as measured by total coefficient of determination (R^2).

Background

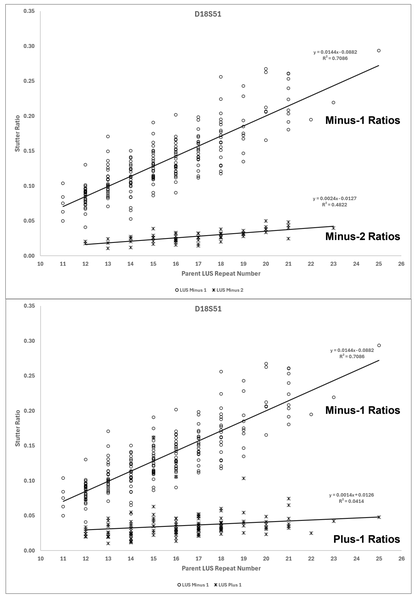
Stutter ratio (SR) models improve when using the longest uninterrupted stretch (LUS) rather than allele number as the independent variable [1]; when modeling at the allele level versus locus-level [2], or when using GLM models such as Beta regression [3]. The ability of massively parallel sequencing (MPS) to detect amplicon sequences facilitates further sequence-aware fine tuning of SR models [4].

MPS technology offers better sensitivity and better allele/artifact discernment than CE methods. However, stutter artifacts generated by MPS are more complicated to interpret and filter than artifacts generated by CE because a wider range of artifacts can be observed. Therefore, software for analyzing MPS data must be sequence-aware to be effective in filtering and displaying stutter and non-stutter artifacts during profile review. Many stutter artifacts revealed by MPS are likely present in CE profiles also, but are commonly unobserved because they are below threshold, or stack with other artifacts or alleles (e.g., N0 stutter) . In many cases, these newly-discernable artifacts form subpopulations that can be separately modeled.

-1, -2, +1 Step Models

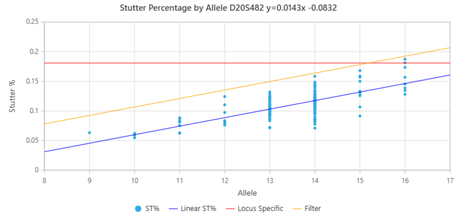


Predictions of plus-1 stutter thresholds based on minus-1 stutter thresholds via $y = x^2$ [5] are only approximately accurate. Minus-2 stutter shows a similar pattern, except that minus-2 stutter was unobserved at some loci.



Stutter ratios of minus-2 and plus-1 LUS stutter artifacts. Minus-2 and plus-1 stutters exhibit similar ratios at most loci and show low dependency on allele LUS. Data points represent instances where both stutter types were observed at the same allele.

Locus Level vs Allele Level Models



Locus and allele-level models for D20S482. Allele- and locus-level thresholds are 2 and 3 standard deviations from the locus mean and allele regression line respectively.

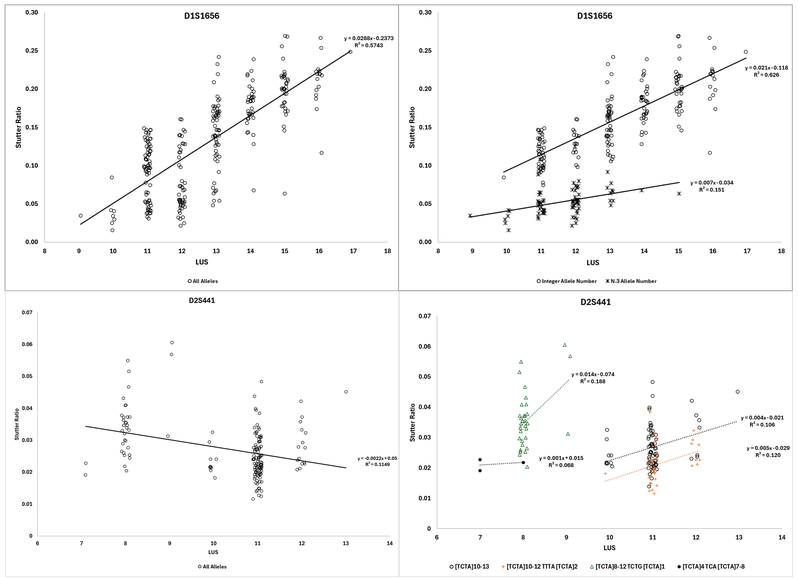
Materials and Methods

Regression models were developed using 175 single source samples from African American, Caucasian and Hispanic subpopulations along with standard DNAs 2800M and 2391d-C. Sequencing was performed using the OmniSTR kit (NimaGen™) and MiSeq™ instruments at Florida International University and NimaGen™ following vendor recommendations. **Data Analysis**: Sequencing error was filtered, and reads were classified into 12,833 distinct sequence types (DSTs), each with a read count intensity (RCI) using MixtureAce™ (NicheVision). Alleles and artifacts were identified using MPS Stutter Finder™ (NicheVision). All modeling was restricted to stutter artifacts that were unambiguously associated with a single "parent" allele. Stutter ratios were defined as SR = $RCI_{artifact}/RCI_{allele}$. **OLS Models**: The independent variables were LUS, or SLUS and the dependent variable was untransformed SR. Stutter calls were manually evaluated by subject matter experts using MixtureAce CSV reports and bar chart displays (i.e., MPS electropherogram equivalents).

References

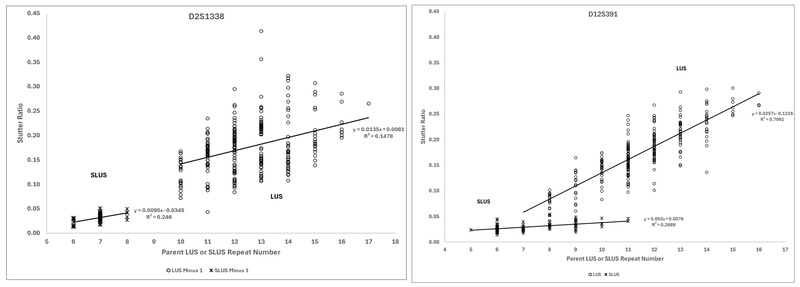
[1] Walsh et al. 1996 Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. Nucleic Acids Res. 24(14):2807-2812.
[2] Kalafut et al. 2018 Implementation and validation of an improved allele specific stutter filtering method for electropherogram interpretation. Forensic Sci Int Genet 35 (2018) 50-56.
[3] Agudo et al. 2022 A comprehensive comparison of MPS-STR stutter artefacts. Forensic Sci Int Genet https://doi.org/10.1016/j.fsigen.2022.102728
[4] Aponte et al 2015 Sequence-based analysis of stutter at STR loci: characterization and utility. http://dx.doi.org/10.1016/j.fsigss.2015.09.181
[5] Verogen Technical Note. Pub. No. VD2021030 September 2021
[6] Murdock et al. 2019 Definitions, methods, and applications in interpretable machine learning. www.pns.org/cgi/doi10.1073/pnas.1900654116.
[7] Agudo et al. 2024 A comparison of likelihood ratios calculated from surface DNA mixtures using MPS and CE technologies. Forensic Sci Genet https://doi.org/10.1016/j.fsigen.2024.103111.
[8] Young et al 2025 Manuscript in preparation.

Modeling Allele Structure Subpopulations



Studies using capillary electrophoresis have noted a non-linear minus-1 stutter for some loci, especially D2S441 where non-linearity drives a negative slope for regression of all stutter ratios vs LUS. Considering allele composition can improve models. D1S1656: Total R^2 increased from 0.12 to 0.48. D2S441: Total R^2 increased from 0.57 to 0.78 and regression slopes are positive.

LUS and SLUS Models

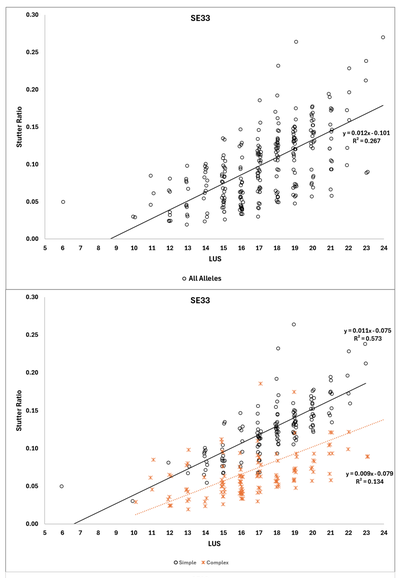


SRs of LUS and SLUS motifs modeled relative to parent allele LUS or SLUS repeat count for D2S1338 (left) and D12S391 (right). Data points represent instances where both stutter types were observed at the same locus.

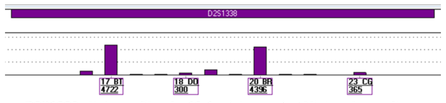
Discussion

The statistical properties of OLS estimators are suboptimal when used with non-normal (untransformed) ratio data. However, simple OLS models have the advantage of explainability [6] when interpreting and defending forensic profiles. When applied to allele subpopulations on a sequence-aware basis, OLS models can improve the accuracy of allele and artifact determinations in forensic profiles.

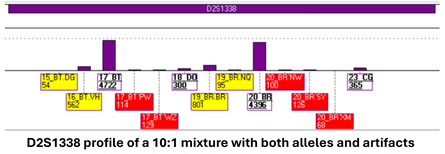
Final evidence about mixed profiles presented in court is likely to come from probabilistic genotyping products such as STRmix NGS (ESR). Failure to identify and remove unexplained artifacts can degrade LRs generated by probabilistic genotyping [7]. Additionally, examiners still have a need to review profiles for suitability and NOC estimation. Examiners must be prepared to understand and defend individual allele/artifact calls that support inclusions or exclusions. Graphic displays such as bar charts with labeled alleles and artifacts can support examiners but crucially depend upon accurate allele/artifact determinations [8].



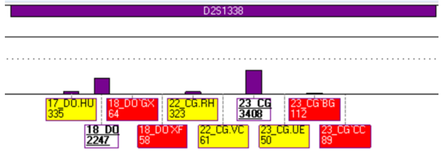
SE33: Total R^2 increased from 0.27 (top) to 0.71 (middle) when considering allele composition. Further subdividing complex alleles by the difference between LUS and SLUS repeat counts may further improve models (bottom), but overfitting is possible.



D2S1338 allele profile of a 10:1 mixture using MixtureAce software (NicheVision). Alleles are labeled on a sequence basis using SID nomenclature comprised of allele numbers and SID letters (NN_LL).



D2S1338 profile of a 10:1 mixture with both alleles and artifacts marked using MixtureAce. Stutter and non-stutter "child" artifacts are attributed to "parent" alleles using a dot and tic convention (i.e., PP.CC and PP`CC signifying parent-child pairs). Sequencing errors were intentionally uncorrected to show those artifacts.



D2S1338 profile of the minor contributor to the 10:1 mixture with both alleles and artifacts marked using MixtureAce.