Limits of Detection In Forensic DNA Analysis Using MPS

Brian Young[1], Ariana Harrison[2], Karin Crenshaw[2]
[1]NicheVision Forensics; [2]Broward County Sheriff's Office
Corresponding Author: brian@nichevision.com

Abstract

We report methods for calculating lower limits of detection (LOD, aka analytical thresholds) for sequence based STR alleles using positive control data. The methods presented here contrast with previously published methods based on negative control data. Thresholds based on positive controls focus attention on artifacts most likely to be misinterpreted as alleles in forensic samples. Thresholds based on negative controls are a function of contamination level, which ideally is zero.

Four thresholding approaches are presented:

- Mean plus ***k*** times the standard deviation of noise applied to either negative or positive controls,
- Tail probabilities of the zero inflated negative binomial applied to negative controls,
- Tail probabilities by Chebyshev's inequality applied to positive controls,
- Empirical intervals between the most intense artifact, and the least intense allele applied to positive controls.

Signal and Noise in PCR-MPS Methods

PCR-MPS methods do not express "baseline noise" familiar to PCR-CE methods. All instrument responses are sequencing reads arising from DNA entering the method (Figure 1).

- **INFORMATIVE SIGNAL:** sequencing reads that correspond to intended DNA template and back one LUS stutter of DNA template.
  - Back one LUS stutter is included because it is modeled in probabilistic genotyping.
- **NOISE:** aka uninformative signal, is interpreted differently for negative and positive controls.
  - Noise in negative controls: reads from contamination.
  - Noise in positive controls: reads from artifacts.

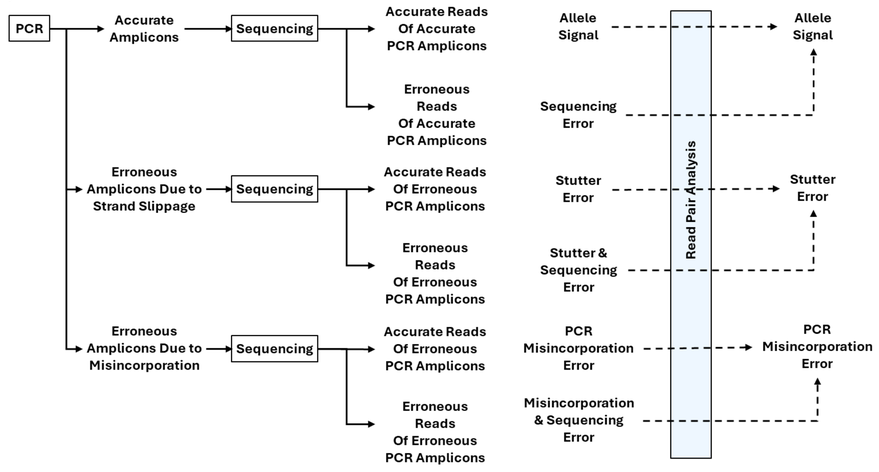

Figure 1. **Illustration of the origin of allele signal and artifacts in STR analysis by PCR-MPS.**

LODs Based on Negative Controls

- Published methods apply Equation 1 to read counts within sample-locus bins [1]. Here, 442 reads were observed from 29 STR loci across 37 samples. LODs ranged from 5 to 22 reads.
- Contamination reads were not normally distributed. An alternative interpretation of the data is that reads are generated by a Poisson process and the sample-locus bins are intervals within which zero or more read events may occur. Under this interpretation, a Poisson distribution may apply, but a mixture model combining a zero-generating component, and the negative binomial distribution fits the data better (Equation 2).
- LODs based on negative controls are sensitive to levels of contamination. Zero contamination results in an LOD of 0, while high contamination results in high LODs. Contamination is stochastic. Thus, LODs based on negative controls are logically specific to a particular sample preparation and sequencing run.
- Outliers can have dramatic effects on the standard deviation and thus on the LOD (Table 1).

Table 1. LOD values calculated by the standard approach, and by the 99.7% upper limit of a zero-inflated negative binomial distribution. Removing the single highest contaminating sequence reduces the LOD (Truncated).

Method	LOD (Reads)		LOD %	
	All Sequences	Truncated	All Sequences	Truncated
Standard Approach: $\mu + 3\sigma$	22	11	3.4%	1.7%
ZINB 99.7% CI Upper Limit	5	<5	0.7%	<0.7%

LODs Based on Positive Controls

- LODs based on positive controls varied depending on the method used and whether the LOD was calculated on a profile-wide or locus-specific basis (Table 2).
- Noise in positive control data consists of method artifacts (Table 3). LODs based on positive controls establish the false positive (aka drop-in) rate due to artifacts.
- LODs based on artifacts in positive controls are characteristic of the kit/method. When expressed in percentage terms they robust to sequencing depth of coverage (data not shown).
- Due to non-normality, tail probabilities by Chebyshev's inequality (Equation 3) are more defensible than by the standard method (Equation 1).
- The empirical interval method provides a window within which reasonable LODs can be set.

Table 2. LODs were calculated from noise in ten positive control samples by equations 1, 3, and 4 on profile-wide and per-locus bases. LODs assigned using Equation 1 have unknown tail proportions (false positive rates) because the data are not normally distributed. The false positive rate can be estimated for LODs assigned using Equation 2. The LOD assigned by the interval method (Equation 3) is the level of the highest noise type found in the training set.

Method	Training Set of 10 Samples			
	Profile-Wide AT		Locus-Specific AT	
	RCI	Pct.	RCI	Pct.
Eq. 1; k=3	18	0.6%	6 – 30	0.2 - 1.6%
Eq. 1; k=10	52	1.8%	14 – 89	0.5 – 4.4%
Eq. 3; 5% tail	29	1.0%	5 - 38	0.2 – 1.8%
Eq. 3; 1% tail	61	2.1%	12 - 84	0.4 – 4.1%
Interval	90	3.1%	16 - 90	0.6 – 3.0%

Table 3. Counts of the most intense noise types found across ten positive control samples. SLUS refers to the second longest uninterrupted stretch of tandem repeats. SE refers to non-stutter artifacts including sequencing error and PCR-misincorporation error. Motif swap refers to artifacts in which apparent reciprocal stutter occurs in tandem repeats that are immediately adjacent, and which differ by a single letter. A single base change can also explain these artifacts.

Artifact Category	Count	Percentage
Forward Stutter	180	63%
Homopolymer Stutter	4	1.4%
Back Stutter	51	18%
Sequence Error Artifacts SLUS Stutter Motif Swap Artifacts	49	18%
Total	**284**	**100%**

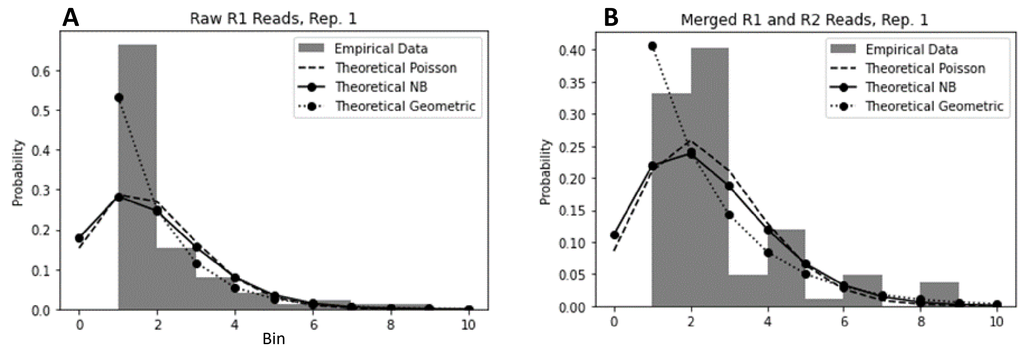

Figure 2. Low level noise in positive controls is not Gaussian and does not fit other distributions well. A) Distributions of uncorrected reads. B) Distributions after sequencing error correction. Frequencies of even-numbered bins increase due to the presence of PCR misincorporation errors present on both R1 and R2 reads.

Discussion and Conclusions

- LODs based on contamination in negative controls are characteristic of the sample preparation and sequencing run and may not apply well to other runs with a different level of contamination. A weakness of negative control-based thresholds is that perfect contamination control, which is desirable, will result in LOD = 0.
- LODs based on artifacts in positive controls, are characteristic of the kit/method and will apply to additional runs of that kit/method to the extent that artifact (e.g., forward stutter) intensities are consistent across runs. We conclude that LODs based on positive controls are more appropriate.
- The read count intensity of both alleles and artifacts increase with increasing depth of coverage (DOC). DOC depends upon the capacity of the sequencer, the number of samples, and the number of contributors to samples among other factors. These factors are difficult to hold constant in forensic laboratories. Thus, percentage-based dynamic thresholds that are robust to changes in DOC have wider applicability than count-based thresholds.
- The distributional properties of noise in both negative and positive controls are not well understood, making it problematic to apply usual parametric approaches. Chebyshev's inequality is useful because it applies to any distribution if the mean and variance are defined.
- The non-parametric interval method provides an empirical interval within which reasonable LODs can be set (Figure 3).
- The choice of LOD depends in part on the ability of available genotyping software to accurately filter artifacts, and on the examiner's tolerance for curating false positives.

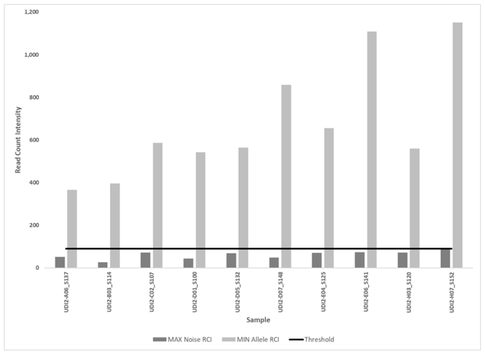

Figure 3. Read count intensities (RCI) of selected distinct sequence types (DST) at each of ten samples. Dark grey: the maximum RCI of noise DSTs. Light grey: the minimum RCI of allelic DSTs. A threshold of 90 RCI is indicated by the black line.

Materials and Methods

37 negative template control and 10 positive control (single source) samples were amplified and sequenced at BCSO using a MiSeq sequencer. One test sample was amplified at NimaGen in triplicate and split for shallow and deep sequencing using a MiSeq™ V3 and NextSeq™ 1000 P1 sequencer respectively. All sequencing was 2x300. The IDseek® OmniSTR™ kit (NimaGen, Nijmegen The Netherlands) was used for all amplification and MixtureAce™ software (NicheVision, Akron) was used for all genotyping analysis. R and Python were used for statistical analysis. All samples were collected under informed consent consistent with BCSO and NimaGen policies. LODs were calculated by three parametric methods and one empirical method: i) a standard approach [2] based on the mean and standard deviation of noise (Equation 1), ii) tail probabilities of the zero-inflated negative binomial distribution (Equation 2) [3], iii) Tail probabilities using Chebyshev's inequality (Equation 3) [4], and iv) the empirical "interval method" consisting of finding the range between the most intense noise sequence type and the least intense allele sequence type (Equation 4).

$$(1)\quad \mathrm{LOD} = \mu_b + k\sigma_b$$

$$(2)\quad f(x) = \begin{cases} p + (1-p)NB(0;\mu,r) & x = 0 \\ (1-p)NB(x;\mu,r) & x > 0 \end{cases}$$

$$(3)\quad Pr(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

$$(4)\quad (\min(\text{allele}) \underset{interval}{\Longleftrightarrow} \max(noise))$$

References

[1] K.M. Stephens, R. Barta, K. Fleming, J.C. Perez, S.F. Wu, J. Snedecor, C.L. Holt, B. LaRue, B. Budowle, Developmental validation of the ForenSeq MainstAY kit, MiSeq FGx sequencing system and ForenSeq Universal Analysis Software, Forensic Sci Int Genet 64 (2023) 102851.
[2] J. Bregu, D. Conklin, E. Coronado, M. Terrill, R.W. Cotton, C.M. Grgicak, Analytical thresholds and sensitivity: establishing RFU thresholds for forensic DNA analysis, J Forensic Sci 58(1) (2013) 120-9.
[3] J.M. Hilbe, Negative binomial regression, 2nd ed., Cambridge University Press, Cambridge, UK ; New York, 2011.
[4] P. Billingsley, Probability and measure, Anniversary ed., Wiley, Hoboken, N.J., 2012.