Analytical Thresholds In Forensic DNA Analysis Using MPS

Brian Young, PhD[1,2], Ariana Harrison, MS[3], Karin Crenshaw, PhD[3]
[1]NicheVision Forensics; [2]Florida International University, [3]Broward County Sheriff's Office
Corresponding Author: brian@nichevision.com

Abstract

We report methods for calculating analytical thresholds (AT) 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.
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 distribution 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.

Analytical Thresholds 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.
- Contamination reads were not normally distributed. A mixture model combining a zero-generating component, and the negative binomial distribution fits the data better (Equation 2).
- Outliers can have dramatic effects on the standard deviation and thus on the AT (Table 1).

Table 1. AT 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 AT (Truncated).

Method	AT (Reads)		AT %	
	All Sequences	Truncated	All Sequences	Truncated
Standard Approach: $\mu + 3\sigma$	18	11	3.4%	1.7%
ZINB 99.7% CI Upper Limit	5	<5	0.7%	<0.7%

Conclusions

- PCR-MPS methods do not express "baseline noise" familiar in PCR-CE methods. Instrument responses are discrete (countable) sequencing reads arising from DNA entering the method.
- The distributional properties of noise in both negative and positive controls are not well understood, making it problematic to apply usual parametric approaches appropriate to Normally distributed continuous data.
- The non-parametric interval method provides an empirical interval within which reasonable LODs can be set (Figure 2).
- Expressing thresholds as percentages renders them robust to changes in depth of sequencing.

Materials and Methods

37 negative template control and 10 positive control (single source) samples were amplified and sequenced at BCSO using a MiSeq sequencer and a standard flow cell. All sequencing was 2x300. The IDseek® OmniSTR™ kit (NimaGen, Nijmegen The Netherlands) was used for all amplification and MixtureAce™ software (NicheVision, Akron Ohio) 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) an empirical "interval method" consisting of finding the range between the most intense noise sequence type and the least intense allele sequence type (Equation 4).

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

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

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

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

Signal and Noise in PCR-MPS Methods

PCR-MPS methods do not express "baseline noise" familiar in PCR-CE methods. Instrument responses are discrete (countable) sequencing reads arising from DNA. Background fluorescence is not a factor in read counting.

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

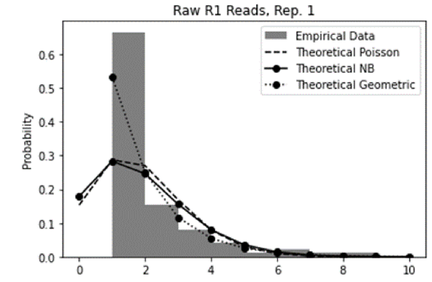

Figure 1. Histogram of low-level noise (RCI ≤ 10) in positive controls

Analytical Thresholds Based on Positive Controls

- ATs based on positive controls varied depending on the method used and whether the AT was calculated on a profile-wide or locus-specific basis (Table 2).
- Noise in positive control data consists of method artifacts (Table 3). ATs based on positive controls can help establish the false positive (aka drop-in) rate due to artifacts.
- Chebyshev's inequality (Equation 3) provides defensible tail probabilities for non-Normal data.
- The empirical interval method provides a window within which reasonable LODs can be set.

Table 2. Analytical Thresholds calculated from noise in ten positive control samples by equations 1, 3, and 4.

Method	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.

Artifact Category	Count	Percentage
Forward Stutter	180	63%
Homopolymer Stutter	4	1.4%
Back Stutter	51	18%
Sequence Error Artifacts Second Longest Stretch Stutter N0 Stutter	49	18%
Total	**284**	**100%**

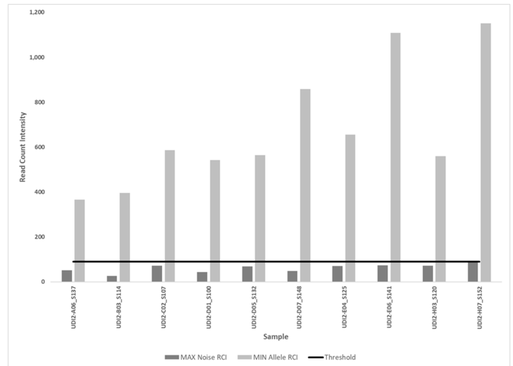

Figure 2. Read count intensities (RCI) of selected distinct sequence types (DST) at each of ten samples. Dark grey: the maximum RCI of noise DSTs. Light grey: the minimum RCI of allelic DSTs. A threshold of 90 RCI is indicated by the black line.

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.