Forensic Science International: Genetics Supplement Series 8 (2022) 88–90

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

journal homepage: www.elsevier.com/locate/fsigss

Calculation and implementation of sample-wide stochastic thresholds for forensic genetic analysis of STRs and SNPs for massively parallel sequencing platforms

Kathryn Stephens *, June Snedecor, Bruce Budowle [1]

Verogen, Inc., San Diego, CA 92121, USA

ARTICLE INFO

Keywords:
Next generation sequencing
Stochastic threshold
STR

ABSTRACT

Capillary electrophoresis (CE) analysis of short tandem repeats (STRs) and single nucleotide polymorphisms (SNPs) use a stochastic threshold to consider the possibility of missing alleles (dropouts) or detecting additional alleles (drop-ins). In CE, this threshold may be approximately 200 RFU, and peak heights are assessed relative to this threshold. In next generation sequencing (NGS), also known as massively parallel sequencing (MPS), STRs are identified by their sequence, and specific alleles are identified by their repeat number and intra-allelic variation. Abundance is approximated by the number of sequence reads for each allele. The total number of reads generated for each marker in a sample depends on factors such as the numbers of samples pooled for sequencing, the number of markers in the assay, the integrity and quantity of the input DNA sample, and the inter-locus balance of the assay. For multiplexes that contain both autosomal and sex-linked markers, the biological sex of the sample also influences total reads per locus. To normalize these variables and better establish a robust stochastic threshold, a sample-wide metric is proposed for estimating the possibility of dropouts or drop-ins based on the variance of the inter-locus balance of the markers across a sample. The intuition is that samples with variable allele balance globally are more likely to have noisier data and therefore require more stringent read count thresholds. This method is robust to sequencing multiplexity, biological sex and manufacturing lot variation.

1. Introduction

Stochastic effects during PCR amplification are evident when DNA inputs are low or DNA is of poor quality. Stochastic thresholds are set for analysis of DNA for human identification to indicate the potential for missing data in compromised samples. The stochastic thresholds for analysis of STRs using CE are understood to be the peak height at which there is a reasonable to high probability of a missing allele. The advent of NGS for STR analysis in human identification has yielded several advantages such as recovery of data from degraded DNA samples, sequence information from repeats and flank regions and the ability to analyze more markers simultaneously from low amounts of DNA. However, setting a stochastic threshold has been more challenging for this technology. Since many variables can impact read counts, static thresholds on read counts like those used for RFUs in CE are not appropriate. The peak heights for alleles for CE are analogous to read numbers for alleles for NGS, but read numbers depend on the number of samples in a library pool (multiplexity), biological sex of the DNA contributors, DNA input and DNA quality/integrity. In this communication, the potential for use of the %CV of inter-locus balance for reads per locus as a stochastic threshold is demonstrated. When the %CV of a sample is above the stochastic threshold, there is imbalance across all loci, which indicates potential missing data: any homozygous locus genotype could be heterozygous with a missing allele.

2. Materials and methods

Libraries were prepared using either the ForenSeq DNA Signature Prep Kit with DNA Primer Mix B (DPMB) or the ForenSeq MainstAY kit

* Corresponding author.
E-mail address: kstephens@verogen.com (K. Stephens).
[1] University of Helsinki, Department of Forensic Medicine, Haartmaninkatu 8, P.O. Box 63, Helsinki 00014, Finland.

https://doi.org/10.1016/j.fsigss.2022.09.032
Received 19 September 2022; Accepted 28 September 2022
Available online 29 September 2022

following the manufacturers reference guides [1,2]. Control DNAs 2800 M (Promega Corporation, Madison, WI, USA), 9948 (MCLAB, South San Francisco, CA, USA), NA24385 and NA12878 (Coriell Institute for Medical Research, Camden, NJ, USA), and SRM 2391d sample B (NIST, Gaithersburg, MD, USA) were used with inputs ranging from 8 pg to 4 ng. A degraded DNA series was purchased from InnoGenomics Technologies (New Orleans, LA, USA). DNA quantity and degradation index were determined using InnoQuant HY real-time qPCR kit (InnoGenomics Technologies, New Orleans, LA, USA) [3,4]. Indigo, hematin, tannic acid, and humic acid PCR inhibitors were purchased from SIGMA-Aldrich (St. Louis, MO, USA) and added at various concentrations to the DNA: indigo (133 μM), hematin (20–50 μM), tannic acid (0.5–4 μM), and humic acid (0.125–2.50 ng/μL) (manuscript in preparation). ForenSeq DNA Signature Prep Kit DPMB libraries were sequenced at plexities of 32 (recommended) or 8, and the ForenSeq MainstAY libraries were sequenced at plexities of 96 (recommended), 66 or 33 libraries, using either the MiSeq FGx Reagent Kit or MiSeq FGx Reagent Micro Kit following the MiSeq FGx reference guide [5]. Sequencing runs were analyzed using the default analysis method for each library preparation kit: the Universal Analysis Software v1.3 (DNA Signature) or v2.4 (MainstAY) [6,7]. Project reports were exported from the software to analyze autosomal and Y-STR coverage and detected alleles for MainstAY and DNA Signature; and X-STR and identity-informative SNP (iiSNP) coverage for DNA Signature. To analyze the ancestry-informative SNPs (aiSNPs) and phenotype-informative SNPs (piSNPs) for DNA Signature, the phenotype reports were exported for each sample and analyzed for coverage and detected alleles. The global proportional inter-locus balance was calculated by determining the total number of typed reads for each locus and dividing by the total number of reads for the marker type (STR or SNP). The coefficient of variance (CV) for a sample was calculated by dividing the standard deviation of all typed reads per locus by the average of all the typed reads per locus for the sample.

$$\%CV = \frac{(Standard \quad Deviation)}{(Mean)} \times 100\%$$

3. Results and discussion

Inter-locus balance is consistent for STR and SNP markers with ForenSeq DNA Signature Prep Kit (DNA Signature) [8] and STR markers with ForenSeq MainstAY kit (MainstAY) (manuscript in preparation) across high quality and high input samples. At lower inputs or with low quality DNA samples, stochastic effects are increased with inter-locus balance worsening and becoming variable with alleles dropping below the Analytical Threshold (AT). A metric for inter-locus balance across multiple loci is the Coefficient of Variance or %CV for the typed reads per locus. %CV normalizes the standard deviation of a set of data by the mean of that set of data since standard deviation scales with the magnitude of the values. %CV is calculated across reads per locus for all loci in the library preparation kits as demonstrated in Supplemental Fig. 1. To test %CV of reads per locus (inter-locus balance) as an indicator of stochastic effects, the %CV for inter-locus balance was calculated across replicate samples in different Sensitivity Studies for all loci. The sample replicates were run with different DNA samples of different biological sexes and were processed by multiple operators for DNA Signature and MainstAY (data not shown). %CV is stable across DNA inputs above 125 pg. %CV increases once DNA input drops below 125 pg for DNA Signature and below 62 pg for MainstAY.

To test whether the %CV could be used to estimate dropouts of alleles or loci, %CV was correlated to the number of alleles that dropped below the AT for the sensitivity studies. The alleles below AT begin to increase under 125 pg with %CV values above 110% for ForenSeq DNA Signature and under 62 pg with %CV values above 65% for ForenSeq MainstAY. These %CV values were tested for use as Stochastic Thresholds for DNA Signature and MainstAY samples in the following studies.

Stochastic Thresholds of 110% for DNA Signature and 65% for MainstAY correlated with loss of alleles across sensitivity studies performed by multiple operators for both male and female DNA samples sequenced at the recommended plexities. To test the robustness of these Stochastic Thresholds across sequencing plexities, Sensitivity Studies for MainstAY (8 pg – 4 ng) sequenced with the 66-plex and 33-plex and DNA Signature (8 pg – 1 ng) sequenced at an 8-plex were analyzed for %CV and allele detection. The thresholds are robust for both DNA Signature (110%) and MainstAY (65%) at these multiplexities with alleles dropping below the AT when the %CV of inter-locus balance was above the thresholds.

To test the %CV threshold for low quality samples, %CV was calculated for degraded and inhibited samples and compared to recovery of expected alleles. The results show that the %CV of coverage indicates that alleles are missing at higher %CV with degraded and inhibited samples and that the stochastic thresholds of 110% and 65% for DNA Signature and MainstAY, respectively, can be operationally functional (Supplemental Fig. 2). The %CV for inter-locus balance for typed reads across the loci in ForenSeq DNA Signature Prep Kit and ForenSeq MainstAY kit seems to function well as an indicator of stochastic effects in low input, degraded or inhibited samples.

%CV of inter-locus balance was calculated for DNA mixtures, male: male, male:female, and female:male with 1 ng input of high quality DNA samples and ratios from 1:1 to 1:99. %CVs were all below the Stochastic Threshold (not shown) indicating that a per locus stochastic threshold will be required for assessing minor contributor alleles in DNA mixture samples with high input or high quality DNA.

Acknowledgments

The authors would like to thank Juan Carlos Perez, Richelle Barta, Keenan Fleming, and Michaela Russo for preparing libraries and performing sequencing runs for this study and Joana Antunes for reviewing this manuscript.

Conflict of interest

Kathryn Stephens and June Snedecor are current employees of Verogen, Inc. where the experiments were planned, performed and analyzed. Bruce Budowle is a consultant for Verogen, Inc.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigss.2022.09.032.

References

[1] Verogen, ForenSeq MainstAY Kit Reference Guide. ⟨https://verogen.com/wp-content/uploads/2022/01/forenseq-mainstay-reference-guide-PCR1-vd2020050-c.pdf⟩, (Accessed September 2022).
[2] Verogen, ForenSeq DNA Signature Prep Reference Guide. ⟨https://verogen.com/wp-content/uploads/2022/01/forenseq-dna-signature-prep-reference-guide-PCR1-vd2018005-d.pdf⟩, (Accessed 2021).
[3] P. Carrasco, et al., Optimizing DNA recovery and forensic typing of degraded blood and dental remains using a specialized extraction method, comprehensive qPCR sample characterization, and massively parallel sequencing, Int. J. Leg. Med. 134 (1) (2020) 79–91.
[4] InnoGenomics, InnoQuant HY Human and Male DNA Quantification and Degration Assessment Kit Using 7500 Real-time PCR System User Guide v1.5. ⟨https://innogenomics.com/wp-content/uploads/files/InnoQuant_HY_Using_7500_Real_Time_PCR_System_User_Guide_v1_5.pdf⟩, (Accessed 2021).
[5] Verogen, MiSeq FGx Sequencing System Reference Guide. Revision F, 2021. ⟨https://verogen.com/wp-content/uploads/2021/02/miseq-fgx-system-reference-guide-vd2018006-f.pdf⟩, (Accessed February 2021).

[6] Verogen, Universal Analysis Software MainstAY Product Line Module Version 2 Reference Guide. Revision A, 2022. ⟨https://verogen.com/wp-content/uploads/2022/08/universal-analysis-software-v2-reference-guide-mainstAY-VD2022001-RevA.pdf⟩, (Accessed June 2022).
[7] Verogen, ForenSeq Universal Analysis Software Guide. ⟨https://verogen.com/wp-content/uploads/2018/08/ForenSeq-Univ-Analysis-SW-Guide-VD2018007-A.pdf⟩, (Accessed 2021).
[8] A.C. Jäger, et al., Developmental validation of the MiSeq FGx forensic genomics system for targeted next generation sequencing in forensic DNA casework and database laboratories, Forensic Sci. Int. Genet. 28 (2017) 52–70, https://doi.org/10.1016/j.fsigen.2017.01.011.