B110 The Use of Receiver Operating Characteristic (ROC) Curves as a Tool to Assess Noise and Zygosity in the Targeted Sequencing of Forensic Short Tandem Repeat (STR) Markers

*Sarah Riman, PhD**, National Institute of Standards and Technology, Gaithersburg, MD 20899; Hariharan Iyer, PhD, Gaithersburg, MD; Lisa Borsuk, MS, National Institute of Standards and Technology, Gaithersburg, MD 20899; Peter M. Vallone, PhD, National Institute of Standards and Technology, Gaithersburg, MD 20899-8314*

Learning Overview: After attending this presentation, attendees will understand how to characterize and understand noise, stutter artifacts, heterozygote imbalance, allelic drop-in, and allelic drop-out in Next Generation Sequencing (NGS) datasets generated from single-source samples.

Impact on the Forensic Science Community: This presentation will impact the forensic science community by showing a framework of statistical tools developed to systematically interpret and understand the characteristics of single-source DNA profiles generated by targeted sequencing.

The sequencing of STR markers provides additional information due to the underlying sequence variation that is typically masked by traditional fragment-based genotyping. The interpretation of STR profiles generated by targeted sequencing methods are susceptible to familiar parameters such as signal noise, stutter artifacts, heterozygote imbalance, and allelic drop-out/in, as well as additional factors introduced by the library preparation workflow.

In this work, data were generated from sensitivity studies using known single-source samples. The DNA extracts were amplified with the PowerSeq 46GY System Prototype with varying DNA target masses ranging from 15 pg to 500 pg. Amplified PCR products were subjected to library preparation using two different library preparation kits: Truseq DNA PCR-Free High Throughput (HT) Library Prep Kit (Illumina) and KAPA Hyper Prep Kit (KAPA Biosystems). Libraries were either normalized or left without normalization. Paired-end sequencing of the STR loci was then performed on the Illumina MiSeq platform, and raw FASTQ data files were analyzed using a modified version of the open source STRait Razor v2.0. The software identified the sequences, allele length, and coverage of the STR markers and regions at a minimum depth of coverage of 1X to capture as much data as possible. Receiver Operating Characteristic (ROC) curves were then used to understand the tradeoff between true positives (alleles) and false positives. False positives were attributed to drop-ins, stutter, and random noise. ROCs were also used to infer and examine zygosity using heterozygote balance (Hb) information to minimize the risks of misidentifying a heterozygote as a homozygote locus or a homozygote as heterozygote locus. Data generated from each library workflow were analyzed globally (all DNA quantities combined), as well as investigated per DNA quantity and per locus.

The aim of this presentation is to share the findings and show how analyses presented here can also be applied to sequence data generated by similar targeted sequence multiplexes and/or sequencing platforms.

Receiver Operating Characteristic, Noise, Next Generation Sequencing