THE USE OF RECEIVER OPERATING CHARACTERISTIC (ROC) CURVES AS A TOOL TO ASSESS NOISE AND ZYGOSITY IN TARGETED SEQUENCING OF FORENSIC STR MARKERS

Sarah Riman[1], Hari Iyer[2], Lisa A. Borsuk[1], Peter M. Vallone[1]
[1]National Institute of Standards and Technology, Applied Genetics Group
[2]National Institute of Standards and Technology, Statistical Design, Analysis, and Modeling Group

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 the same factors as for profiles generated using capillary electrophoresis. These factors include signal noise, stutter artifacts, heterozygote imbalance, and allelic drop-out/in. Our goal is to characterize and understand how these behave in targeted sequence datasets.

Here, we developed a framework using statistical tools to systematically interpret and understand the characteristics of single source DNA profiles generated by targeted sequencing. Data were generated from sensitivity studies using known single source samples amplified with the PowerSeq 46GY System Prototype with varying DNA target masses ranging from 15 pg to 500 pg. The STR loci were sequenced on the Illumina MiSeq platform and raw FASTQ data files were analyzed in STRait Razor without applying any thresholds (an i.e. at a coverage ≥ 1). Receiver Operating Characteristic (ROC) curves were then used to understand the tradeoff between true positives and false positives. False positives were attributed to drop-ins and noise (stutter and random causes). 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. These data were analyzed globally (all DNA quantities combined), as well as investigated per DNA quantity and per locus. Analyses presented can be applied to sequence data generated by similar targeted sequence multiplexes and/or NGS platforms.