CHARACTERIZATION OF STR ALLELOTYPING DATA FROM SECOND GENERATION SEQUENCING (SGS) WORKFLOWS

Esley Heizer, Mark Hester, Angela Minard-Smith, Dan Bornman, Brian Young, Battelle

SGS is being applied to forensic DNA analysis because of the informativeness provided by sequence-level data, and its capacity for exploiting multiple forensic markers such as STRs and ancestral SNPs in single assays. Recently, the feasibility of STR allelotyping by SGS has been demonstrated by several groups; however a critical analysis of all data generated by a SGS-based STR typing assay including the types of reads generated, sequence error characterization, and discerning allele abundance ratios has not been performed. In this study, we analyzed and characterized sequence data generated by PCR amplification of an 18-STR locus panel followed by SGS from 13 individuals and two standard reference materials (SRM) from NIST using the Illumina MiSeq platform. Only about 40% of the reads were informative for allelotyping. The remaining 60% of reads are uninformative due to error inherent to the analysis process. The major sources of error were PCR stutter, base substitution, including base substitution due to sequencing error and misclassification of PhiX controls. However, when considering only reads assignable to a specific STR locus, this analysis revealed that true allele reads were approximately 10X as abundant as stutter reads and 100X as abundant as non-stutter erroneous sequences. In addition, reads assignable to the true allele exhibited abundance ratios generally greater than 0.6, thereby confirming the suitability of SGS-based methods for STR allelotyping. The discriminatory power of allele sequencing was demonstrated through the observation of variant sequences. In four individuals, a locus was shown to be heterozygous by sequence, but homozygous by allele size; and 23 apparently novel sequence variants were found overall. In sum, because of the multitude of uninformative short-reads in SGS data, some of which are only subtly different, it is critical to develop high-fidelity allelotyping workflows that yield correct and deterministic results.