USE OF THE LUS TO FACILITATE PROBABILISTIC INTERPRETATION OF SEQUENCE-BASED STR DATA

R. Just[1,2], J. Irwin[1]
[1]DNA Support Unit, Federal Bureau of Investigation Laboratory
[2]Counterterrorism and Forensic Science Research Unit, Federal Bureau of Investigation Laboratory

Next generation sequencing (NGS) technologies continue to be investigated for forensic applications, and have the potential to improve typing of short tandem repeat (STR) loci in multiple respects. Some of the expected advantages of NGS for STR typing include enhanced mixture detection and genotype resolution via sequence variation among non-homologous alleles of the same length. However, at the same time that NGS methods for forensic DNA typing have advanced in recent years, many caseworking laboratories have implemented or are transitioning to probabilistic genotyping to assist the interpretation of complex autosomal STR typing results. Current probabilistic software programs are designed for data produced using length-based typing technologies, and do not accommodate sequence strings or other common shorthand notations of the sequence as the signal input. Yet to leverage the benefits of NGS for enhanced genotyping and mixture deconvolution, the sequence variation among same-length products must be utilized in some form.

To achieve this aim in the near term, we propose use of the longest uninterrupted stretch (LUS) in allele designations as a simple method to represent sequence variation within the STR repeat regions and enable probabilistic interpretation of sequence-based data in software programs similar to those that currently exist. An examination of published population data indicates that using repeat unit plus LUS can capture approximately 60-75% of the increase in the number of distinct alleles observed by length versus complete sequence information, and can represent greater than 80% of the total alleles detected by sequencing. Though using repeat unit plus LUS as the allele designator does not capture variation that occurs outside of the core repeat regions, the concept could serve as a valuable intermediate step towards the development of probabilistic genotyping based on complete STR sequence. The approach maintains both 1) numeric allele designations familiar to casework practitioners and 2) a clear relationship between parent alleles and their stutter products, while avoiding the algorithmic complexities that come with string based searches. This straightforward approach would permit a substantial portion of known STR sequence variation to be used for mixture deconvolution, resulting in more informative mixture statistics. Ultimately, the method could bridge the gap from current probabilistic systems to facilitate broader near-term adoption of NGS by forensic DNA testing laboratories.