

THE SEQUENCE IDENTIFIER (SID) SEQUENCE LABELING METHOD: AN OPERATIONAL NOMENCLATURE FOR MIXED-DNA CASEWORK USING MPS

Brian Young, Tom Faris, Luigi Armogida
NicheVision Forensics, LLC.

The use of MPS in casework has been hindered by the lack of a nomenclature that meets the operational needs of routine casework. The SID labeling system addresses this gap. PCR-MPS methods produce dozens to hundreds of distinct sequence types at each locus. In single-source samples, all non-allelic and non-(N-1)-stutter sequence types can be filtered without considering sequences using a suitably high analytical threshold (AT); and the widely used 1.5% threshold works well in this regard. However, in mixed casework samples, the number of contributors is unknown as are the proportion ratios. Thus, every distinct sequence type in a mixed PCR-MPS data set must be considered as a possible allele until assigned otherwise by typing rules. Consequently, typing/deconvolution software must label all sequence types in order to process them. Numerous nomenclature systems have been proposed, but all proposed systems are suitable for labeling alleles or stutter but not the sequencing errors that must be filtered from PCR-MPS mixtures. The SID nomenclature fills this role and offers a compact labeling system for computer analysis of mixed MPS casework data. SID is a SHA-256 hash value converted to a base-26 value comprised solely of capital letters from the English alphabet. When dynamically allocated within loci, unique and deterministic labeling of all sequences can be accomplished using typically only the first 2 letters of the full SID label; making the nomenclature as compact as the allele number nomenclature. For example, a fragment equivalent to STRSeq GenBank accession MK570033.1 (locus D1S1656) is labeled "15.3 JE", where the 2-letter SID is prepended with the CE-equivalent length. SID labels can be used in software systems to filter artifacts. For example, the ArmedXpert™ software system implements a dot connector (aka period) to denote a "parent.child" relationship between alleles and stutter artifacts. For example, the N-1 LUS stutter is labeled "15.3 JE.AC". A C->T sequencing error 18 bp in the upstream flank, results in the label "15.3 JE`WE" with a tick connector (aka grave accent) denoting a parent-child relationship between allele and sequence artifacts. Once labeled, alleles and artifacts can be compactly displayed and efficiently filtered. The SID labeling system is available for non-commercial use and can be implemented in any mixture analysis method that is able to accept character-based nomenclature for allele profiles.