PopSEQ – THE STR SEQUENCE DIVERSITY DATABASE

Seth A. Faith, NC State University

Massively parallel or "next-generation" sequencing technologies are undergoing validations in some forensic DNA laboratories. Should this technology be applied to short tandem repeat (STR) analysis, the population frequencies of sequence-based alleles will need to be known for statistical conclusions of forensic samples. To meet the growing needs of the criminal justice community we have built an open source website, PopSEQ, for STR sequence data and bioinformatics tools for self-service analytics. The current database, v1.0, represents autosomal- and Y-STR loci for over 1000 samples from 10 population groups. Data was collected using various next-generation sequencing platforms and extensively reviewed for quality. The interactive self-service tools allow users to login and view data (allele frequencies) determined by categorical identifiers of interest, such as sequencing kit, sample type, sex, loci, and geographic location. As previous reports have demonstrated, some STR loci in PopSEQ display high nucleotide sequence diversity in the core repeat, and may allow for opportunities in sample analysis beyond fragment size interpretations. The PopSEQ STR sequence diversity database is the first open source tool to readily provide high quality STR sequence allele frequency data for laboratories that will utilize next-generation sequencing methods in their workflows.