DNA Profiling of Database Reference Samples Using Second Generation Sequencing

Carey Davis and Bruce Budowle
Institute of Applied Genetics, Department of Forensic and Investigative Genetics, University of North Texas Health Science Center, Fort Worth, Texas, USA

Fourteen years ago, a core set of forensic markers was selected for the United States national databank, Combined DNA Index System (CODIS). This databank houses over 10,400,000 DNA reference profiles comprised of autosomal STRs from convicted felons and arrestees. These profiles have been used to develop many investigative leads for a variety of crimes. The database size continues to grow and additional search strategies have been considered. The expanded applications that CODIS has experienced over its 14 years warranted a reconsideration of whether the current core loci are sufficient. There is general agreement that the core STR markers for CODIS needs to be increased; but there are differences of opinion on what criteria and how best to proceed with core marker selection. While choosing a core set of markers is useful for formalizing a common set for data exchange, this concrete set inadvertently can limit progress and stifle innovation for alternate markers that may serve well the specialized forensic community needs. However, these discussions on a fixed core set of loci and unintentional stymied growth of novel marker sets can be rendered moot with the advent of second generation sequencing (SGS). With this technology, it is possible to analyze a large battery of forensically relevant genetic markers simultaneously with economies of scale once not thought possible. Furthermore, the high throughput capacity of NGS technology makes possible multiplexing of 12 to 384 individuals in a single reaction. Additional barcodes can increase sample throughput an order of magnitude. Instead of focusing only on a core set, a comprehensive (although small by SGS capabilities) panel of 31 autosomal STRs, 26 X STRs, 29 Y STRs, and 379 forensically relevant SNPs (both identity and bioancestry) was created using two separate sample preparation methods. Sample preparation (including library generation) is still a labor intensive process and this study attempts to reduce the front end labor component by comparing current approaches with the Haloplex (Agilent). Analysis was carried out on an Illumina GAIIx instrument which has a throughput in 50-100 gigabase range and far surpasses coverage requirements. This presentation describes the issues that confront STR and SNP typing by SGS which include library preparation, ease of use, read length, sampling of fragments encompassing intact alleles, and cost.