A UNIVERSAL ANALYSIS SOFTWARE FOR NEXT GENERATION SEQUENCING DATA: MODERN FORENSIC BIOINFORMATICS

John Walsh, Kirby Bloom, Jocelyne Bruand, Felix Schlesinger, Joe Varlaro, Steven Lee, Cydne Holt

Illumina Inc. San Diego, CA

One important step toward the implementation of NGS into routine forensic genomics analysis is the development of streamlined and user-friendly software. We describe here a complete sample to answer pipeline for simultaneous analysis of autosomal, X and Y STRs, identity, phenotypic and biogeographical ancestry informative SNPs. The ForenSeq Universal Analysis Software is part of the MiSeq FGx Forensic Genomics System, the first fully-validated NGS system specifically designed for use in forensic genomics applications.

The validated software solution provides a simple yet comprehensive program for analyzing, viewing, and reporting forensic genomics data for identification and investigative genetic leads. The software includes algorithms to accurately type >200 genetic markers, including autosomal, X, and Y STRs (including the 24 CODIS STRs), STR sequence variation from deep sequencing reads, multilocus genotyping of small, identity informative SNPs, estimate biogeographical ancestry, hair and eye color, and QC metrics. Default thresholds and filters for the limit of detection (analytical), stochastic (interpretation) and stutter (repeat slippage) were set based on extensive empirical testing, and can be customized based on a lab's internal validation.

Analysis is geared toward both single source DNA samples as well as challenging forensic samples. Algorithms will be described that perform demultiplexing, sequence alignment, STR and SNP genotyping, allele counting, and quality control indicators that help to identify DNA mixtures and genotype call quality. For example, each STR and SNP read is aligned to the hg19 human DNA reference sequence corresponding to the locus in the ForenSeq primer mix to determine locus and repeat length. After alignment, potential alleles are counted from read numbers for digital quantification. The STR sequence algorithm designates the STR allele repeat number, identifies sequence variants within STR repeats, and identifies stutter.

A multinomial logistic regression model estimates hair and eye color using phenotypic informative SNPs. Bio-geographical ancestry estimation is obtained by principal component analysis.

The ForenSeq Universal Analysis Software and its underlying bioinformatics pipeline facilitates streamlined, robust analysis of sequencing data for forensics genomics. It allows users to harness the additional power provided by sequencing and to retain backwards compatibility with established standards, nomenclature and databases created using CE-based STR typing.