Targeted Next-Generation Sequencing for Forensic Genomics

www.illumina.com/applications/forensics.ilmn

illumina®

Table of Contents

I. Targeted Next-Generation Sequencing for Forensic Genomics

DNA analysis has become the cornerstone of contemporary forensic science. Today, most forensic DNA testing utilizes PCR and capillary electrophoresis (CE)-based analysis methods to detect fragment length variation in short tandem repeat (STR) markers. A small percentage of forensic investigations also use CE-based Sanger sequencing to analyze specific regions of mitochondrial DNA (mtDNA).

Advances in genomic technologies have outpaced the methods that were first introduced into forensic DNA testing laboratories over 20 years ago. As the number of samples profiled has increased dramatically, the fundamentals of CE-based testing have remained relatively static. The advent of national DNA databases and the demonstration of their ability to aid investigations and identify suspects has encouraged more countries to establish criminal intelligence databases. As a result, the number of casework samples requiring DNA processing is steadily increasing. In parallel, the critical role DNA testing plays in the identification of missing persons, kinship testing, ancestry investigations, and other human identification applications continues to drive interest in new and more powerful analysis methods. This increasing demand strains the fixed capabilities of CE-based methods.

Illumina sequencing by synthesis (SBS) technology[1] offers a massively parallel approach for sequencing PCR amplicons and allows forensic scientists worldwide to harness the full power of targeted next-generation sequencing (NGS). With the highest yield of error-free reads, best performance in repetitive sequence regions, and lowest base-by-base price, Illumina SBS sequencing is the most widely adopted chemistry in the industry.[2-6] Using contemporary NGS systems, examiners can generate data that span the entire genome and address a wider range of questions in a single, targeted assay. Moreover, NGS-generated STR allele calls are fully compatible with current database formats, providing a seamless link between CE-based and NGS data. Early studies are providing a promising view of the advantages NGS brings to the analysis of even the smallest, most compromised, and highly mixed evidentiary samples.

II. The Basic NGS Amplicon Workflow

In principle, the concept behind NGS technology is similar to CE sequencing—DNA polymerase catalyzes the incorporation of fluorescently labeled deoxyribonucleotide triphosphates (dNTPs) into a DNA template strand during sequential cycles of DNA synthesis. During each cycle, at the point of incorporation, the nucleotides are identified by fluorophore excitation. One critical difference is that, rather than sequencing a single DNA fragment, NGS extends this process across millions of fragments in a massively parallel fashion.

While many NGS applications call for whole-genome sequencing (WGS), forensic genomic applications use a targeted sequencing approach where forensically relevant sequences are targeted through an amplicon-based workflow. The Illumina NGS amplicon workflow includes four basic steps (Figure 1):

1. **Library Preparation**—With amplicon sequencing, the library is prepared by performing a two-step amplification. The first PCR step uses sequence-specific, tagged primer pairs for each forensically relevant target sequence in the DNA sample. During the second PCR step, indexes and adapters are incorporated into the amplicons. The amplicon libraries are then purified, pooled into a single tube, and linearized.
2. **Cluster Generation**—During cluster generation, the library is injected into a flow cell where fragments are captured on a lawn of surface-bound oligos complementary to the library adapters. Each fragment is then amplified into distinct, clonal clusters through bridge amplification. With cluster generation complete, the templates are ready for sequencing.
3. **Sequencing**—Illumina SBS technology uses a reversible terminator-based method that detects single bases as they are incorporated into DNA template strands.[1] As all four reversible, terminator-bound dNTPs are present during each sequencing cycle, natural competition minimizes incorporation bias and greatly reduces raw error

rates compared to other technologies.[3-5] The result is highly accurate base-by-base sequencing that virtually eliminates sequence-context-specific errors, even within repetitive sequence regions and homopolymers.[4,5] The quality and accuracy of NGS data are important in forensic genomics, particularly when reporting results for mixed DNA samples, mtDNA heteroplasmy, or SNP data.

4. **Data Analysis**—During data analysis and alignment, the sequence reads are aligned to a reference genome. Following alignment, different types of analysis are possible such as single nucleotide polymorphism (SNP), STR typing, mtDNA analysis, phylogenetic or metagenomic analysis, and more.

A detailed animation of SBS technology is available at www.youtube.com/watch?v=womKfikWlxM.

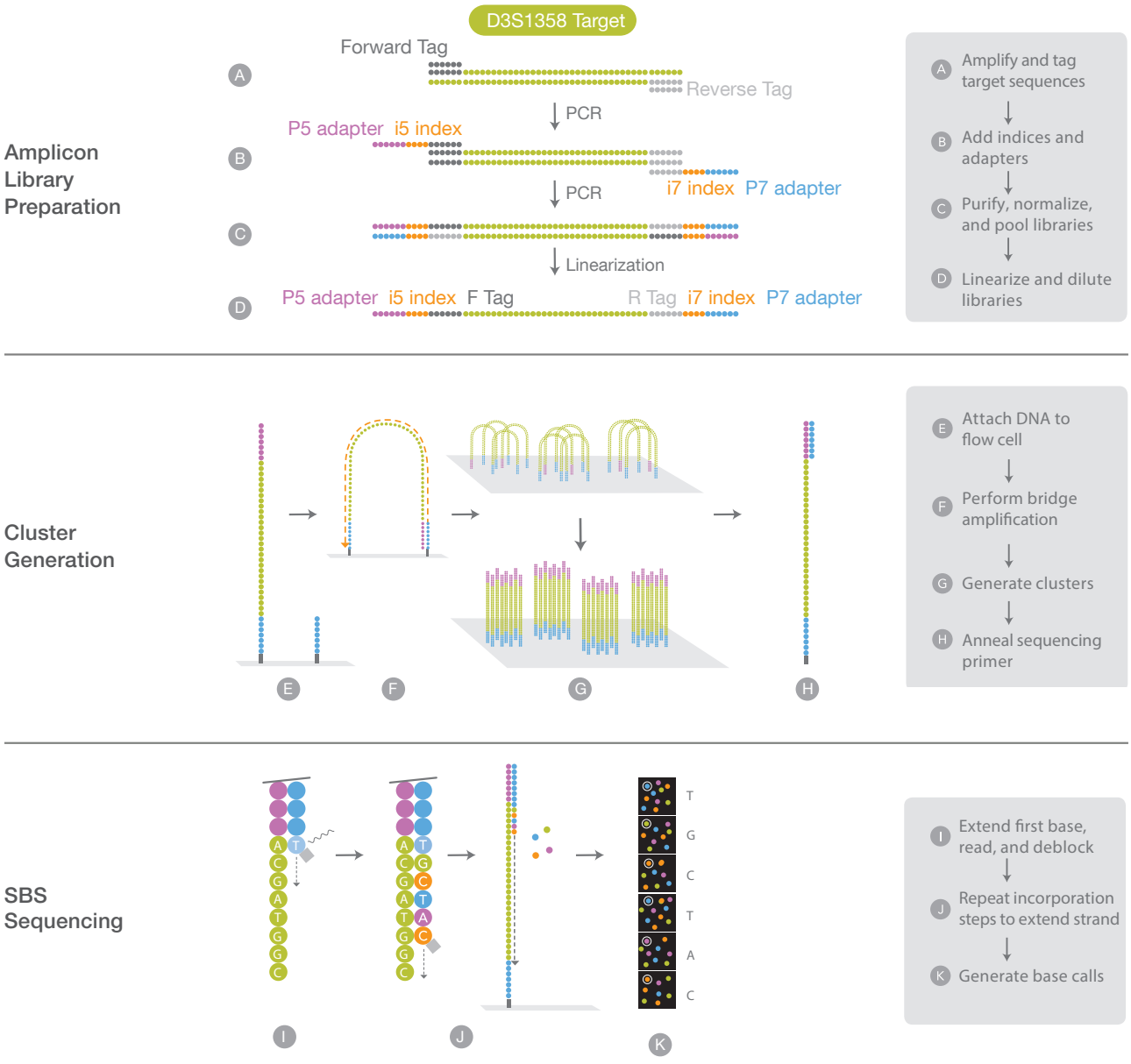

Figure 1: Basic NGS Amplicon Sequencing Overview.

III. NGS Solutions for Forensic Genomics

a. Human Identification with STRs and SNPs

Developed in collaboration with the forensic community and leveraging the proven technology of the MiSeq® System, Illumina created the MiSeq FGx™ Forensic Genomics System—the first fully validated[7] sequencing system designed for forensic genomics applications. The system provides a complete workflow for the analysis of forensic DNA samples from DNA-to-Data (Figure 2).

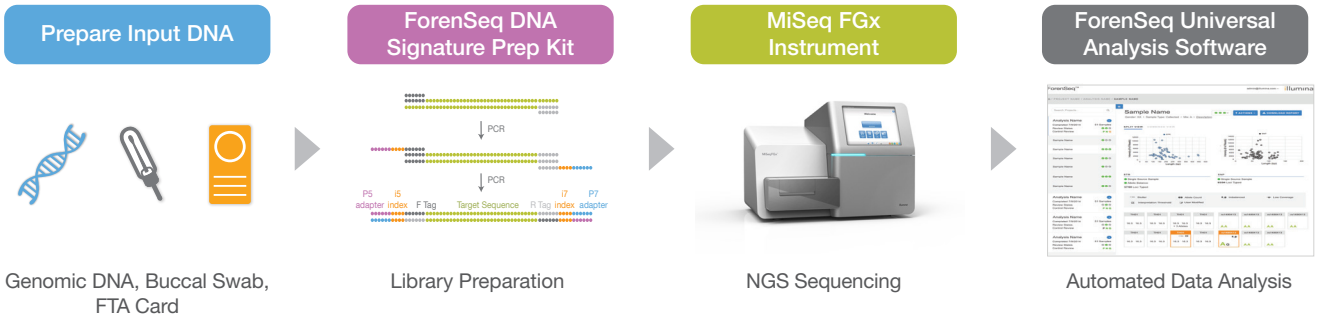

Figure 2 MiSeq FGx Forensic Genomics System Workflow. The MiSeq FGx Forensic Genomics System is a fully integrated, DNA-to-Data solution, including library preparation, DNA sequencing platform, and data analysis software designed for forensic genomics. The MiSeq FGx System offers the most complete, integrated workflow currently available.

The workflow begins with the ForenSeq™ DNA Signature Prep Kit, which includes reagents required to prepare the DNA library for sequencing using a simple, plate-based format and standard lab equipment. The ForenSeq DNA Signature Prep Kit includes over 200 forensically relevant genetic markers in a single, streamlined workflow (Table 1), eliminating the need for multiple STR kits.[8] The kit not only consolidates the autosomal STR markers currently utilized around the world for casework and criminal DNA databasing, it also includes Y- and X-STRs and SNP marker sets not routinely available with traditional CE-based methods. These include a dense set of identity-informative single nucleotide polymorphisms (iiSNPs),[9,10] that are informative for source attribution, especially with degraded, mixed, or PCR-inhibited samples. They also include phenotypic-informative SNPs (piSNPs),[11] which provide estimates of eye color (blue, intermediate, brown) and hair color (brown, red, black, blond), as well as biogeographical ancestry-informative SNPs (aiSNPs).[12,13]

Table 1. Forensic Loci Included in ForenSeq DNA Signature Prep Kit

Feature	Number of Markers[a]	Amplicon Size Range (bp)
Global Autosomal STRs	27	61–467
Y-STRs	24	119–390
X-STRs	7	157–462
Identity SNPs	95	63–231
Phenotypic SNPs[b]	22	73–227
Biogeographical Ancestry SNPs[b]	56	67–200

a. SNP and STR chromosome locations can be found in the ForenSeq DNA Signature Prep Kit User Guide (support.illumina.com/downloads/forenseq-dna-signature-prep-guide-15049528.html).

b. Two piSNPs used for hair/eye color are also used in the aiSNP marker set.

The MiSeq FGx Reagent Kit provides SBS sequencing reagents, an RFID labeled reagent cartridge, and wash solution, which are then loaded onto the MiSeq FGx instrument along with the sequencing-ready DNA libraries. The workflow concludes with the ForenSeq Universal Analysis Software, which delivers a powerful suite of forensic analysis capabilities. In addition to STR and SNP analysis for human identification, the ForenSeq software also includes automatic detection of mixed DNA samples, generation of population statistics, rapid sample comparison, and automated report generation. The ForenSeq software also enables estimation of visible traits and biogeographical ancestry markers that can provide crucial investigative leads in "no suspect" cases. To learn more about ForenSeq software features and capabilities, see the ForenSeq Universal Analysis Software Guide (15053876).

b. Mitochondrial DNA Analysis

Human remains exposed to the environment for extended periods can be challenging due to erosion. DNA samples from bone fragments or hair samples can be highly degraded or present in small quantities. Because mtDNA is present in hundreds of copies per cell, it can survive environments where nuclear DNA does not, and can therefore be a powerful tool for human identification. The Illumina workflow for mtDNA sequencing begins with DNA isolation and continues through final data analysis (Figure 3).

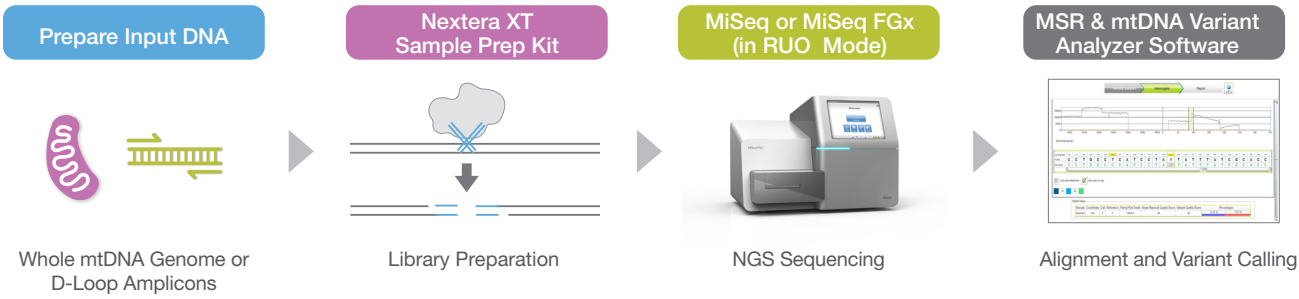

Figure 3: Illumina mtDNA Sequencing and Analysis Workflow. The Illumina mtDNA Sequencing and Analysis workflow provides an integrated solution for each step of the NGS sequencing workflow from DNA through final data analysis. The MiSeq FGx instrument, in research use only (RUO) mode, can be used for mtDNA analysis as well as a broad range of applications.

Following DNA extraction, mtDNA is amplified with a D-loop amplification protocol or with a whole mitochondrial genome protocol.[14,15] The sequencing library is prepared with the Nextera® XT DNA Library Prep Kit (Figure 4).

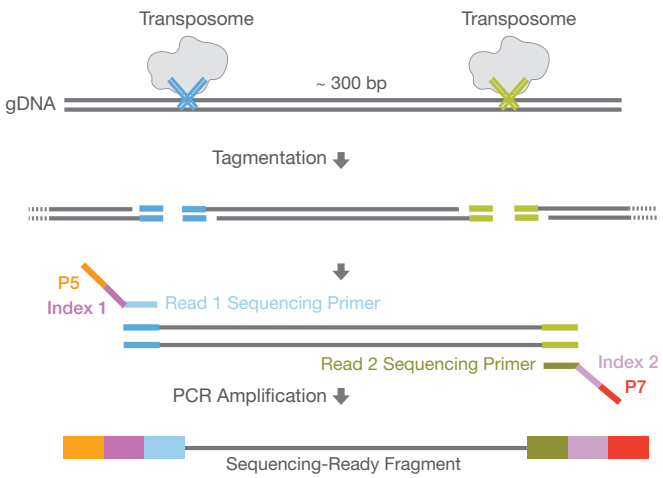

Figure 4: Nextera XT Library Prep Workflow. Nextera chemistry simultaneously fragments and tags DNA in a single step. A simple PCR amplification then appends sequencing adapters and indexes to each fragment.

The Nextera XT kit is designed to isolate specific genomic regions in low-quantity samples with maximum efficiency. For example, the Nextera XT kit allows rapid library preparation of mtDNA target regions from as low as 1 ng input gDNA. Using a single enzymatic "tagmentation" reaction to fragment and tag amplicons with sequencing adapters, the complete Nextera XT protocol takes less than 90 minutes (~15 minutes of hands-on time). With a multiplexing capacity of up to 384 libraries per sequencing run, the Nextera XT DNA Library Prep Kit transforms the time consuming Sanger sequencing process into a simple, streamlined workflow.

Following library preparation, mtDNA libraries can be sequenced on the MiSeq FGx System (in RUO mode) or on the original MiSeq System. While the MiSeq FGx System was designed for forensic analysis, it also supports additional applications in RUO mode allowing access to the full range of MiSeq NGS applications from RNA sequencing to human exome sequencing and more.

IV. The Advantages of Targeted Sequencing for Forensic Genomics

Whole-genome sequencing (WGS) reveals all the allelic differences between individuals across the whole human genome, including variations that occur in coding, regulatory, and intronic regions. While WGS is immensely valuable in the study of human biology and disease and delivers the most comprehensive genomic data, it also requires the largest data management and data analysis efforts. Rather than taking such a broad genomic view, forensic scientists typically perform targeted sequencing of forensically relevant loci. By sequencing a targeted subset of the genome, casework and database efforts are directed toward the genomic regions that best answer forensic questions. This approach relieves privacy concerns, produces manageable amounts of data, and simplifies data analysis—a common bottleneck in current forensic DNA workflows. Targeted sequencing also relieves many of the limitations imposed by CE where genotyping is based on fragment length detection.

a. Detection of Intra-STR SNPs

A clear advantage of NGS in forensic genomics is the ability to resolve alleles that are identical by size, but different by sequence. Intra-STR SNPs, mtDNA SNPs, and the complete sequence of unexpected signals or data artifacts can be evaluated at the nucleotide level providing a powerful and precise method for casework and human identification applications. The ForenSeq Universal Analysis Software user interface provides easy data visualization of intra-STR SNPs. The Locus Detail screen displays intensity charts, shows base-by-base DNA target sequences, and more (Figure 5).

b. Simultaneous Analysis of Multiple Polymorphisms

NGS overcomes a major limitation of CE typing: different classes of polymorphisms cannot be analyzed together, causing forensic laboratories to validate and maintain multiple PCR-based systems. CE typing often necessitates multiple rounds of testing, which can be impossible on limited or poor quality material, or may not yield sufficient information for a conclusive result (Figure 6). This creates the burden of maintaining redundant procedures, each requiring its own QA/QC and training programs.

NGS systems streamline testing by simultaneously analyzing large numbers of globally relevant STR markers and dense SNP sets in a single test. Furthermore, SNP markers from the ForenSeq DNA Signature Prep Kit improve the analysis of degraded samples due to their short amplicon length. They also enable phenotypic analysis of visible traits such as hair color, eye color, and biogeographic ancestry. NGS systems provide more data, while NGS-based STRs maintain standard allele nomenclature and compatibility with existing databases—enabling the consolidation of overlapping marker sets in use worldwide.

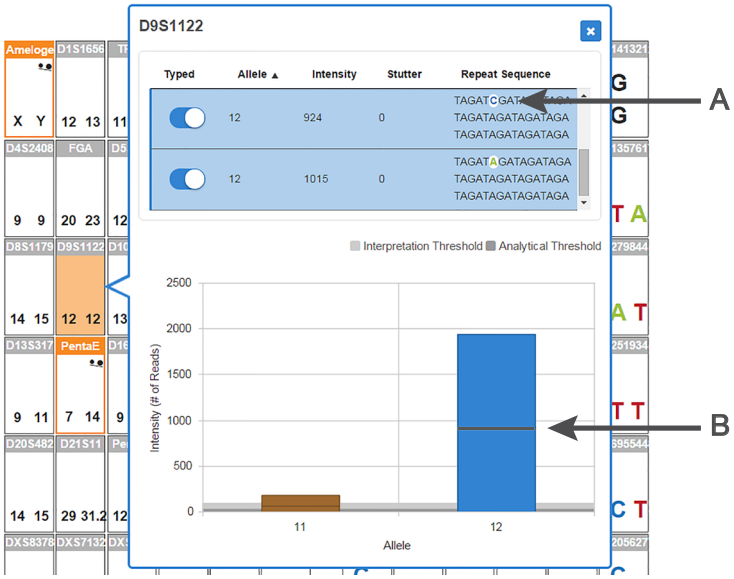

Figure 5: ForenSeq Universal Analysis Software- Locus Detail Screen and High-Resolution Genotyping. Click any Locus Genotype Box to see a pop-up Locus Details Screen. In this example, the Locus Details Screen shows that two 12 alleles of the same length, but with different DNA sequences, were detected at D9S1122. There are two indicators of intra-STR SNPs (isometric heterozygotes). A) The Repeat Sequence column shows the base-by-base sequence of each 12 allele and highlights intra-STR SNPs. B) The horizontal bar across the intensities of the two 12 alleles, shows the quantitative split in the genotype chart between the two 12 alleles that differ by sequence. CE cannot identify these different STR alleles because it performs genotyping based on STR amplicon length alone.

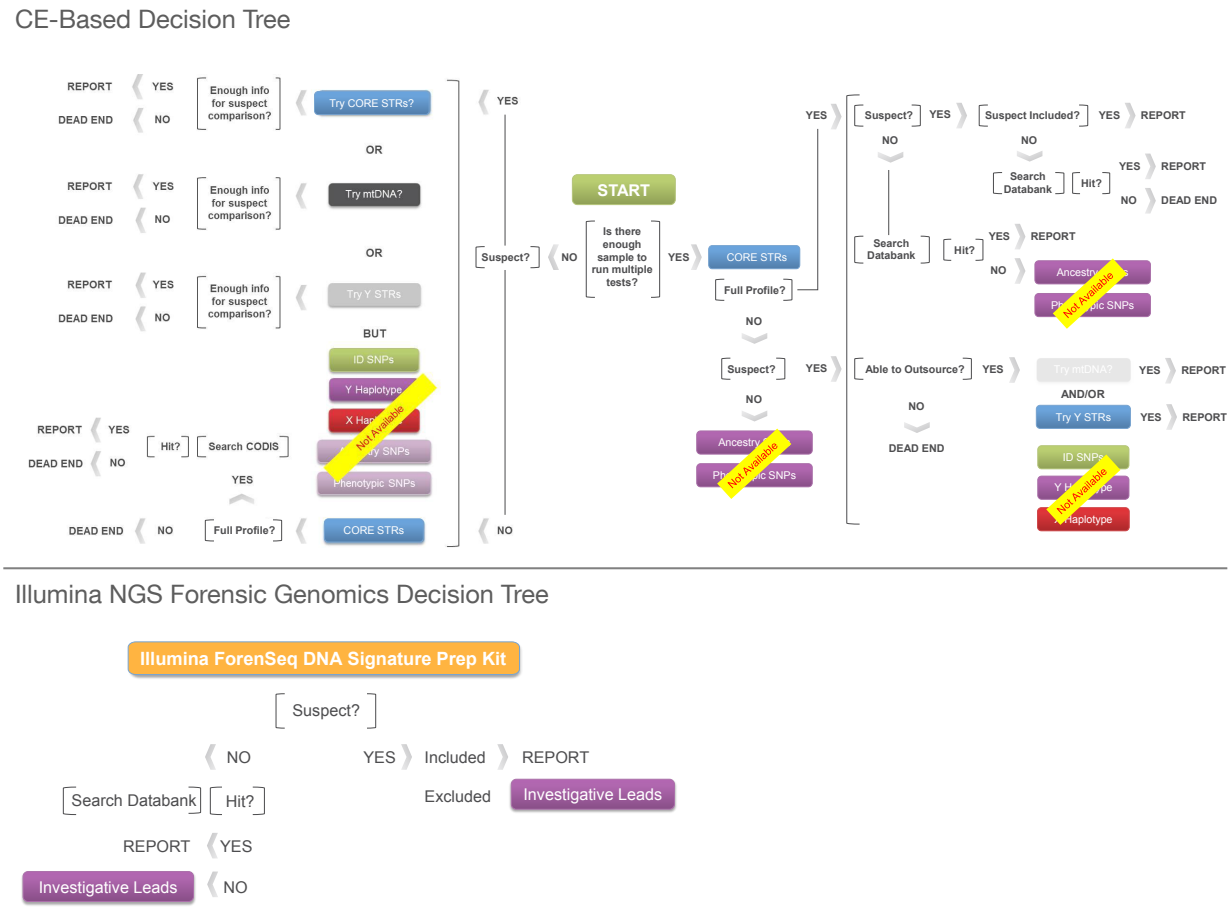

Figure 6. With ~200 genetic markers in a single workflow, the Illumina ForenSeq DNA Signature Prep Kit offers the simplest and most straightforward path to human identification.

c. Higher Sensitivity with Digital Data

CE-based systems produce analog metrics such as peak color, size, shape, and height, whereas all Illumina NGS systems deliver precise digital data (ie., discrete read counts). The digital nature of NGS and the ability to tune the sensitivity of an experiment by increasing or decreasing coverage level supports an unlimited dynamic range. Digital read counts and deep sequencing provide high sensitivity for quantitative applications such as detection of minor DNA contributors in complex mixtures, which can be missed or only partially detected using CE-based methods. For example, when performing mtDNA heteroplasmy analysis, NGS deep sequencing can detect minor variant frequencies of ~1% of the major compared to > 10–20% minor variant frequencies with CE-based sequencing.

d. Higher Throughput with Library Multiplexing

NGS enables library multiplexing through the attachment of unique barcodes (index sequences) to support scalable throughput at a level not possible with CE methods (Figure 7). Forensic libraries can be pooled in a controlled fashion to simultaneously sequence up to 96 libraries with the ForenSeq DNA Signature Prep Kit (384 potential) or 384 libraries with the Nextera XT DNA Library Prep Kit.

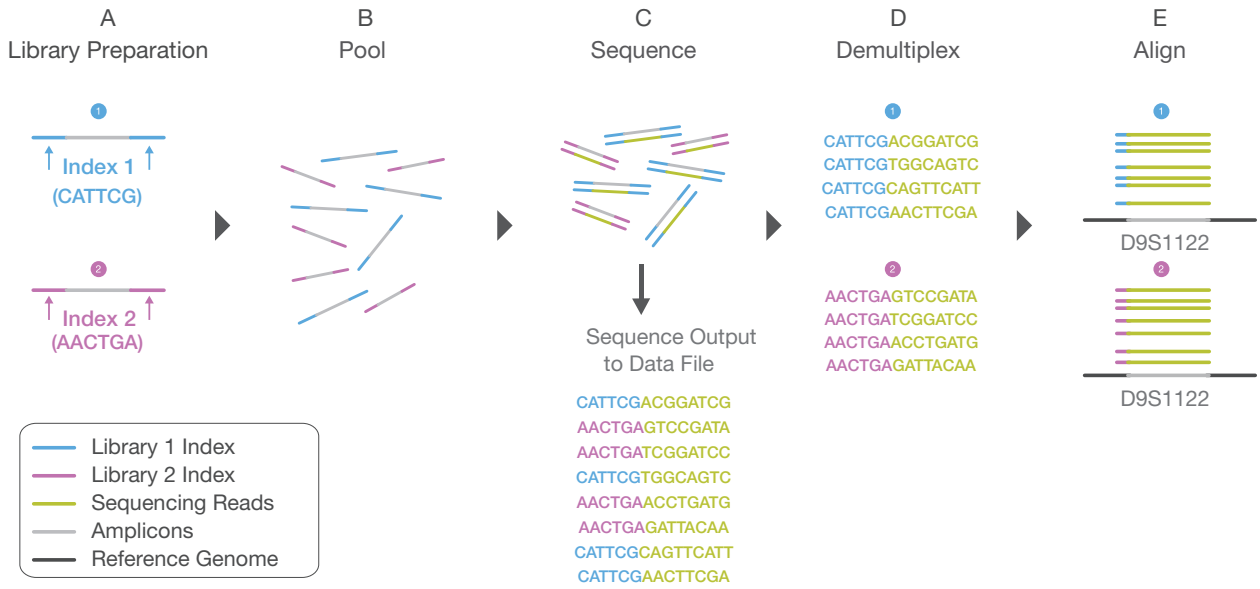

Figure 7: Library Multiplexing Overview.

A. Unique index sequences are attached to two distinct samples. Index sequences are attached during library preparation.

B. Indexed libraries are pooled together and loaded into the flow cell for the MiSeq FGx instrument.

C. Libraries are sequenced together during a single instrument run. Sequences are exported to a single output file.

D. De-multiplexing sorts the reads into different files according to their indexes.

E. Each set of reads is aligned to the appropriate ForenSeq target sequence.

V. Looking to the Future

Current forensic workflows artificially truncate the power of genomics, using techniques that pre-date the Human Genome Project and that require multiple rounds of analysis to produce complete genetic profiles. With NGS, forensic scientists have access to a greater number of informative loci, superior analysis of degraded samples, higher resolution sequencing, and greater overall throughput with library multiplexing. These advances will help solve more cases in a shorter amount of time and will produce investigative leads for cases that would have reached dead ends. With Illumina SBS technology at its core, the MiSeq FGx System is the first NGS system designed specifically for forensic genomic applications.

DNA database organizations are expanding their marker sets to enable more efficient and collaborative work with global and regional law enforcement. Research teams around the world continue to develop new capabilities using NGS technology in forensic case work and human identity testing. Current research efforts are evaluating piSNPs that provide association with attributes such as shape of the nose, lips, ears, and overall facial morphology, as well as with straight, or curly hair.[16,17] As forensic genomic methods continue to evolve, Illumina is dedicated to both supporting and advancing these efforts.

VI. Glossary

Adapters: Oligos bound to the 5'and 3'end of each DNA fragment in a sequencing library. The adapters are complementary to the lawn of oligos present on the surface of Illumina sequencing flow cells.

Artifact: Non-allelic products of the amplification process (eg. stutter or minus A) or anomaly of CE-based detection process (eg. spectral bleed through (pull up) or fluorescent spike).

Bridge Amplification: An amplification reaction that occurs on the surface of an Illumina flow cell. During flow cell manufacturing, the surface is coated with a lawn of two distinct oligonucleotides referred to as "p5" and "p7." In the first step of bridge amplification, a single-stranded sequencing library is injected into the flow cell. Individual molecules in the library bind to complementary oligos as they "flow" across the oligo lawn. Priming occurs as the free end of a ligated fragment bends over and "bridges" to another complementary oligo on the surface. Repeated denaturation and extension cycles (similar to PCR) results in localized amplification of single molecules into millions of unique, clonal clusters across the flow cell. This process, also known as "clustering," occurs on-board the MiSeq FGx instrument.

Buccal Swab: A Q-tip, cotton bud, or brush-like collector tool used to collect cells, in a noninvasive manner, from the inside of mouth/cheeks for DNA typing.

Capillary Electrophoresis (CE): The current method of choice for separating STRs based on size distribution.

Clusters: A clonal grouping of template DNA bound to the surface of a flow cell. Each DNA template strand that binds to the flow cell acts as a seed and is clonally amplified through bridge amplification until the cluster has roughly 1000 copies. Each cluster on the flow cell produces a single sequencing read. For example, 10,000 clusters on the flow cell would produce 10,000 single reads.

Coverage Level: The average number of sequenced bases that align to each base of the ForenSeq target loci. For example, a forensic DNA library sequenced at 30× coverage means that each base within the ForenSeq target loci was sequenced, on average, 30 times.

Flow Cell: A glass slide coated with a lawn of surface bound, adapter-complimentary oligos. A pool of 8–384 multiplexed libraries can be sequenced simultaneously, depending on application parameters.

Indexes/Barcodes/Tags: A unique DNA sequence ligated to fragments within a sequencing library for downstream *in silico* sorting and identification. Libraries with unique indexes can be pooled together, loaded into one lane of a sequencing flow cell, and sequenced in the same run. Reads are later identified and sorted via software.

Read: The process of next-generation DNA sequencing involves using instruments to determine the DNA sequence of a sample. In general terms, a sequence "read" refers to the data string of "A,T, C, and G" bases corresponding to the sample DNA. With Illumina technology, millions of reads are generated in a single sequencing run. In specific terms, each cluster on the flow cell produces a single sequencing read. For example, 10,000 clusters on the flow cell would produce 10,000 single reads.

Short Tandem Repeat (STR): Also known as microsatellites, they are repeating sequences of approximately 2–6 bp (ex: AATG- AATG-AATG) sequences that occur every 6–10 kb or so throughout the human genome. They have a high variability in the human population and serve as unique identifiers for individuals.

Single Nucleotide Polymorphism (SNP): A DNA sequence variation occurring when a single nucleotide (A, T, C, or G) in the genome differs between members of a species or between paired chromosomes in an individual. Due to variations between human populations, SNP alleles that are common in one geographical or ethnic group can be rare in another.

VII. References

1. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456(7218):53–9.
2. Nakazato T, Ohta T, Bono H. Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. *PLoS One* (2013)22;8(10):e77910.
3. Sebastian J, Fritz JS, Karola P, et al. Updating benchtop sequencing performance comparison. *Nat Biotechnol. 2013;*31:294–296.
4. Loman NJ, Misra RV, Dallman TJ, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol*. 2012;30:434–439.
5. Quail MA, Smith M, Coupland P, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences, and Illumina MiSeq sequencers. *BMC Genomics. 2012;*13:341.
6. Liu L, Li Y, Li S, et al. Comparison of next-generation sequencing systems. *J Biomed Biotechnol.* 2012;2012:251364.
7. The full MiSeq FGx System workflow is validated per the Scientific Working Group on DNA Analysis Methods (SWGDAM) guidelines (www.swgdam.org).
8. Illumina (2014) ForenSeq DNA signature prep kit data sheet (www.illumina.com/products/forenseq-dna-signature-kit.ilmn).
9. Kidd KK, Pakstis AJ, Speed WC, et al. Developing a SNP panel for forensic identification of individuals. *Forensic Sci Int.* 2006;164(1):20–32.
10. Sanchez JJ, Phillips C, Børsting C, et al. A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis*. 2006;27(9):1713–1724.
11. Walsh S, Liu F, Wollstein A, et al. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Sci Int Genet*. 2013;7(1):98–115.
12. Kidd KK, Speed WC, Pakstis AJ, et al. Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci Int Genet.* 2013;10:23–32.
13. Phillips C, Prieto L, Fondevila M, et al. Ancestry analysis in the 11-M Madrid bomb attack investigation. *PLoS One.* 2009;4(8):e6583.
14. Illumina. Human mtDNA genome protocol (support.illumina.com/downloads/human_mtdna_genome_guide_15037958.ilmn).
15. Illumina. Human mtDNA D-loop Hypervariable region protocol. (support.illumina.com/downloads/human_mtdna_d_loop_hypervariable_region_guide_15034858.ilmn).

16. Illumina (2012) By digging deeper into the genome, next-generation sequencing may yield more forensic clues. Interview. (applications.illumina.com/content/dam/illumina-marketing/documents/products/other/interview_budowle.pdf).

17. Claes P, Hill H, and Shriver MD. Toward DNA-based facial composites: preliminary results and validation. *Forensic Sci Int Genet.* 2014;13:208–16.

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH, FORENSIC, OR PATERNITY USE ONLY– NOT FOR ANY CLINICAL OR THERAPEUTIC USE IN HUMANS OR ANIMALS

 Pub. No. 770-2012-034 Current as of 05 February 2015

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACG
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTA
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACG
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGACG
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA