# NGS Data

# Next Generation Sequencing

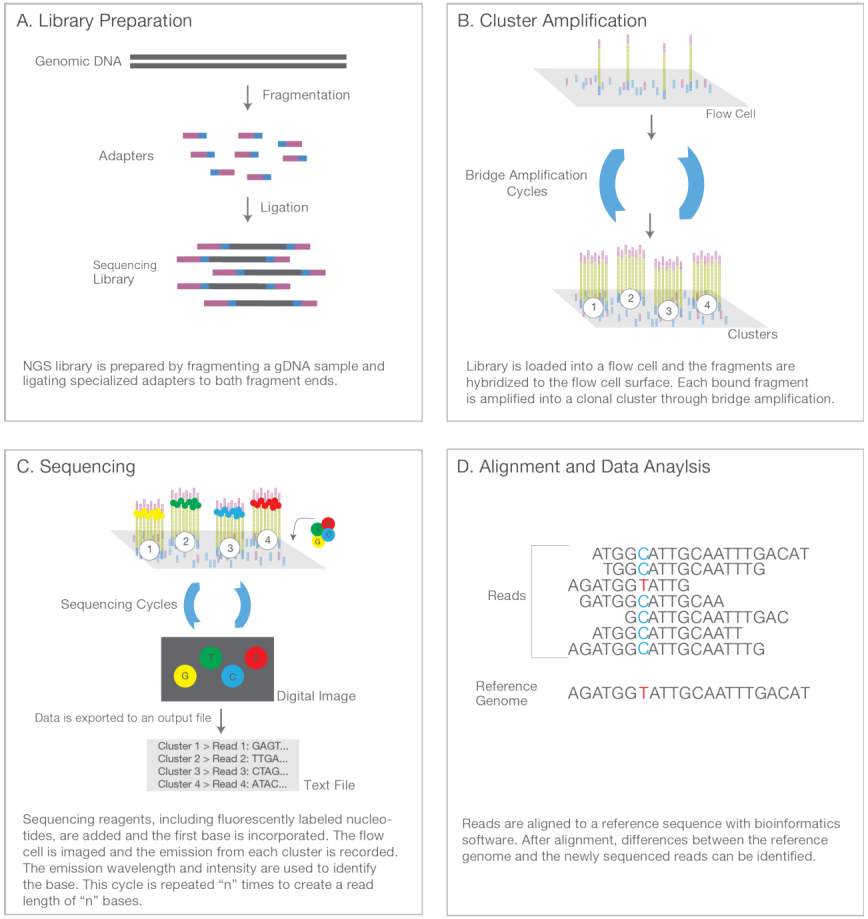
The introduction of *Next Generation Sequencing (NGS)* added a new dimension to the field of forensic genetics, providing distinct advantages over traditional CE systems in terms of captured information.

| **Locus** | **Allele number** | **Allele sequence** |
|---|---|---|
| D3S1358 | 15 | $[TCTA][TCTG]_3[TCTA]_{11}$ |
| D3S1358 | 15 | $[TCTA][TCTG]_2[TCTA]_{12}$ |
| D18S51 | 20 | $[AGAA]_{20}$ |
| D18S51 | 20 | $[AGAA]_{16}GGAA[AGAA]_3$ |

NGS is also referred to as Massively Parallel Sequencing (MPS), Second Generation Sequencing (SGS) or High-Throughput (HTP) sequencing.

# NGS Workflow

By far the biggest player in the field of sequencing instruments is Illumina. Their workflow includes four basic steps:




Source: An Introduction to NGS Technology (Illumina, 2015).

# NGS Data

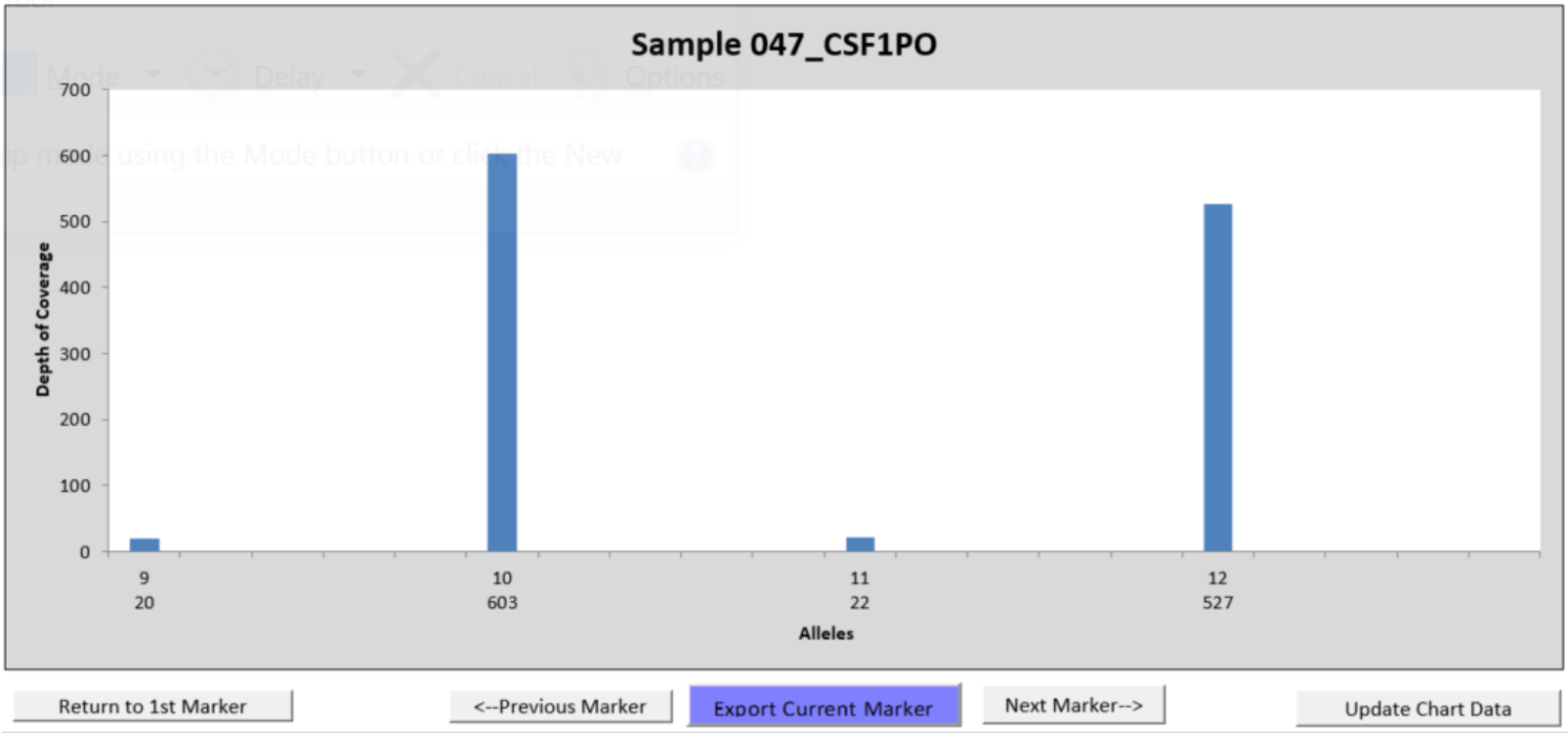
Results from sequencing platforms usually entail raw data, and need to be translated into information suitable for further (statistical) analysis.

- Software tools are available that align the reads to a reference sequence (**alignment**);
- Detect variations in the individual's genome (**variant calling**);
- And annotate the data using external information, resulting in a summarized data structure (**annotation**).

Instead of aligning to a reference sequence, sequence-searching techniques can be used that will use flanking sequences to detect STRs.

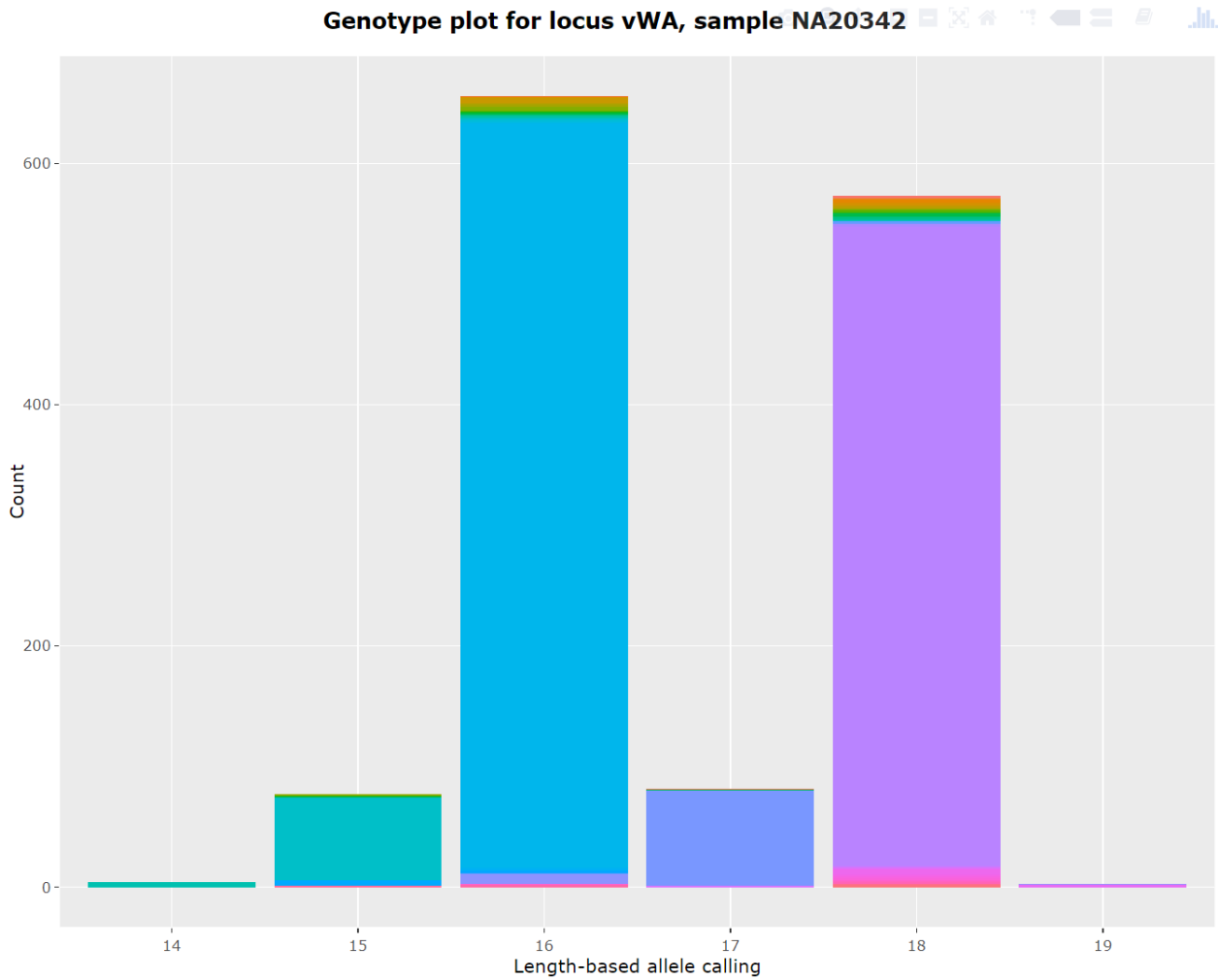
# NGS Data Output

A DNA profile can be visualized similar to an epg:



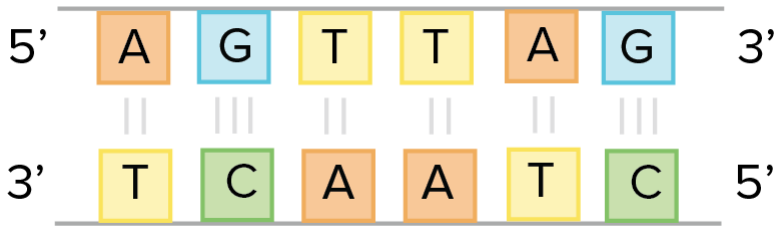
# NGS Data Output

A DNA profile can be visualized similar to an epg:



# NGS Considerations

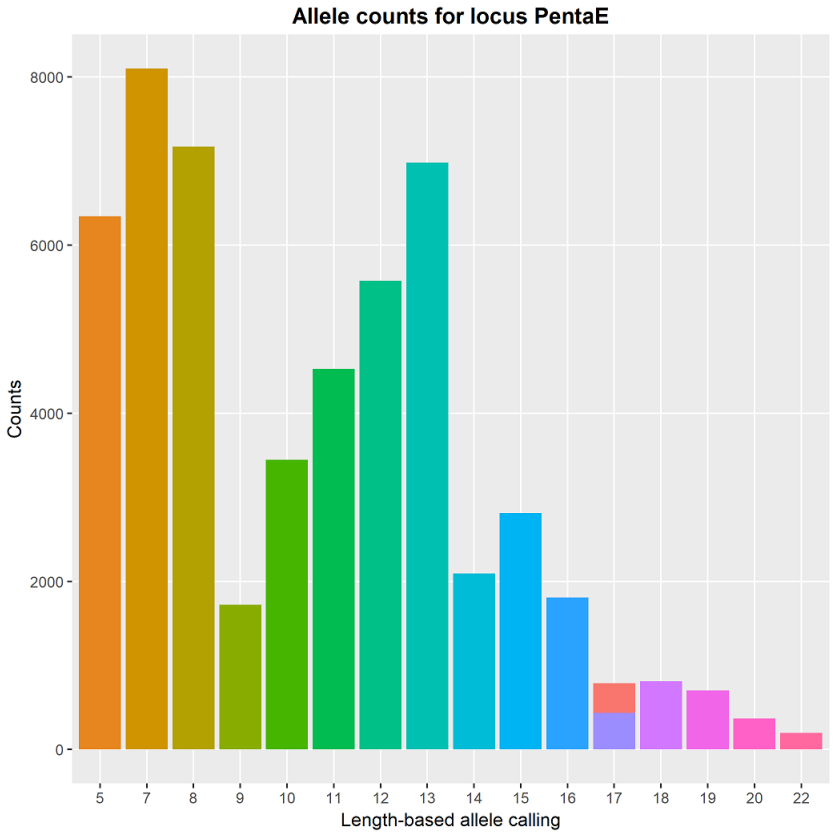
- Reads vs. peaks (discrete vs. continuous data)
- Discovery of previously unknown alleles and more variability
- New system of nomenclature needed
- Direction of strand reporting

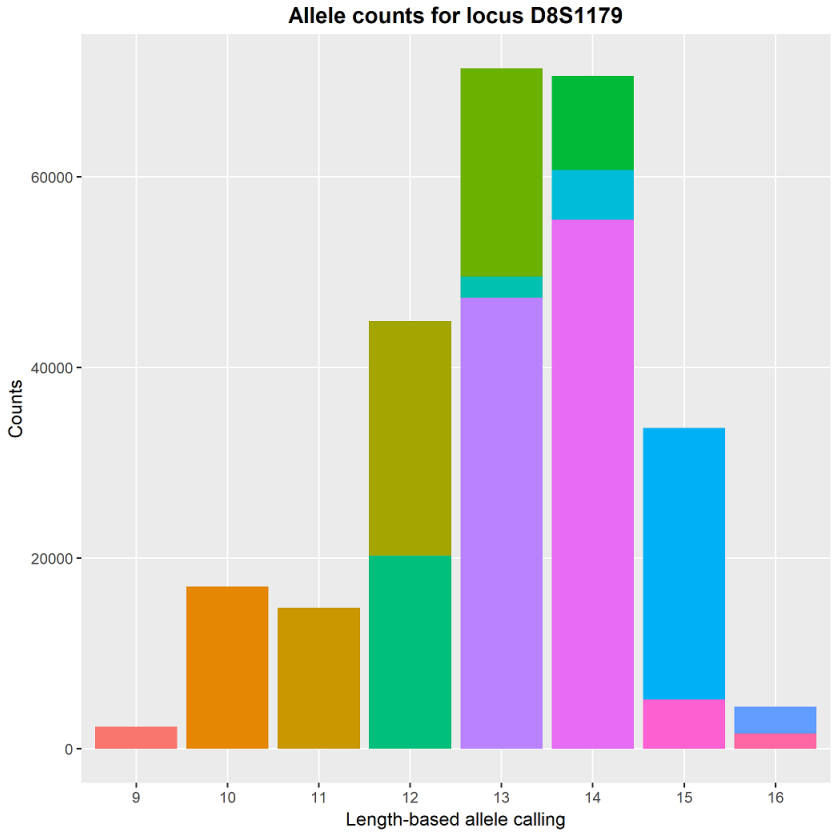



Source: `https://www.khanacademy.org/science/biology/dna-as-the-genetic-material/dna-replication/a/molecular-mechanism-of-dna-replication`.

# LB vs. SB Allele Callings

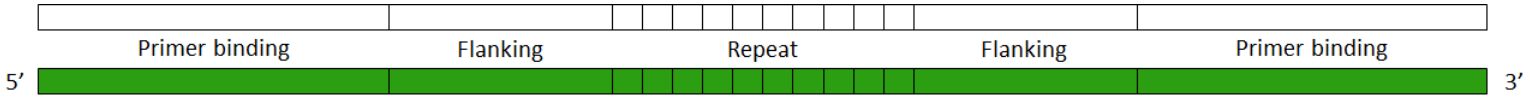
Locus Penta E is already quite polymorphic, so NGS data does not lead to significant improvements. For locus D8S1179, sequencing leads to a substantial increase in variability.



# Flanking Region SNPs

Additional variation has been found in the flanking regions adjacent to repeat motifs.



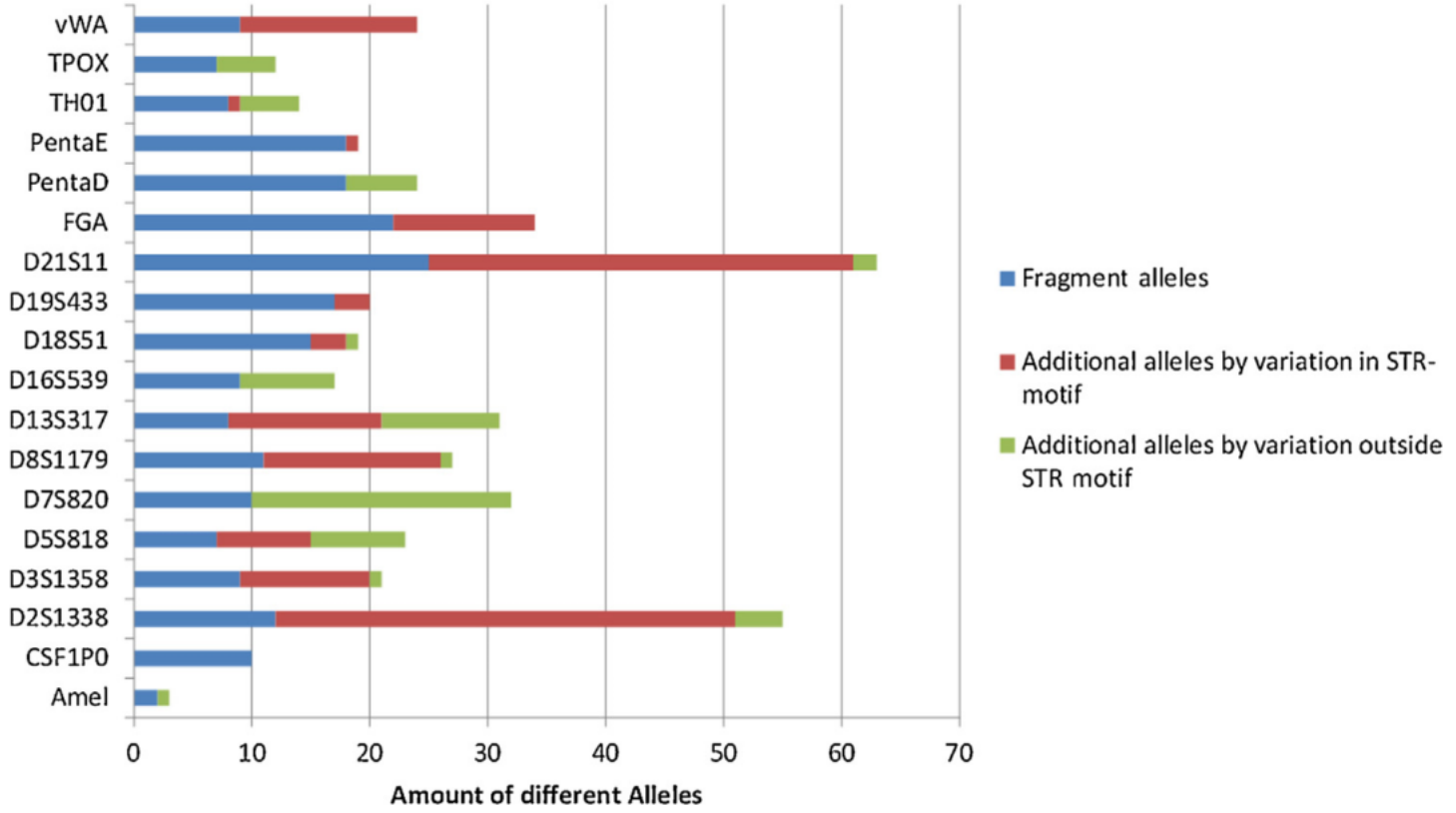

Source: Forensic DNA Evidence Interpretation (Buckleton et al., 2016).

For STR loci in which repeat regions do not display sequence differences, flanking region SNPs may still add substantial variability. Knowledge of these variants can be utilized in primer design to ensure optimal positioning during the PCR process.

| Locus | LB Allele | SB Allele | SB Allele with SNPs |
|---|---|---|---|
| D16S539 | 11 | $[GATA]_{11}$ | $[GATA]_{11}$rs11642858[A] |
| D16S539 | 11 | $[GATA]_{11}$ | $[GATA]_{11}$rs11642858[C] |

# Observed Sequence Variation

STR sequence variation divided in length variation, additional sequence variation, and SNP variation:




Source: Massively parallel sequencing of short tandem repeats (van der Gaag et al., 2016).