Whole Genome Sequencing (WGS)

Overview

- Sequence entire genome (~3 billion bases in humans)
- Captures all genetic variation including non-coding regions
- Most comprehensive genomic analysis method

Coverage	Cost	Time
30-50X	**$600-1000**	**1-3 days**
Clinical grade	Per sample	Sequencing + analysis

Applications

Clinical

- Rare disease diagnosis
- Cancer genomics
- Pharmacogenomics
- Prenatal screening

Research

- Population genetics
- Evolution studies
- GWAS studies
- Structural variants

Detects SNVs, indels, CNVs, and structural variants genome-wide

WGS Workflow Principles

1

DNA Extraction

High-quality genomic DNA isolated from blood, saliva, or tissue samples

↓

2

Library Preparation

DNA fragmented into ~300-500bp pieces, adapters ligated to both ends

↓

3

Sequencing

Millions of fragments sequenced in parallel using NGS platforms

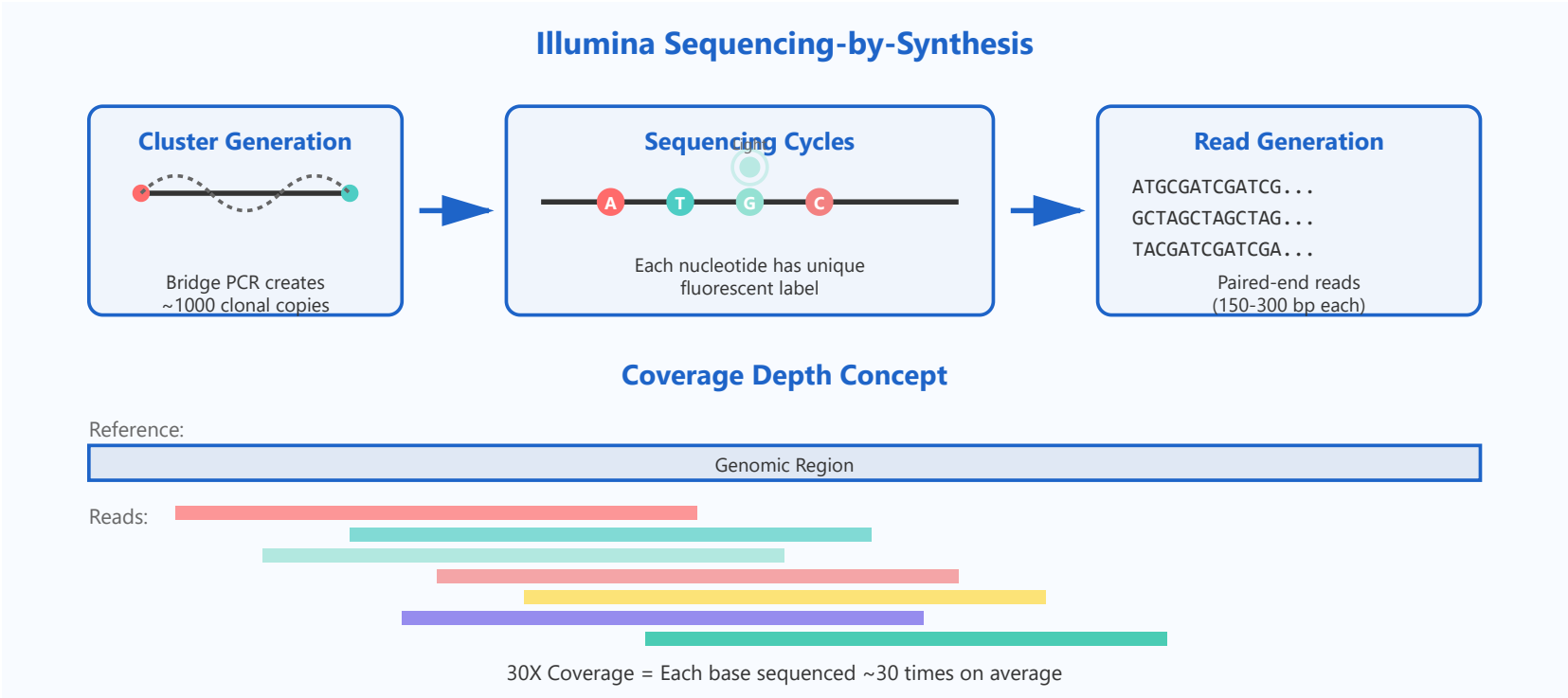
↓

4

Bioinformatics Analysis

Alignment to reference genome, variant calling, and annotation

Sequencing-by-Synthesis Principle



Detailed Process Explanation

1. Library Preparation

Genomic DNA is randomly fragmented into smaller pieces (300-500 bp). Adapter sequences are ligated to both ends of each fragment. These adapters contain:

- Sequences complementary to primers on the flow cell
- Index sequences for sample identification (multiplexing)
- Sequencing primer binding sites

2. Cluster Generation

Library fragments are loaded onto a flow cell surface coated with complementary oligonucleotides. Bridge amplification occurs:

- Fragment hybridizes to flow cell oligonucleotide
- DNA polymerase creates complementary strand forming a "bridge"
- Double-stranded bridge denatures, creating two single strands
- Process repeats ~1000 times, creating clonal cluster
- Result: Millions of spatially separated clusters across flow cell

---

3. Sequencing by Synthesis

Modified nucleotides (A, T, G, C) are added sequentially. Each nucleotide has:

- Unique fluorescent label for identification
- Reversible terminator preventing multiple incorporations

Process per cycle:

- All four nucleotides added simultaneously
- DNA polymerase incorporates complementary nucleotide
- Unincorporated nucleotides washed away
- Fluorescence captured by high-resolution camera
- Fluorescent label and terminator cleaved chemically
- Next cycle begins

---

4. Data Analysis Pipeline

Raw sequencing data undergoes comprehensive bioinformatics processing:

- **Base Calling:** Fluorescence signals converted to nucleotide sequences (FASTQ format)
- **Quality Control:** Reads filtered by quality scores (typically Q30+ retained)

- **Alignment:** Reads mapped to reference genome using algorithms (BWA, Bowtie2)
- **Variant Calling:** SNVs, indels, CNVs identified by comparing to reference
- **Annotation:** Variants annotated with functional, clinical, and population data
- **Interpretation:** Variants filtered and prioritized based on clinical significance

Key Technical Considerations

Coverage Uniformity

Not all regions covered equally. GC-rich regions, repetitive sequences, and structural variants may have lower coverage. 30X average ensures most regions adequately covered.

Read Length

Longer reads (150-300 bp paired-end) improve alignment accuracy, especially in repetitive regions. Insert size typically 300-500 bp for optimal genome coverage.

Error Rate

Illumina sequencing: ~0.1-1% error rate per base. High coverage depth allows confident variant calling by distinguishing true variants from sequencing errors.

Data Volume

Human WGS at 30X generates ~100 GB raw data per sample. Requires substantial computational resources for storage and analysis.

WGS provides comprehensive genome-wide view enabling detection of all variant types, making it the gold standard for genomic medicine and research