Whole Exome Sequencing (WES)

Overview

- Sequences only protein-coding regions (exons)
- Covers ~1-2% of genome (~30-50 Mb)
- Captures ~85% of known disease-causing variants

WES Advantages

- Lower cost than WGS
- Higher coverage per dollar
- Easier data analysis
- Smaller file sizes

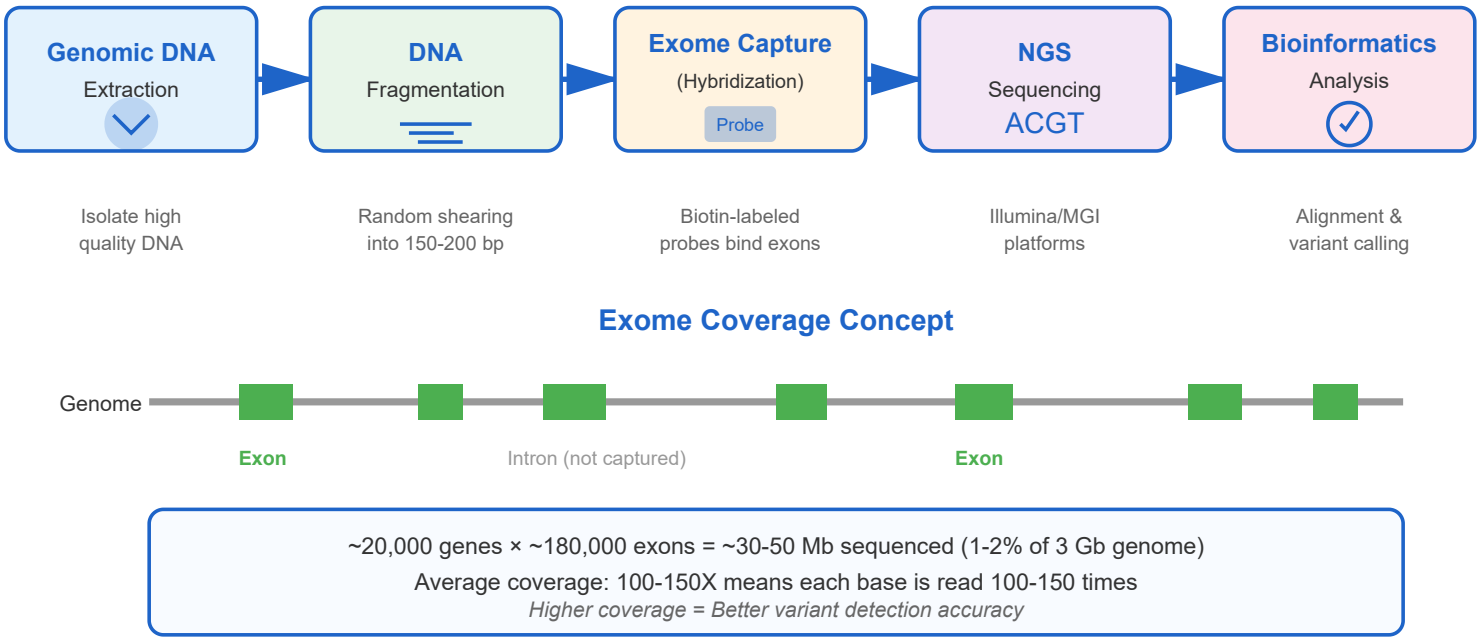
WES Limitations

- Misses regulatory variants
- Limited structural variant detection
- Capture bias
- Non-coding regions excluded

Coverage	Cost	Diagnostic Yield
100-150X	**$300-500**	**25-40%**

Preferred for Mendelian disorders and cancer driver mutations

WES Workflow Principles



Step-by-Step Principles

1. DNA Extraction & Quality Control

- Extract high molecular weight genomic DNA from blood, tissue, or saliva
- Quality assessment: DNA concentration (>50 ng/μL), purity (A260/280 ratio ~1.8), and integrity
- Typically requires 1-3 μg of input DNA

2. Library Preparation

- DNA fragmentation: Mechanical shearing (sonication) or enzymatic digestion to 150-200 bp fragments
- End repair: Create blunt ends and add 5' phosphate groups
- A-tailing: Add adenine bases to 3' ends
- Adapter ligation: Attach platform-specific adapters with unique barcodes (for multiplexing)

3. Exome Capture (Enrichment)

- **Key Technology:**Uses biotinylated RNA or DNA probes complementary to exonic sequences
- **Hybridization:**Library fragments hybridize with probes in solution (65°C, 16-24 hours)
- **Capture:**Streptavidin-coated magnetic beads bind biotin-labeled probe-target complexes
- **Washing:**Remove non-target DNA fragments through stringent washing steps
- **Popular kits:**Agilent SureSelect, Illumina Nextera, Twist Bioscience
- Enrichment efficiency: Typically 60-80% on-target rate

4. Next-Generation Sequencing

- **Platform:**Primarily Illumina (NovaSeq, NextSeq) or MGI sequencers
- **Chemistry:**Sequencing by synthesis (SBS) with fluorescent nucleotides
- **Read configuration:**Paired-end sequencing (2 × 150 bp most common)
- **Coverage target:**Mean depth of 100-150X for clinical applications
- **Output:**FASTQ files containing millions of short sequence reads
- Run time: 12-48 hours depending on platform and throughput

5. Bioinformatics Analysis Pipeline

- **Quality control:**FastQC analysis, adapter trimming, quality filtering
- **Alignment:**Map reads to reference genome (hg19/GRCh37 or hg38/GRCh38) using BWA or Bowtie2
- **Post-alignment processing:**Mark duplicates, base quality score recalibration (GATK)
- **Variant calling:**Identify SNVs and indels using GATK HaplotypeCaller, FreeBayes, or similar
- **Annotation:**Functional impact prediction (ANNOVAR, VEP, SnpEff)
- **Filtering:**Remove common variants, prioritize pathogenic mutations
- **Interpretation:**Clinical significance assessment using ACMG guidelines

Technical Considerations

Capture Efficiency Factors

- GC content bias
- Probe design quality
- Hybridization temperature
- DNA input quality
- Target region complexity

Coverage Uniformity

- Not all exons covered equally
- 90-95% of targets at >20X
- Some regions difficult to capture
- GC-rich regions may need higher depth

Limitations to Consider

- Cannot detect balanced translocations
- Misses copy number variants <1 kb
- Poor detection of repeat expansions
- Limited mtDNA analysis

Quality Metrics

- On-target rate: 60-80%
- Mean coverage: 100-150X
- Uniformity: >80% at 20X
- Duplication rate: <20%

Primary Clinical Applications

Rare Diseases
Mendelian disorders, developmental delays

Cancer Genomics
Somatic mutations, driver genes

Carrier Screening
Recessive disease alleles