Lecture 4:

Next-Generation Sequencing and Genomics

Ho-min Park

homin.park@ghent.ac.kr

powersimmani@gmail.com

Lecture Contents

Part 1/3:

Sequencing Technologies

1. Sanger Sequencing Recap
2. NGS Revolution Overview
3. Illumina Sequencing
4. Library Preparation
5. Paired-end vs Single-end Sequencing
6. Long-read Sequencing (PacBio)
7. Nanopore Sequencing

Sanger Sequencing Recap

Method

Chain termination sequencing using dideoxynucleotides (ddNTPs)

Year Introduced

1977 by Frederick Sanger (Nobel Prize 1980)

Read Length

400-900 base pairs per read

Accuracy

99.9% accuracy (very high)

Key Characteristics

- Gold standard for verification and validation
- Low throughput - sequences one fragment at a time
- Relatively expensive per base (~$500 per sample)
- Takes several hours to complete
- Best for targeted sequencing of specific genes

Clinical Use Today

Still widely used for confirming genetic variants and clinical diagnostics

NGS Revolution Overview

Sanger (Traditional)	
Throughput	**~1 Kb/day**
Cost per Mb	**~$500,000**
Parallelization	**Single reaction**
Time	**Hours-Days**

NGS (Next-Gen)	
Throughput	**~1 Tb/run**
Cost per Mb	**~$0.01**
Parallelization	**Millions of reads**
Time	**Hours-Days**

NGS Key Advantages

- ✓ Massive parallelization - sequence millions of fragments simultaneously
- ✓ Cost-effective - made genome sequencing affordable ($1000 genome)
- ✓ High throughput - entire human genome in 1-2 days
- ✓ Comprehensive - discover novel variants and structural changes
- ✓ Versatile - DNA, RNA, epigenetic, metagenomic applications

Illumina Sequencing (SBS)

1

Library Preparation

DNA fragments are prepared with adapters attached to both ends

2

Cluster Generation

DNA fragments bind to flow cell surface and amplify into clusters

3

Sequencing by Synthesis

Fluorescently labeled nucleotides are added one at a time

4

Imaging & Data Analysis

Camera captures fluorescent signals and converts to sequence data

Read Length **50-300 bp**	Accuracy **>99%**
Throughput	Market Share

Up to 6 Tb/run	~80% of NGS

Library Preparation

1. DNA Fragmentation

Break DNA into smaller fragments (200-600 bp)

↓

2. End Repair

Create blunt ends on DNA fragments

↓

3. Adapter Ligation

Attach sequencing adapters to both ends

↓

4. Size Selection

Select fragments of desired length

↓

5. PCR Amplification

Amplify library for sequencing

Quality Control

Check library size distribution and concentration using Bioanalyzer or TapeStation

Critical Factors

Input DNA quality, fragmentation method, and adapter ligation efficiency

Paired-end vs Single-end Sequencing

Single-end (SE)

Method: Sequence from one end only

Read Length: 50-150 bp

Cost: Lower ($)

Time: Faster

Use Case: Gene expression, small RNA-seq

Paired-end (PE)

Method: Sequence from both ends

Read Length: 2 × (75-300) bp

Cost: Higher ($$)

Time: Longer

Use Case: Variant calling, de novo assembly, structural variants

Paired-end Advantages

- ✓ Better alignment accuracy - confirms read location
- ✓ Detect structural variants and rearrangements
- ✓ Improved de novo assembly quality

✓ Span repetitive regions more effectively

Long-read Sequencing (PacBio)

PacBio SMRT Technology

- Single Molecule Real-Time (SMRT) sequencing
- Watches DNA polymerase in real-time
- Zero-mode waveguides (ZMWs) for detection

Read Length	Accuracy	Throughput
10-30 Kb	**99.9% (HiFi)**	**~30 Gb/run**

Advantages

- Sequence through repetitive regions
- Detect structural variants and complex rearrangements
- Better genome assembly - fewer gaps
- Native base modification detection (methylation)

Nanopore Sequencing

Technology Principle

- DNA/RNA passes through protein nanopore
- Changes in electrical current identify bases
- Real-time sequencing - no synthesis required

Read Length

Ultra-long reads: up to 2 Mb

Average: 10-100 Kb

Accuracy

Raw: ~95%

With consensus: >99%

Key Features

- Portable device (MinION USB sequencer)
- Real-time data analysis
- Direct RNA sequencing without reverse transcription
- Detect base modifications natively
- Rapid sequencing for outbreak response

Part 2: Data Processing

Part 2/3:

Data Processing

1. FASTQ Format
2. Quality Control (FastQC)
3. Read Alignment
4. SAM/BAM Formats
5. Variant Calling
6. VCF Format
7. Annotation Tools

FASTQ Format

FASTQ File Structure


```
@SEQ_ID (Sequence identifier)
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+ (Separator)
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Line 1: @Identifier

Unique read ID with instrument and run information

Line 2: Sequence

Raw nucleotide sequence (A, T, C, G, N)

Line 3: +

Separator (sometimes repeats identifier)

Line 4: Quality Scores

Phred quality scores (ASCII encoded)

Phred Score: $Q = -10 \times \log_{10}(P)$ | Q30 = 99.9% accuracy, Q40 = 99.99%

Quality Control (FastQC)

FastQC Metrics

- Per base sequence quality - quality drops at read ends
- Per sequence quality scores - overall read quality distribution
- Per base sequence content - nucleotide balance
- Sequence duplication levels - PCR duplicates
- Adapter content - leftover adapter sequences
- Overrepresented sequences - contamination check

Good Quality

- Phred score >30
- Balanced GC content
- Low duplication
- No adapter contamination

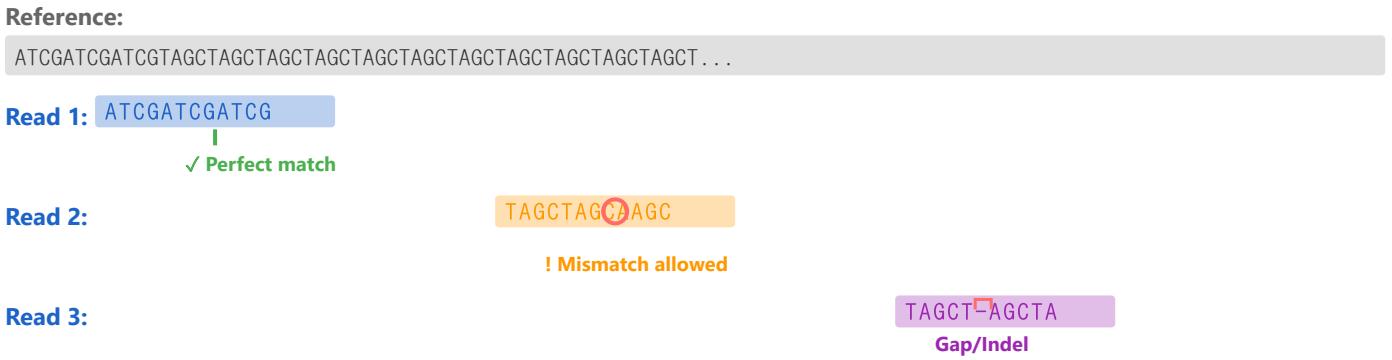
Poor Quality

- Phred score <20
- GC bias
- High duplication (>50%)
- Adapter sequences present

Common Tools: FastQC, MultiQC, Trimmomatic, Cutadapt

Read Alignment

Alignment Process



- Map sequencing reads to reference genome
- Find best matching position for each read
- Allow for mismatches and gaps (indels)
- Handle multi-mapping and unique reads

BWA

DNA-seq

Burrows-Wheeler Aligner

Bowtie2

DNA-seq

Fast, gapped alignment

STAR

RNA-seq

Splice-aware aligner

Key Considerations

- Read length
- Error rate
- Computational resources
- Paired-end vs single-end

Quality Metrics

- Mapping rate (>80% good)
- Properly paired (%)
- Coverage uniformity
- Duplicate rate

SAM/BAM Formats

SAM (Sequence Alignment/Map) Format

```
Header: @HD VN:1.6 SO:coordinate
@SQ SN:chr1 LN:248956422
Alignment: READ1 99 chr1 10001 60 76M = 10052 127 ACGT... IIII...
```

SAM (Text)

- Human-readable
- Tab-delimited
- Large file size
- 11 mandatory fields

BAM (Binary)

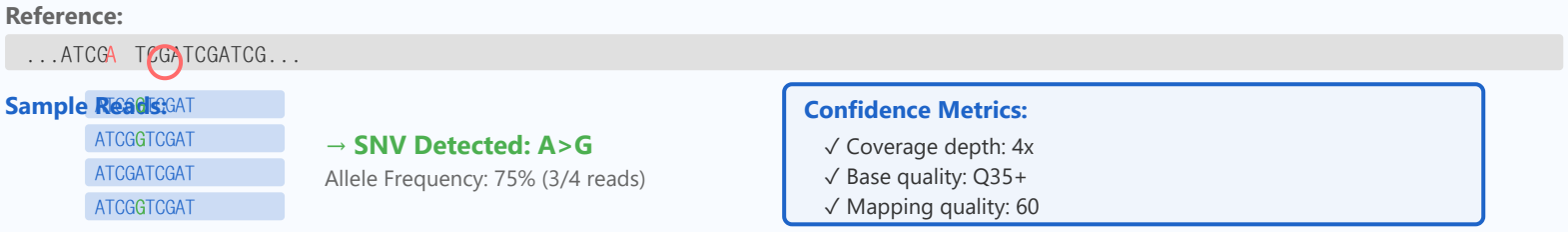
- Compressed SAM
- ~3-5x smaller
- Faster to process
- Requires indexing (.bai)

Key SAM Fields

- QNAME - Read name
- FLAG - Bitwise flag (paired, mapped, reverse, etc.)
- RNAME - Reference sequence name (chromosome)
- POS - Alignment position
- MAPQ - Mapping quality score

- CIGAR - Alignment string (M=match, I=insertion, D=deletion)

Variant Calling



Variant Calling Process

- Identify differences between sample and reference genome
- Distinguish true variants from sequencing errors
- Calculate confidence scores for each variant
- Filter low-quality and false positive calls

GATK

Gold Standard

Genome Analysis Toolkit

FreeBayes

Bayesian

Haplotype-based

DeepVariant

Deep Learning

Google AI method

Variant Types

SNVs/SNPs

Single nucleotide variants - most common (~50M per genome)

Indels

Small insertions/deletions - 1-50 bp

Structural Variants

Large deletions, duplications, inversions, translocations (>50 bp)

Copy Number Variants

Changes in gene copy number

VCF Format

VCF (Variant Call Format)

```
##fileformat=VCFv4.2
##reference=GRCh38
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1
chr1 10177 . A AC 50 PASS DP=32;AF=0.5 GT:DP:GQ 0/1:32:50
chr1 10352 rs123 T A 100 PASS DP=45;AF=1.0 GT:DP:GQ 1/1:45:99
```

VCF Columns

CHROM
Chromosome

POS
Position

REF
Reference

ALT
Alternate

QUAL
Quality

INFO
Annotations

Genotype (GT)

0/0 = homozygous reference

0/1 = heterozygous

1/1 = homozygous alternate

Key INFO Fields

DP = Total depth

AF = Allele frequency

AC = Allele count

AN = Total alleles

Annotation Tools

Variant Annotation Purpose

- Predict functional effect of variants
- Add gene names and transcript information
- Include population frequency data
- Clinical significance and disease associations
- Conservation scores and pathogenicity predictions

VEP

Ensembl

Variant Effect Predictor

ANNOVAR

Comprehensive

Multiple databases

SnpEff

Fast

Genomic annotations

Annotation Databases

Population Databases

• gnomAD (global frequencies)
• 1000 Genomes
• ExAC, dbSNP

Clinical Databases

• ClinVar (pathogenicity)
• OMIM (disease-gene)
• COSMIC (cancer)

Prediction Tools

- SIFT (deleteriousness)
- PolyPhen-2
- CADD scores

Conservation

- PhyloP
- GERP++
- PhastCons

Part 3: Applications

Part 3/3:

Applications

1. Whole Genome Sequencing
2. Whole Exome Sequencing
3. Targeted Panels
4. RNA-seq Overview
5. ChIP-seq
6. ATAC-seq
7. Metagenomics
8. Clinical Sequencing

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

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

Targeted Gene Panels

Overview

- Sequence specific set of genes related to condition
- Highly focused - typically 10-500 genes
- Very high coverage for selected regions (>500X)

Common Panel Types

Cancer
50-500 genes
Oncology hotspots

Cardio
50-200 genes
Heart conditions

Neuro
100-300 genes
Epilepsy, ataxia

Advantages

- Cost-effective ($100-300)
- Very high depth
- Faster turnaround
- Detect low-frequency variants

Use Cases

- Hereditary cancer screening
- Pharmacogenetic testing
- Carrier screening
- Targeted diagnostics

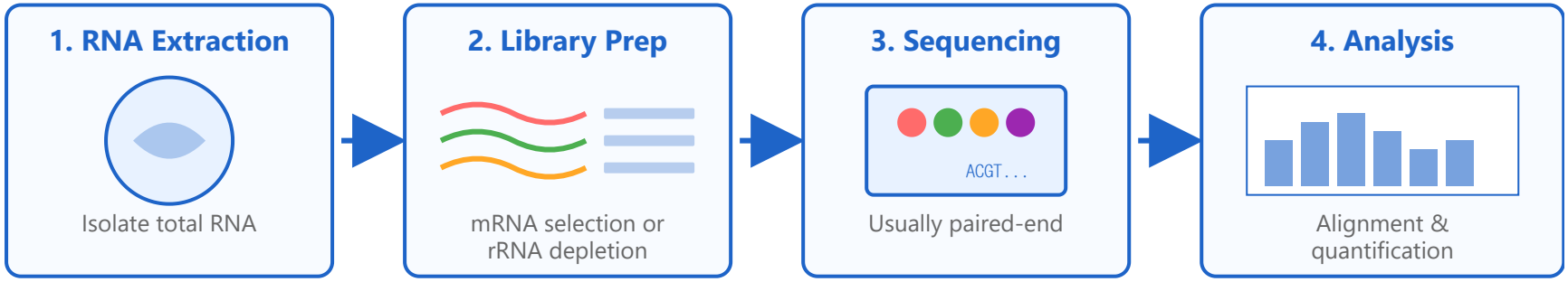
Best for known genes associated with specific phenotypes

RNA-seq Overview

What is RNA-seq?

- Sequence all RNA molecules in a sample
- Quantify gene expression levels
- Discover novel transcripts and splice variants
- Study transcriptome dynamics

RNA-seq Workflow



Applications

- Differential expression
- Alternative splicing

Key Tools

- STAR, HISAT2 (alignment)
- featureCounts (quantification)

- Novel transcript discovery
- Allele-specific expression

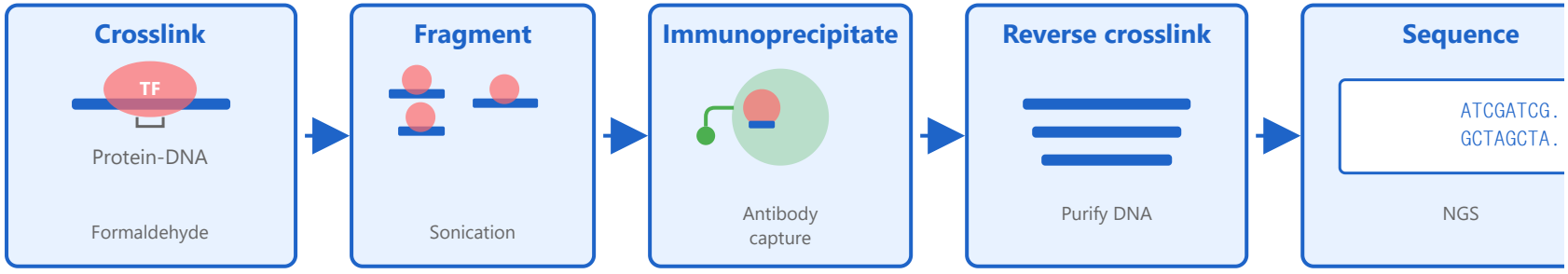
- DESeq2, edgeR (DE analysis)
- Salmon, kallisto (pseudo-alignment)

ChIP-seq (Chromatin Immunoprecipitation Sequencing)

Overview

- Identify protein-DNA binding sites genome-wide
- Study transcription factor binding
- Map histone modifications (epigenetics)
- Understand gene regulation mechanisms

ChIP-seq Workflow



Common Targets

Analysis Tools

- Transcription factors (TFs)
- H3K4me3 (active promoters)
- H3K27ac (active enhancers)
- H3K27me3 (repression)

- MACS2 (peak calling)
- deepTools (visualization)
- Homer (motif discovery)
- DiffBind (differential binding)

Requires high-quality antibodies and input control samples

ATAC-seq (Assay for Transposase-Accessible Chromatin)

Overview

- Map open chromatin regions genome-wide
- Identify active regulatory elements
- Requires fewer cells than ChIP-seq (500-50,000)
- No antibodies needed - uses Tn5 transposase

ATAC-seq Advantages

Technical Benefits

- Fast protocol (~3 hours)
- Low cell input
- No immunoprecipitation
- Less hands-on time

Biological Insights

- Nucleosome positioning
- TF footprinting
- Regulatory landscape
- Gene activity prediction

Cell Input	Protocol Time	Read Depth
500-50K	**~3 hours**	**50M reads**

Popular for single-cell studies (scATAC-seq) and epigenetic profiling

Metagenomics

What is Metagenomics?

- Study genetic material from environmental samples
- Analyze entire microbial communities
- No need to culture individual organisms
- Understand microbiome composition and function

Approaches

16S rRNA Sequencing

- Amplicon-based
- Taxonomic profiling only
- Cheaper, faster
- Bacterial/archaeal identification

Shotgun Metagenomics

- Whole genome sequencing
- Taxonomy + function
- All domains of life
- Discover novel genes/species

Applications

Clinical

Microbiome

Environmental

Ecology

Industrial

Biotechnology

Disease associations

Soil, water studies

Novel enzymes

Tools: Kraken2, MetaPhlAn, QIIME2, HUMAnN3

Clinical Sequencing

Clinical NGS Applications

- Diagnosis of rare genetic diseases
- Cancer precision medicine and treatment selection
- Pharmacogenomics - drug response prediction
- Prenatal and newborn screening
- Infectious disease identification

Clinical Considerations

Quality Standards

- CLIA/CAP certification
- High coverage (>30X)
- Validated pipelines
- Quality control metrics

Interpretation

- ACMG variant classification
- Clinical significance
- Actionable findings
- Secondary findings reporting

Ethical Issues

- Informed consent
- Incidental findings

Reimbursement

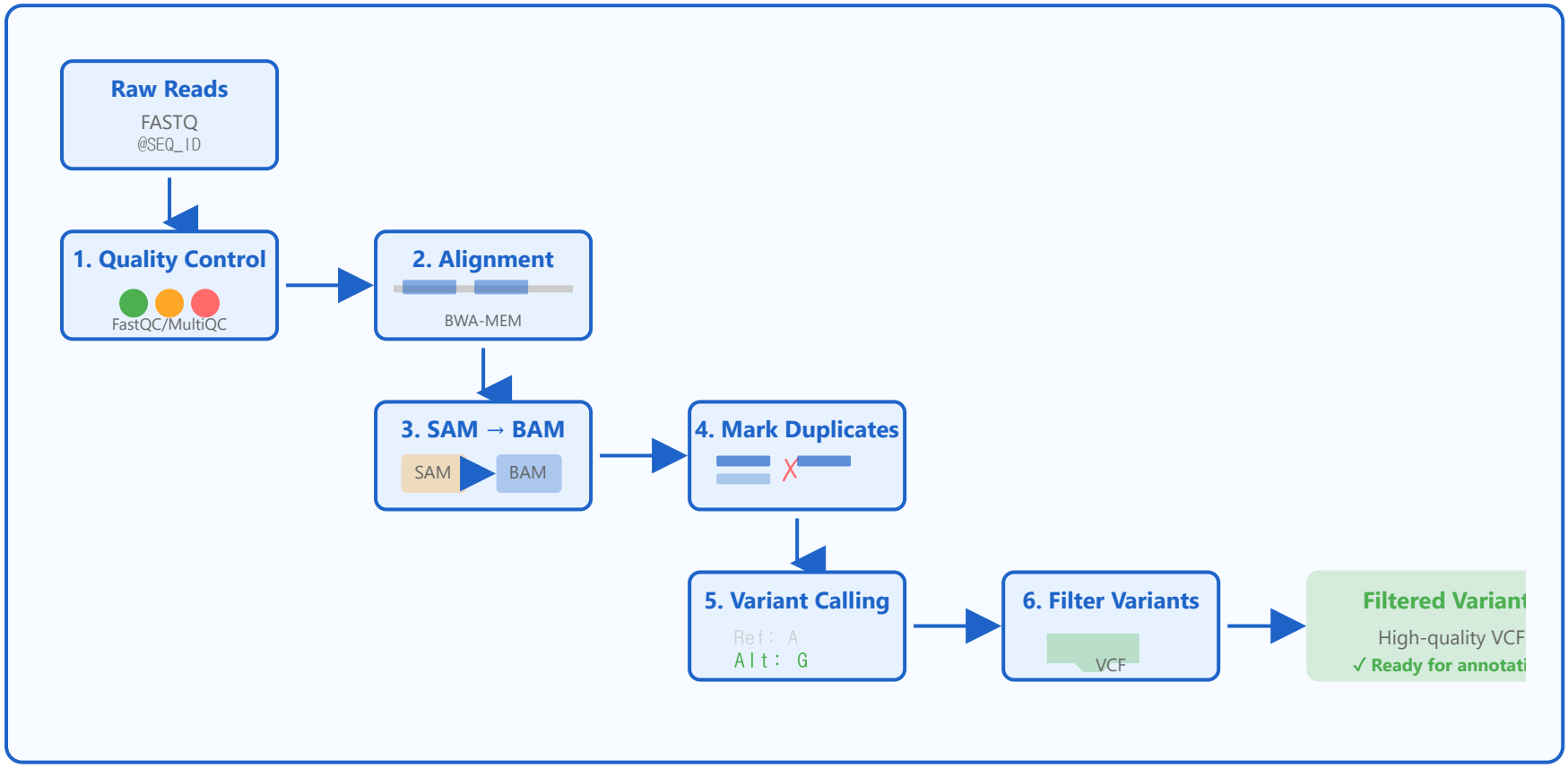
- Insurance coverage
- CPT codes

- Data privacy
- Genetic counseling

- Medical necessity
- Prior authorization

Requires multidisciplinary team: clinicians, geneticists, bioinformaticians, counselors

Hands-on: NGS Pipeline



Standard NGS Analysis Pipeline

```
# 1. Quality Control
fastqc sample_R1.fastq.gz sample_R2.fastq.gz
multiqc .
```

```
# 2. Read Alignment
bwa mem -t 8 reference.fa sample_R1.fastq.gz sample_R2.fastq.gz > sample.sam

# 3. Convert SAM to BAM and Sort
samtools view -bS sample.sam | samtools sort -o sample.sorted.bam
samtools index sample.sorted.bam

# 4. Mark Duplicates
gatk MarkDuplicates -I sample.sorted.bam -O sample.dedup.bam -M metrics.txt

# 5. Variant Calling
gatk HaplotypeCaller -R reference.fa -I sample.dedup.bam -O sample.vcf

# 6. Variant Filtering
gatk VariantFiltration -R reference.fa -V sample.vcf -O sample.filtered.vcf
```

Required Software

FastQC, BWA, SAMtools, GATK, Picard

Typical Runtime

4-24 hours depending on coverage and compute resources

Hands-on: Galaxy Platform

Galaxy: Web-based NGS Analysis

- User-friendly interface - no command line required
- Pre-installed tools and workflows
- Reproducible analysis with workflow sharing
- Public server: usegalaxy.org

Galaxy Workflow Example

Step 1: Upload Data

Upload FASTQ files from your computer or URL

Step 2: Quality Control

Run FastQC → Review reports → Trim if needed

Step 3: Alignment

Map with BWA-MEM → Select reference genome

Step 4: Variant Calling

FreeBayes or GATK → Generate VCF

Step 5: Annotation

SnpEff → Download annotated results

Access Galaxy training materials at training.galaxyproject.org

Thank you

Ho-min Park

homin.park@ghent.ac.kr

powersimmani@gmail.com