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

NGS Principles (Sequencing by Synthesis)

1. **Library Preparation** — Fragment DNA and attach adapters
2. **Cluster Generation** — Immobilize fragments and amplify
3. **Sequencing** — Add fluorescent labeled nucleotides
4. **Data Analysis** — Convert fluorescent signals to sequence

Sequencing by Synthesis Process

Cycle 1

```
Template: 5'-ATCGTAGC-3'
Growing: 3'-T
```

Red Signal (A)

Cycle 2

```
Template: 5'-ATCGTAGC-3'
Growing: 3'-TA
```

Green Signal (T)

Cycle 3

```
Template: 5'-ATCGTAGC-3'
Growing: 3'-TAG
```

Cycle 4

```
Template: 5'-ATCGTAGC-3'
Growing: 3'-TAGC
```

Yellow Signal (G)

Blue Signal (C)

Detailed Principle Explanation

Library Preparation

DNA Fragmentation: Genomic DNA is fragmented into 200-600bp sizes using physical or enzymatic methods.

Adapter Ligation: Universal primer sequences (adapters) are attached to both ends of fragments by ligation.

Size Selection: Optimal-sized fragments are selected using gel electrophoresis or bead-based methods.

Cluster Generation

Surface Attachment: DNA fragments bind to oligonucleotides immobilized on the flow cell surface.

Bridge PCR: Single DNA molecules are amplified to millions of copies at the same location.

Cluster Formation: Each fragment forms an independent cluster, amplifying the signal for detection.

Sequencing Chemistry

Reversible Terminators: dNTPs labeled with specific fluorophores and 3' blockers are added to each base.

Imaging: Fluorescent signals from each cluster are detected and imaged using lasers.

Cleavage and Repeat: Fluorophores and blockers are chemically removed, then the next cycle repeats.

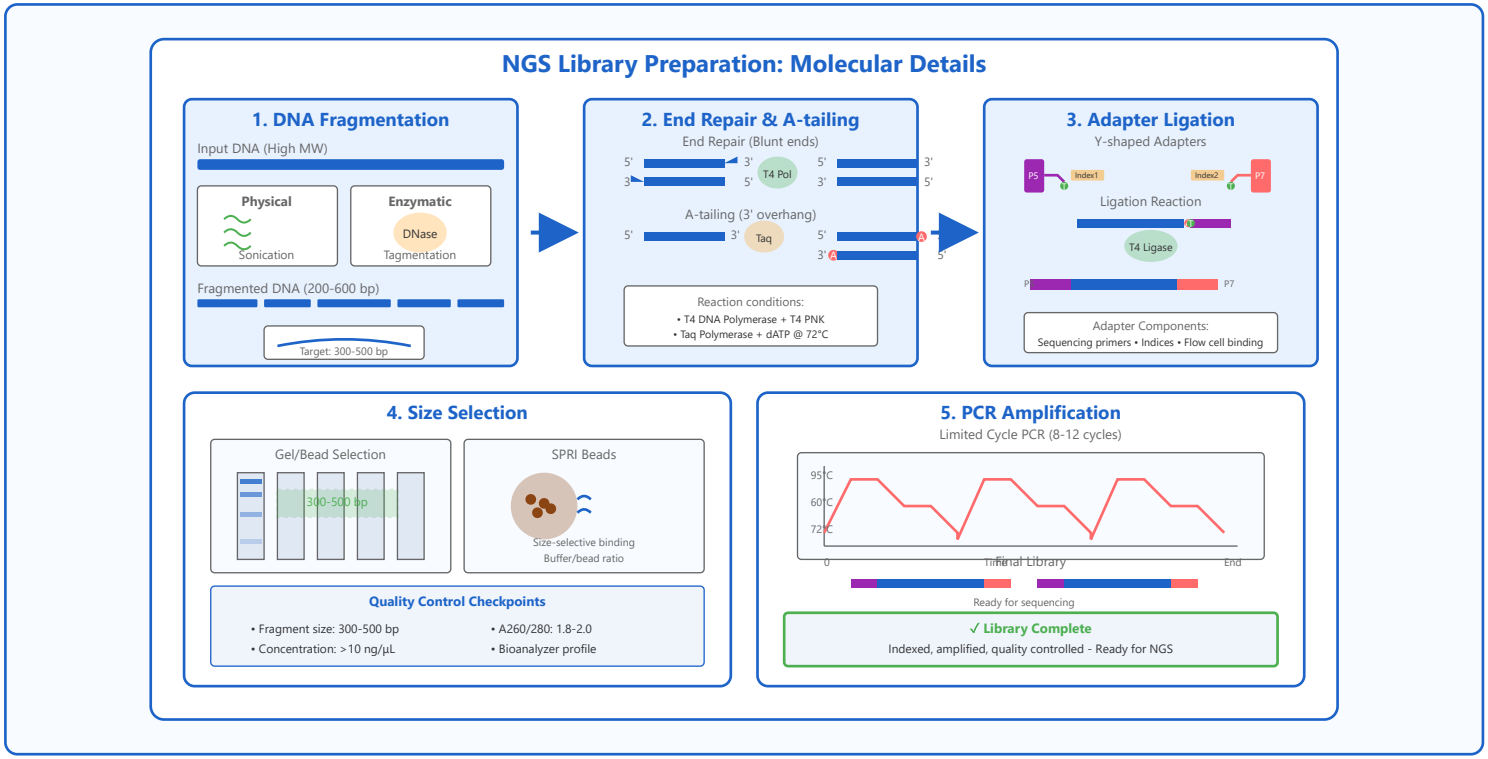
Data Processing

Base Calling: Fluorescent signal intensity and color from each cycle are analyzed and converted to ACGT sequences.

Quality Score: Accuracy of each base is calculated using Phred scores to assess reliability.

Alignment: Reads are aligned to reference genomes and variants (SNPs, Indels, etc.) are analyzed.

Library Preparation



Critical Factors

- Input DNA quality and quantity
- Fragment size distribution
- Adapter ligation efficiency
- Minimal PCR cycles to avoid bias

Library Types

- Whole genome libraries
- PCR-free libraries (reduce bias)
- Mate-pair libraries (long-range)
- Targeted capture libraries

Comprehensive explanations and diagrams for each libr

Detailed Step-by-Step Guide

1 DNA Fragmentation - Detailed Explanation

DNA fragmentation is the critical first step in NGS library preparation, where high molecular weight genomic DNA is broken into smaller fragments suitable for sequencing platforms. The target fragment size typically ranges from 200-600 base pairs, with most applications optimized for 300-500 bp inserts.

Physical Fragmentation Methods

- **Acoustic Shearing (Covaris):** Uses focused ultrasonic energy to generate controlled cavitation events that shear DNA. This method provides the most uniform size distribution and is highly reproducible.
- **Nebulization:** Forces DNA through a small aperture under high pressure, creating shear forces. Less expensive but offers less control over fragment size distribution.
- **Hydrodynamic Shearing:** DNA is forced through narrow channels, causing mechanical breakage. Provides good control but requires specialized equipment.

Enzymatic Fragmentation

- **Tagmentation (Nextera):** Uses a hyperactive Tn5 transposase that simultaneously fragments DNA and adds adapter sequences. This "tagmentation" process reduces library prep time significantly.
- **DNase I Digestion:** Controlled digestion with DNase I in the presence of Mg^{2+} ions. Fragment size is controlled by enzyme concentration and incubation time.
- **Restriction Enzymes:** Uses specific or frequent-cutting restriction enzymes to create defined breakpoints.

Fragmentation Methods Comparison

Physical Methods

Acoustic Shearing

- Most uniform distribution
- Highly reproducible
- 30-60 min processing
- High equipment cost

Enzymatic Methods

Tagmentation

- One-step process
- Fast (5-15 min)
- Low input DNA (1-50 ng)
- Some sequence bias

Parameter	Acoustic	Enzymatic
Uniformity	Excellent	Good
Speed	30-60 min	5-15 min
DNA Input	100 ng - 5 µg	1-50 ng
Sequence Bias	Minimal	Some bias
Equipment Cost	High ($$$)	Low ($)

2 End Repair & A-tailing - Detailed Ex

After fragmentation, DNA fragments have heterogeneous ends (5' overhangs, 3' overhangs, or blunt ends). End repair converts all fragments to blunt-ended, 5'-phosphorylated molecules suitable for adapter ligation. A-tailing then adds a single adenine (A) nucleotide to the 3' ends of these blunt fragments.

End Repair Process

The end repair reaction uses three enzymatic activities simultaneously:

- **T4 DNA Polymerase:** Fills in 5' overhangs with its 5'→3' polymerase activity and removes 3' overhangs with its 3'→5' exonuclease activity.
- **Klenow Fragment:** Fills in 5' overhangs and provides additional 3'→5' exonuclease activity.
- **T4 Polynucleotide Kinase:** Phosphorylates 5' ends, which is essential

End Repair & A-tai

Step 1: End

Before Repair (5' overhang):

5'
3'
3' 5'-P

↓ T4 DNA Polymerase

5'
3'
3'
5'-P

Enzyme Cocktail: T4 DNA Fragment • T4 PNK

Step 2: A-t

Adding 3' A-o

5'-P 3'
3'-OH

↓ Klenow (exo-) + dATP

5'-P A
3' T

✓ **Result:** 3' A-overhang comp adapters for effici

⚠ **Critical Consideration:** The fragmentation method impacts downstream bias. Acoustic shearing provides the most random fragmentation, while enzymatic methods may show sequence-specific preferences.

✓ **Best Practice:** Always validate fragment size distribution using a Bioanalyzer, TapeStation, or Fragment Analyzer before proceeding to the next step.

for subsequent ligation reactions.

A-tailing Purpose and Mechanism

A-tailing adds a single deoxyadenosine to the 3' end of blunt fragments using Klenow fragment (3'→5' exo-minus) or Taq polymerase. This creates a T overhang that is complementary to the T overhangs on adapter molecules, ensuring proper directional ligation.

- **Prevents Adapter Dimer Formation:** Only fragments with A-tails can ligate to T-overhang adapters.
- **Increases Ligation Efficiency:** T-A base pairing is more stable than blunt-end ligation.
- **Directional Cloning:** Ensures adapters ligate in the correct orientation.

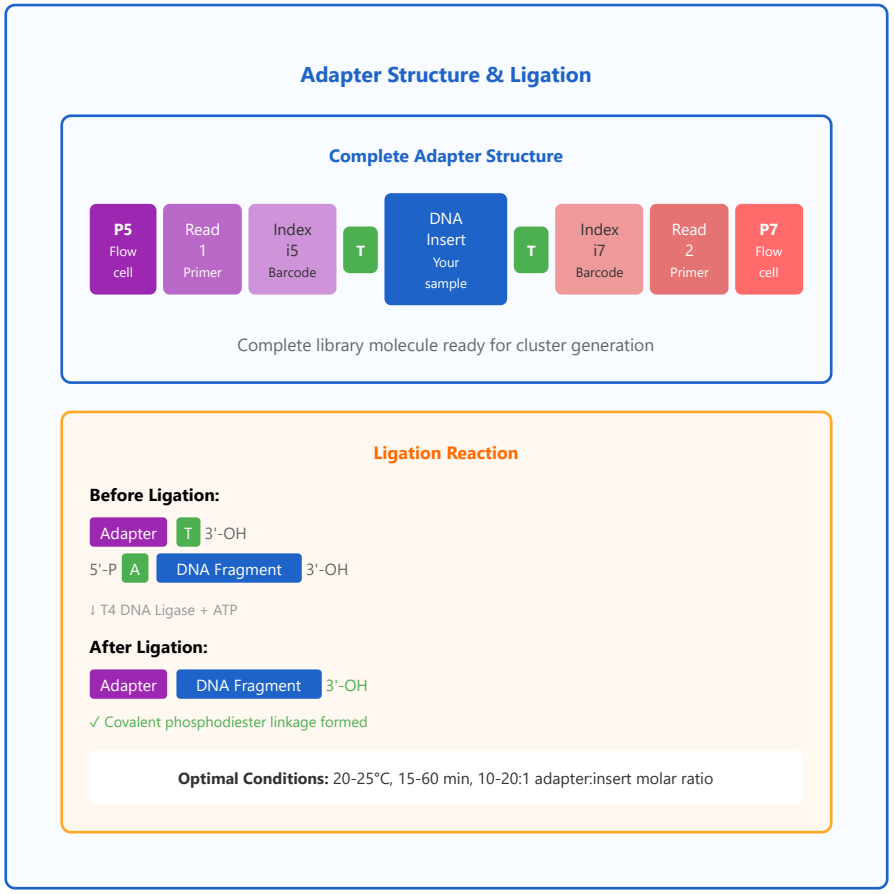
⚠ **Temperature Sensitivity:** End repair is typically performed at 20-25°C, while A-tailing requires 37-72°C depending on the enzyme. Improper

temperatures lead to incomplete reactions.

✓ **Quality Check:** Incomplete end repair or A-tailing dramatically reduces library yield. Some kits now offer combined end repair/A-tailing reactions to streamline the workflow and reduce sample loss.

3 Adapter Ligation - Detailed Explanation

Adapter ligation is the process of attaching synthetic oligonucleotide adapters to both ends of the prepared DNA fragments. These adapters contain several critical elements necessary for NGS sequencing and are the defining feature that converts fragmented DNA into a "sequencing library."

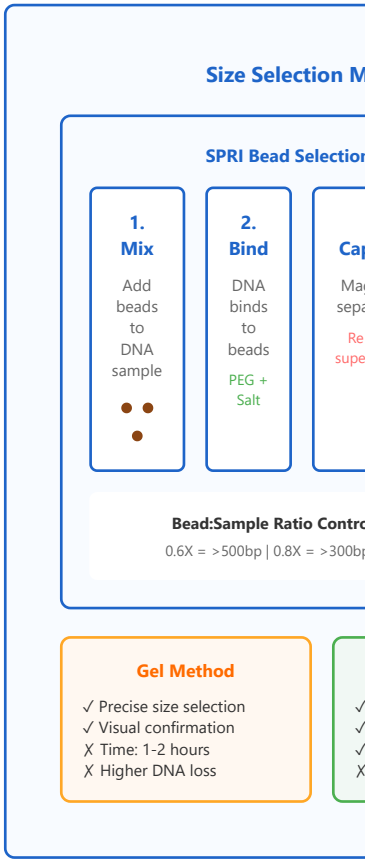


Adapter Structure and Components

- **P5 and P7 Binding Sites:** Complementary sequences to oligonucleotides on the flow cell surface, enabling cluster generation during bridge amplification.
- **Sequencing Primer Binding Sites:** Allow

4 Size Selection - Detailed Explanatio

Size selection is a critical quality control step that removes unwanted DNA fragments from the library preparation, including adapter dimers, very short fragments, and excessively long fragments. Proper size selection ensures optimal cluster density, sequencing quality, and data output.



Methods for Size Selection

1. SPRI Beads (AMPure XP)

SPRI beads are paramagnetic beads coated with carboxyl groups that reversibly bind DNA in the presence of polyethylene glycol (PEG) and salt. This is currently the most popular method.

- **Mechanism:** DNA binding efficiency depends on the bead:sample volume ratio.

sequencing primers to hybridize and initiate sequencing reactions.

- **Index Sequences (Barcodes):** 6-10 bp unique identifiers that enable sample multiplexing. Can be single-indexed (one barcode) or dual-indexed (two barcodes for added specificity).
- **T-overhangs:** Single thymine nucleotides at the 3' ends that complement the A-tails on DNA fragments.

Ligation Chemistry

T4 DNA Ligase catalyzes the formation of phosphodiester bonds between the 3'-OH of the A-tailed insert and the 5'-phosphate of the T-overhang

- **Size selection strategy:**
  - 0.6X bead ratio: Selects fragments >500 bp
  - 0.8X bead ratio: Selects fragments >300 bp
  - 1.8X bead ratio: Selects fragments >100 bp
  - Double-sided selection (0.5X then 0.7X): Narrow range (300-500 bp)
- **Advantages:** Fast (15-20 min), scalable, minimal DNA loss

2. Gel Electrophoresis

Traditional agarose gel electrophoresis provides the most precise size selection but is labor-intensive.

- **Procedure:** Run library on 2% agarose gel, visualize, excise target band (400-500 bp), purify

adapter. The reaction:

- Requires ATP as a cofactor
- Is typically performed at 20-25°C for 15-60 minutes
- Benefits from PEG (polyethylene glycol) to create molecular crowding and increase effective concentration

Multiplexing Strategy

Index sequences enable pooling of multiple samples in a single sequencing run. During data analysis, reads are "demultiplexed" based on their index sequences to assign them back to individual samples. Dual indexing (i5 and i7 indices) provides additional accuracy and can correct for index hopping in patterned flow cells.

DNA from gel slice

- **Advantages:** Visual confirmation, precise selection, complete adapter dimer removal
- **Disadvantages:** Time-consuming (1-2 hours), higher DNA loss

⚠ Adapter Dimer Problem: Adapter dimers (~120 bp) are highly problematic because they:

- Compete with insert-containing molecules during cluster generation
- Sequence more efficiently than longer fragments
- Can consume 20-80% of sequencing capacity if not removed
- Provide zero useful data

⚠ **Adapter Dimers:** A major challenge is adapter-adapter ligation without insert DNA, creating short molecules (~120 bp) that sequence efficiently but provide no useful data. Proper insert:adapter ratios and size selection steps are critical to minimize dimers.

✓ **Optimization Tip:** Use a 10-20 molar excess of adapters to DNA fragments to ensure complete ligation while minimizing adapter dimers. Always verify complete removal of the ~120 bp adapter dimer peak before sequencing!

✓ **Quality Control:** Use a Bioanalyzer, TapeStation, or Fragment Analyzer to verify:

- Complete removal of adapter dimers (~120 bp peak)
- Tight size distribution around target size (e.g., 400-500 bp)
- Adequate DNA concentration for sequencing (typically >2-5 nM)

include a size selection step post-ligation to remove excess unligated adapters and adapter dimers.

5 PCR Amplification - Detailed Explanation

PCR amplification is the final step of library preparation, enriching the adapter-ligated DNA fragments to generate sufficient material for sequencing. This step must be carefully optimized to achieve adequate library concentration while minimizing amplification bias and artifacts.

Purpose of Library Amplification

- **Increase DNA Quantity:** Generate enough library material (typically 10-50 nM in 25-50 µL) for accurate quantification and sequencing
- **Enrich Properly Ligated Molecules:** Selectively amplify fragments with adapters on both ends
- **Add Flow Cell Binding Sites:** Incorporate complete P5 and P7 sequences needed for cluster generation
- **Introduce Additional Indices:** Some protocols add sample indices during PCR rather than ligation

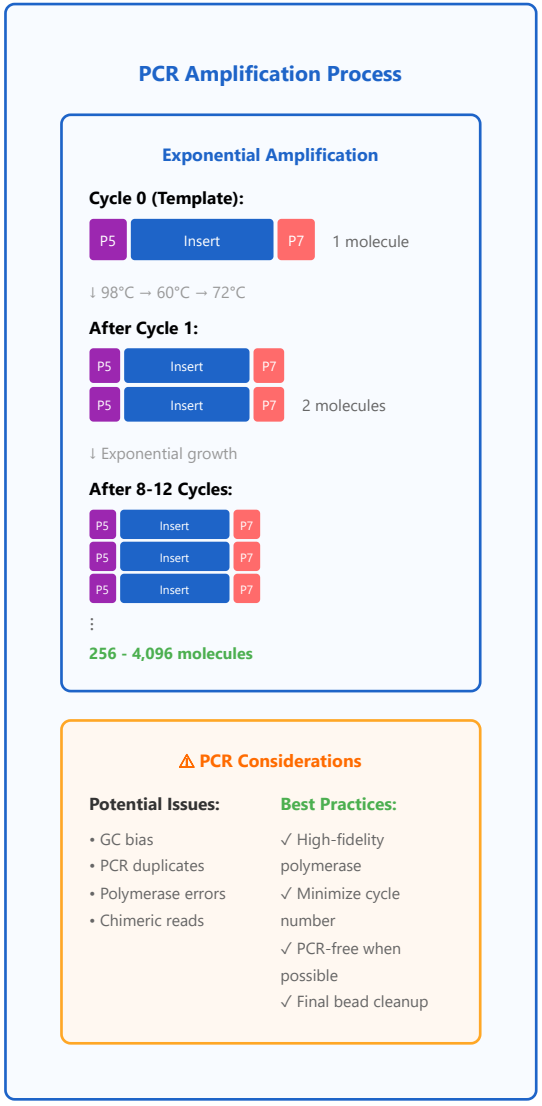
PCR Cycling Parameters

Typical library amplification uses a "limited cycle" protocol (8-12 cycles):

- **Initial Denaturation:** 98°C for 30 seconds
- **Cycling (8-12 cycles):**
  - Denaturation: 98°C for 10 seconds
  - Annealing: 60-65°C for 30 seconds
  - Extension: 72°C for 30 seconds
- **Final Extension:** 72°C for 5 minutes

Cycle Number Optimization

- **High input (>100 ng):** 4-6 cycles
- **Medium input (10-100 ng):** 8-10 cycles
- **Low input (1-10 ng):** 12-15 cycles
- **Very low input (<1 ng):** 15-18 cycles (increased bias risk)



PCR-Free Libraries

When sufficient input DNA is available (typically >1 µg), PCR-free library preparation is preferred because it:

- Eliminates PCR-induced GC bias and errors
- Provides more uniform genome coverage
- Reduces duplicate reads
- Improves variant calling accuracy

⚠ **PCR Artifacts:** Excessive PCR cycles can introduce:

- **GC Bias:** AT-rich regions amplify more efficiently than GC-rich regions
- **PCR Duplicates:** Same molecule amplified multiple times, reducing effective coverage
- **Chimeric Reads:** Template switching during PCR creates artificial rearrangements
- **PCR Errors:** Polymerase errors become permanent and can be mistaken for variants

✓ **Best Practice:**

- Always use high-fidelity polymerase (e.g., KAPA HiFi, Q5)
- Minimize cycle number to the absolute minimum needed
- Perform final bead cleanup (0.8-1.0X) to remove primers
- Quantify library by qPCR (most accurate for NGS)
- Check library quality on Bioanalyzer: expect single peak at ~420-520 bp

Complete Workflow Overview

1

Fragmentation

Break DNA into 200-600 bp fragments

Time: 15-60 min

2

End Repair

Create blunt ends + A-tails

Time: 30-45 min

3

Ligation

Attach indexed adapters

Time: 15-60 min

4

Size Selection

Remove adapter dimers

Time: 15-90 min

5

Amplification

PCR enrichment (8-12 cycles)

Time: 30-60 min

✓ Critical Success Factors

- **High-quality input DNA:** A260/280 ratio 1.8-2.0, no degradation
- **Proper fragment size:** 300-500 bp insert for optimal sequencing
- **Complete adapter removal:** No adapter dimers in final library
- **Minimal PCR cycles:** Reduce bias and maintain library complexity
- **Accurate quantification:** Use qPCR for precise library molarity
- **Quality control at each step:** Bioanalyzer/TapeStation validation

⏱ Total Workflow Time

Manual Protocol: 6-8 hours (1 day)
Automated Protocol: 4-5 hours (same day)
Rapid Protocol (Tagmentation): 90-120 minutes

Final Library Quality Specifications

Fragment Size	Concentration	Adapter Dimers
400-520 bp	**10-50 nM**	**< 5%**
(300-400 bp insert + adapters)	(by qPCR quantification)	(ideally 0%)

Key Takeaways

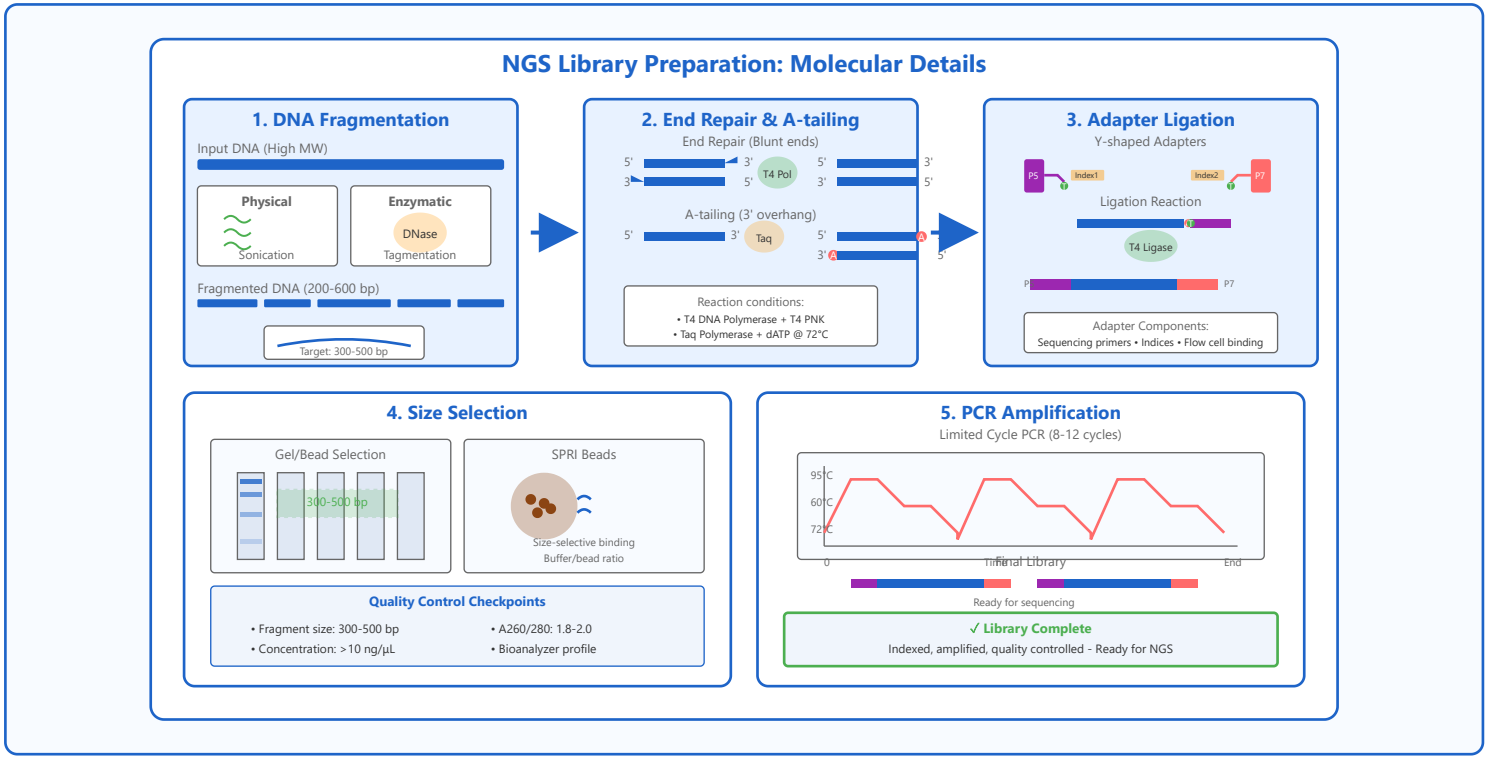
1. Library preparation is a multi-step process that converts genomic DNA into sequencing-ready libraries with adapters, indices, and optimal fragment sizes.

2. Each step requires careful optimization and quality control to ensure high-quality sequencing data with minimal bias.

3. The choice of fragmentation method, cycle number, and size selection strategy significantly impacts library quality and downstream data analysis.

4. Modern library preparation kits have simplified and accelerated the workflow, but understanding the underlying molecular mechanisms is essential for troubleshooting and optimization.

Library Preparation



Critical Factors

- Input DNA quality and quantity
- Fragment size distribution
- Adapter ligation efficiency
- Minimal PCR cycles to avoid bias

Library Types

- Whole genome libraries
- PCR-free libraries (reduce bias)
- Mate-pair libraries (long-range)
- Targeted capture libraries

Comprehensive explanations and diagrams for each libr

Detailed Step-by-Step Guide

1 DNA Fragmentation - Detailed Explanation

DNA fragmentation is the critical first step in NGS library preparation, where high molecular weight genomic DNA is broken into smaller fragments suitable for sequencing platforms. The target fragment size typically ranges from 200-600 base pairs, with most applications optimized for 300-500 bp inserts.

Physical Fragmentation Methods

- **Acoustic Shearing (Covaris):** Uses focused ultrasonic energy to generate controlled cavitation events that shear DNA. This method provides the most uniform size distribution and is highly reproducible.
- **Nebulization:** Forces DNA through a small aperture under high pressure, creating shear forces. Less expensive but offers less control over fragment size distribution.
- **Hydrodynamic Shearing:** DNA is forced through narrow channels, causing mechanical breakage. Provides good control but requires specialized equipment.

Enzymatic Fragmentation

- **Tagmentation (Nextera):** Uses a hyperactive Tn5 transposase that simultaneously fragments DNA and adds adapter sequences. This "tagmentation" process reduces library prep time significantly.
- **DNase I Digestion:** Controlled digestion with DNase I in the presence of Mg^{2+} ions. Fragment size is controlled by enzyme concentration and incubation time.
- **Restriction Enzymes:** Uses specific or frequent-cutting restriction enzymes to create defined breakpoints.

Fragmentation Methods Comparison

Physical Methods

Acoustic Shearing

- Most uniform distribution
- Highly reproducible
- 30-60 min processing
- High equipment cost

Enzymatic Methods

Tagmentation

- One-step process
- Fast (5-15 min)
- Low input DNA (1-50 ng)
- Some sequence bias

Parameter	Acoustic	Enzymatic
Uniformity	Excellent	Good
Speed	30-60 min	5-15 min
DNA Input	100 ng - 5 µg	1-50 ng
Sequence Bias	Minimal	Some bias
Equipment Cost	High ($$$)	Low ($)

2 End Repair & A-tailing - Detailed Ex

After fragmentation, DNA fragments have heterogeneous ends (5' overhangs, 3' overhangs, or blunt ends). End repair converts all fragments to blunt-ended, 5'-phosphorylated molecules suitable for adapter ligation. A-tailing then adds a single adenine (A) nucleotide to the 3' ends of these blunt fragments.

End Repair Process

The end repair reaction uses three enzymatic activities simultaneously:

- **T4 DNA Polymerase:** Fills in 5' overhangs with its 5'→3' polymerase activity and removes 3' overhangs with its 3'→5' exonuclease activity.
- **Klenow Fragment:** Fills in 5' overhangs and provides additional 3'→5' exonuclease activity.
- **T4 Polynucleotide Kinase:** Phosphorylates 5' ends, which is essential

End Repair & A-tai

Step 1: End

Before Repair (5' overhang):

5'
3'
3' 5'-P

↓ T4 DNA Polymerase

5'
3'
3'
5'-P

Enzyme Cocktail: T4 DNA Fragment • T4 PNK

Step 2: A-t

Adding 3' A-o

5'-P 3'
3'-OH

↓ Klenow (exo-) + dATP

5'-P A
3' T

✓ **Result:** 3' A-overhang comp adapters for effici

⚠ **Critical Consideration:** The fragmentation method impacts downstream bias. Acoustic shearing provides the most random fragmentation, while enzymatic methods may show sequence-specific preferences.

✓ **Best Practice:** Always validate fragment size distribution using a Bioanalyzer, TapeStation, or Fragment Analyzer before proceeding to the next step.

for subsequent ligation reactions.

A-tailing Purpose and Mechanism

A-tailing adds a single deoxyadenosine to the 3' end of blunt fragments using Klenow fragment (3'→5' exo-minus) or Taq polymerase. This creates a T overhang that is complementary to the T overhangs on adapter molecules, ensuring proper directional ligation.

- **Prevents Adapter Dimer Formation:** Only fragments with A-tails can ligate to T-overhang adapters.
- **Increases Ligation Efficiency:** T-A base pairing is more stable than blunt-end ligation.
- **Directional Cloning:** Ensures adapters ligate in the correct orientation.

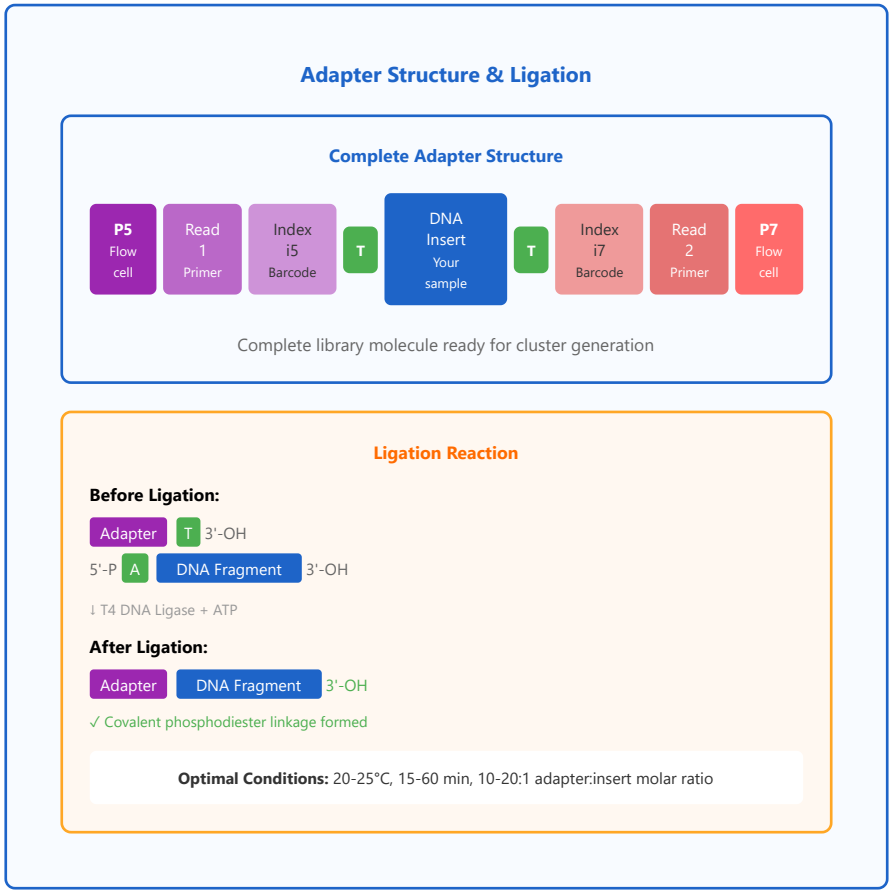
⚠ **Temperature Sensitivity:** End repair is typically performed at 20-25°C, while A-tailing requires 37-72°C depending on the enzyme. Improper

temperatures lead to incomplete reactions.

✓ **Quality Check:** Incomplete end repair or A-tailing dramatically reduces library yield. Some kits now offer combined end repair/A-tailing reactions to streamline the workflow and reduce sample loss.

3 Adapter Ligation - Detailed Explanation

Adapter ligation is the process of attaching synthetic oligonucleotide adapters to both ends of the prepared DNA fragments. These adapters contain several critical elements necessary for NGS sequencing and are the defining feature that converts fragmented DNA into a "sequencing library."

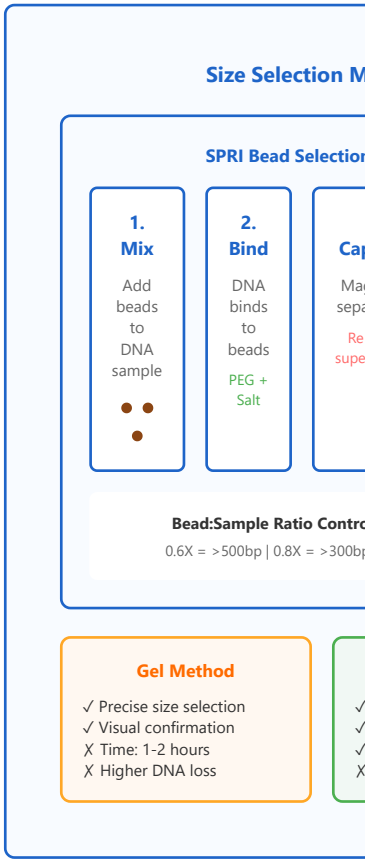


Adapter Structure and Components

- **P5 and P7 Binding Sites:** Complementary sequences to oligonucleotides on the flow cell surface, enabling cluster generation during bridge amplification.
- **Sequencing Primer Binding Sites:** Allow

4 Size Selection - Detailed Explanatio

Size selection is a critical quality control step that removes unwanted DNA fragments from the library preparation, including adapter dimers, very short fragments, and excessively long fragments. Proper size selection ensures optimal cluster density, sequencing quality, and data output.



Methods for Size Selection

1. SPRI Beads (AMPure XP)

SPRI beads are paramagnetic beads coated with carboxyl groups that reversibly bind DNA in the presence of polyethylene glycol (PEG) and salt. This is currently the most popular method.

- **Mechanism:** DNA binding efficiency depends on the bead:sample volume ratio.

sequencing primers to hybridize and initiate sequencing reactions.

- **Index Sequences (Barcodes):** 6-10 bp unique identifiers that enable sample multiplexing. Can be single-indexed (one barcode) or dual-indexed (two barcodes for added specificity).
- **T-overhangs:** Single thymine nucleotides at the 3' ends that complement the A-tails on DNA fragments.

Ligation Chemistry

T4 DNA Ligase catalyzes the formation of phosphodiester bonds between the 3'-OH of the A-tailed insert and the 5'-phosphate of the T-overhang

- **Size selection strategy:**
  - 0.6X bead ratio: Selects fragments >500 bp
  - 0.8X bead ratio: Selects fragments >300 bp
  - 1.8X bead ratio: Selects fragments >100 bp
  - Double-sided selection (0.5X then 0.7X): Narrow range (300-500 bp)
- **Advantages:** Fast (15-20 min), scalable, minimal DNA loss

2. Gel Electrophoresis

Traditional agarose gel electrophoresis provides the most precise size selection but is labor-intensive.

- **Procedure:** Run library on 2% agarose gel, visualize, excise target band (400-500 bp), purify

adapter. The reaction:

- Requires ATP as a cofactor
- Is typically performed at 20-25°C for 15-60 minutes
- Benefits from PEG (polyethylene glycol) to create molecular crowding and increase effective concentration

Multiplexing Strategy

Index sequences enable pooling of multiple samples in a single sequencing run. During data analysis, reads are "demultiplexed" based on their index sequences to assign them back to individual samples. Dual indexing (i5 and i7 indices) provides additional accuracy and can correct for index hopping in patterned flow cells.

DNA from gel slice

- **Advantages:** Visual confirmation, precise selection, complete adapter dimer removal
- **Disadvantages:** Time-consuming (1-2 hours), higher DNA loss

⚠ Adapter Dimer Problem: Adapter dimers (~120 bp) are highly problematic because they:

- Compete with insert-containing molecules during cluster generation
- Sequence more efficiently than longer fragments
- Can consume 20-80% of sequencing capacity if not removed
- Provide zero useful data

⚠ **Adapter Dimers:** A major challenge is adapter-adapter ligation without insert DNA, creating short molecules (~120 bp) that sequence efficiently but provide no useful data. Proper insert:adapter ratios and size selection steps are critical to minimize dimers.

✓ **Optimization Tip:** Use a 10-20 molar excess of adapters to DNA fragments to ensure complete ligation while minimizing adapter dimers. Always verify complete removal of the ~120 bp adapter dimer peak before sequencing!

✓ **Quality Control:** Use a Bioanalyzer, TapeStation, or Fragment Analyzer to verify:

- Complete removal of adapter dimers (~120 bp peak)
- Tight size distribution around target size (e.g., 400-500 bp)
- Adequate DNA concentration for sequencing (typically >2-5 nM)

include a size selection step post-ligation to remove excess unligated adapters and adapter dimers.

5 PCR Amplification - Detailed Explanation

PCR amplification is the final step of library preparation, enriching the adapter-ligated DNA fragments to generate sufficient material for sequencing. This step must be carefully optimized to achieve adequate library concentration while minimizing amplification bias and artifacts.

Purpose of Library Amplification

- **Increase DNA Quantity:** Generate enough library material (typically 10-50 nM in 25-50 µL) for accurate quantification and sequencing
- **Enrich Properly Ligated Molecules:** Selectively amplify fragments with adapters on both ends
- **Add Flow Cell Binding Sites:** Incorporate complete P5 and P7 sequences needed for cluster generation
- **Introduce Additional Indices:** Some protocols add sample indices during PCR rather than ligation

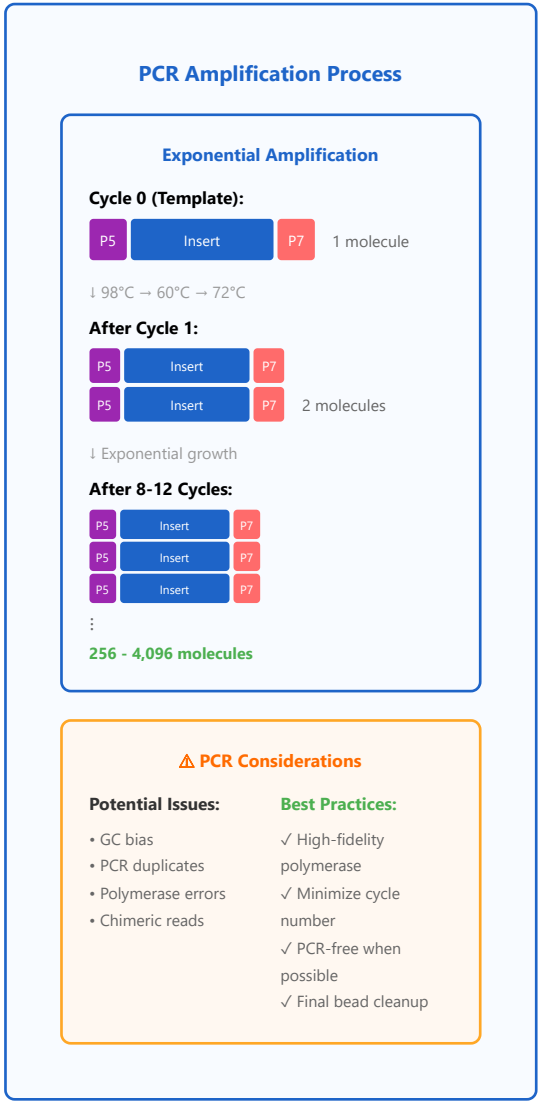
PCR Cycling Parameters

Typical library amplification uses a "limited cycle" protocol (8-12 cycles):

- **Initial Denaturation:** 98°C for 30 seconds
- **Cycling (8-12 cycles):**
  - Denaturation: 98°C for 10 seconds
  - Annealing: 60-65°C for 30 seconds
  - Extension: 72°C for 30 seconds
- **Final Extension:** 72°C for 5 minutes

Cycle Number Optimization

- **High input (>100 ng):** 4-6 cycles
- **Medium input (10-100 ng):** 8-10 cycles
- **Low input (1-10 ng):** 12-15 cycles
- **Very low input (<1 ng):** 15-18 cycles (increased bias risk)



PCR-Free Libraries

When sufficient input DNA is available (typically >1 µg), PCR-free library preparation is preferred because it:

- Eliminates PCR-induced GC bias and errors
- Provides more uniform genome coverage
- Reduces duplicate reads
- Improves variant calling accuracy

⚠ **PCR Artifacts:** Excessive PCR cycles can introduce:

- **GC Bias:** AT-rich regions amplify more efficiently than GC-rich regions
- **PCR Duplicates:** Same molecule amplified multiple times, reducing effective coverage
- **Chimeric Reads:** Template switching during PCR creates artificial rearrangements
- **PCR Errors:** Polymerase errors become permanent and can be mistaken for variants

✓ **Best Practice:**

- Always use high-fidelity polymerase (e.g., KAPA HiFi, Q5)
- Minimize cycle number to the absolute minimum needed
- Perform final bead cleanup (0.8-1.0X) to remove primers
- Quantify library by qPCR (most accurate for NGS)
- Check library quality on Bioanalyzer: expect single peak at ~420-520 bp

Complete Workflow Overview

1

Fragmentation

Break DNA into 200-600 bp fragments

Time: 15-60 min

2

End Repair

Create blunt ends + A-tails

Time: 30-45 min

3

Ligation

Attach indexed adapters

Time: 15-60 min

4

Size Selection

Remove adapter dimers

Time: 15-90 min

5

Amplification

PCR enrichment (8-12 cycles)

Time: 30-60 min

✓ Critical Success Factors

- **High-quality input DNA:** A260/280 ratio 1.8-2.0, no degradation
- **Proper fragment size:** 300-500 bp insert for optimal sequencing
- **Complete adapter removal:** No adapter dimers in final library
- **Minimal PCR cycles:** Reduce bias and maintain library complexity
- **Accurate quantification:** Use qPCR for precise library molarity
- **Quality control at each step:** Bioanalyzer/TapeStation validation

⏱ Total Workflow Time

Manual Protocol: 6-8 hours (1 day)
Automated Protocol: 4-5 hours (same day)
Rapid Protocol (Tagmentation): 90-120 minutes

Final Library Quality Specifications

Fragment Size	Concentration	Adapter Dimers
400-520 bp	**10-50 nM**	**< 5%**
(300-400 bp insert + adapters)	(by qPCR quantification)	(ideally 0%)

Key Takeaways

1. Library preparation is a multi-step process that converts genomic DNA into sequencing-ready libraries with adapters, indices, and optimal fragment sizes.

2. Each step requires careful optimization and quality control to ensure high-quality sequencing data with minimal bias.

3. The choice of fragmentation method, cycle number, and size selection strategy significantly impacts library quality and downstream data analysis.

4. Modern library preparation kits have simplified and accelerated the workflow, but understanding the underlying molecular mechanisms is essential for troubleshooting and optimization.

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

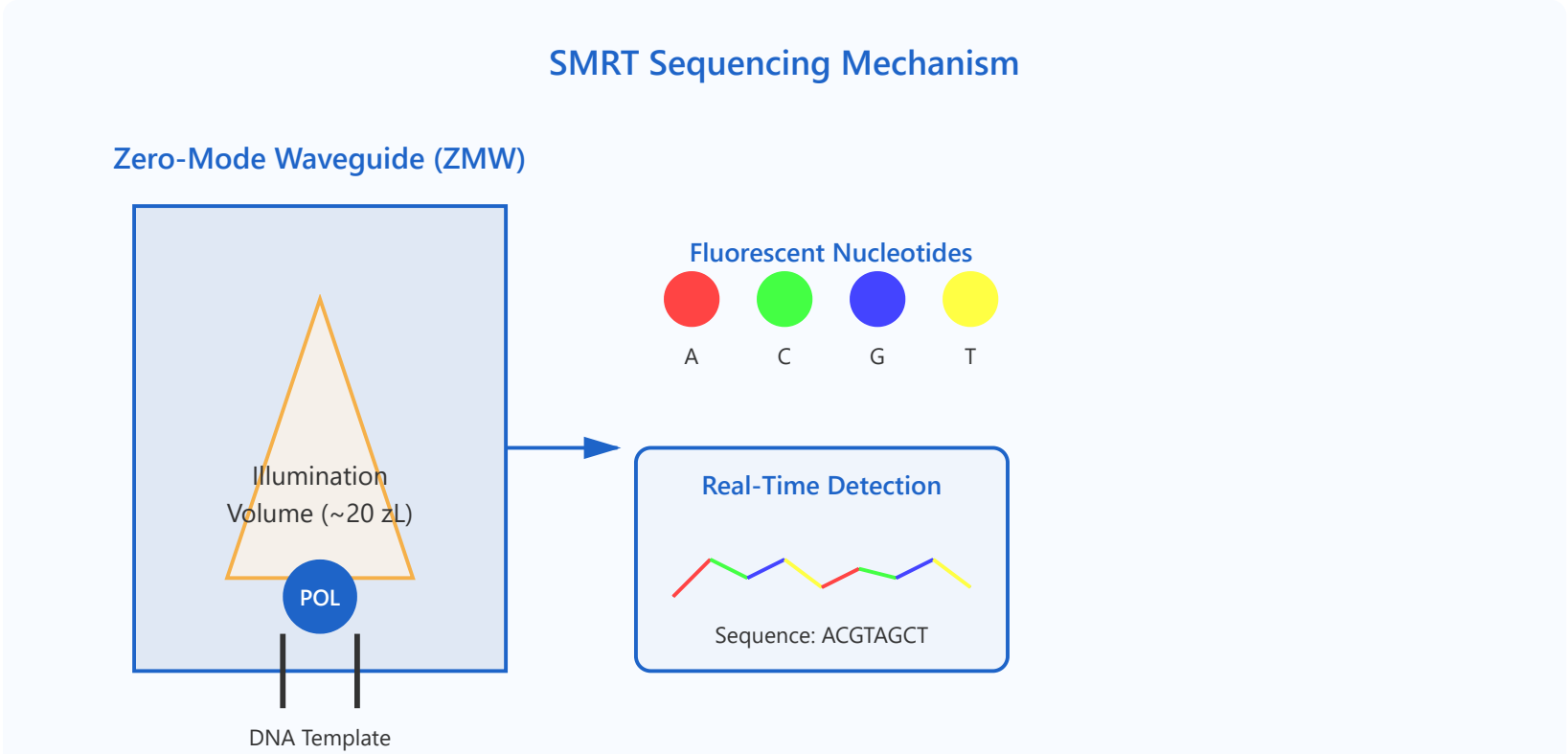
- Single Molecule Real-Time (SMRT) sequencing
- Watches DNA polymerase in real-time
- Zero-mode waveguides (ZMWs) for detection

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)

1. Single Molecule Real-Time (SMRT) Sequencing



How SMRT Sequencing Works: The technology uses zero-mode waveguides (ZMWs), which are tiny wells with a diameter of approximately 70 nanometers at the bottom of the well. Light is directed through the bottom, creating an illumination volume of only about 20 zeptoliters (10^{-21} liters). This extremely small detection volume allows observation of single DNA polymerase molecules at work.

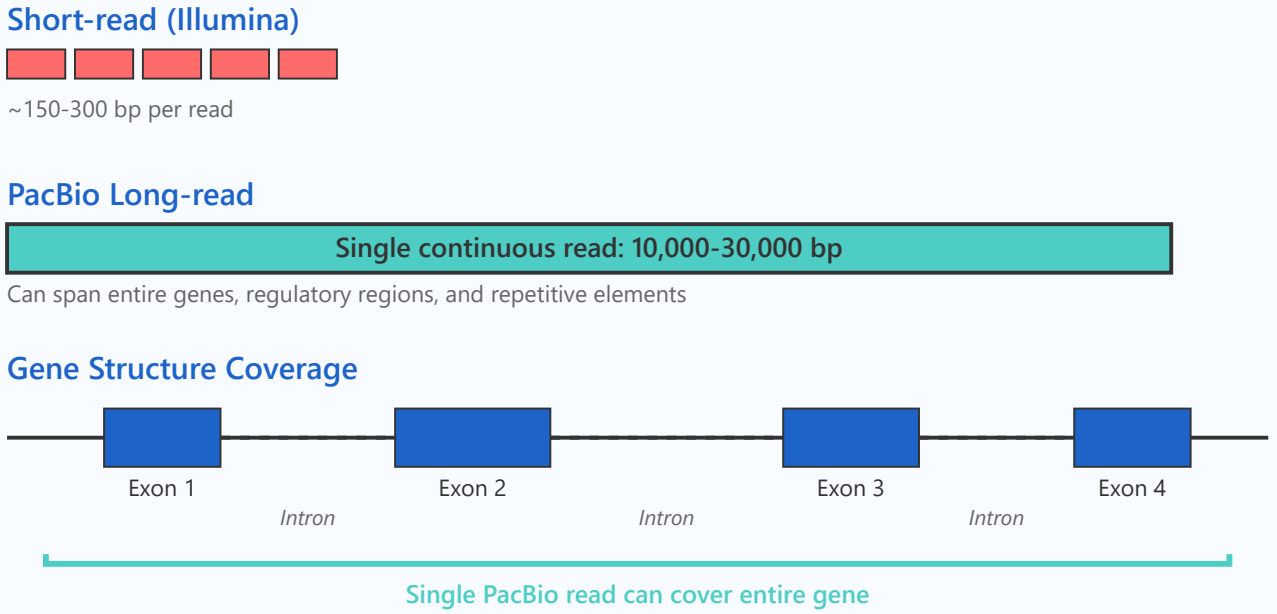
Each of the four DNA bases (A, C, G, T) is attached to a different fluorescent dye. When the polymerase incorporates a nucleotide into the growing DNA strand, the fluorescent label emits a light pulse that is detected in real-time. The dye is then cleaved off, allowing the next nucleotide to be incorporated without interference.

Example Application:

During sequencing of a bacterial genome, the SMRT system can continuously monitor a single DNA polymerase for several hours, generating reads of 10-30 kilobases without interruption. The real-time nature means that the sequencing speed is only limited by the natural rate of DNA polymerase (approximately 10 nucleotides per second).

2. Long Read Length (10-30 Kb)

Read Length Comparison



Advantages of Long Reads: The extended read length of PacBio sequencing (10-30 kb compared to 150-300 bp for Illumina) provides several critical advantages. Long reads can span entire genes including all exons and introns, making it possible to determine full-length transcript isoforms without computational assembly.

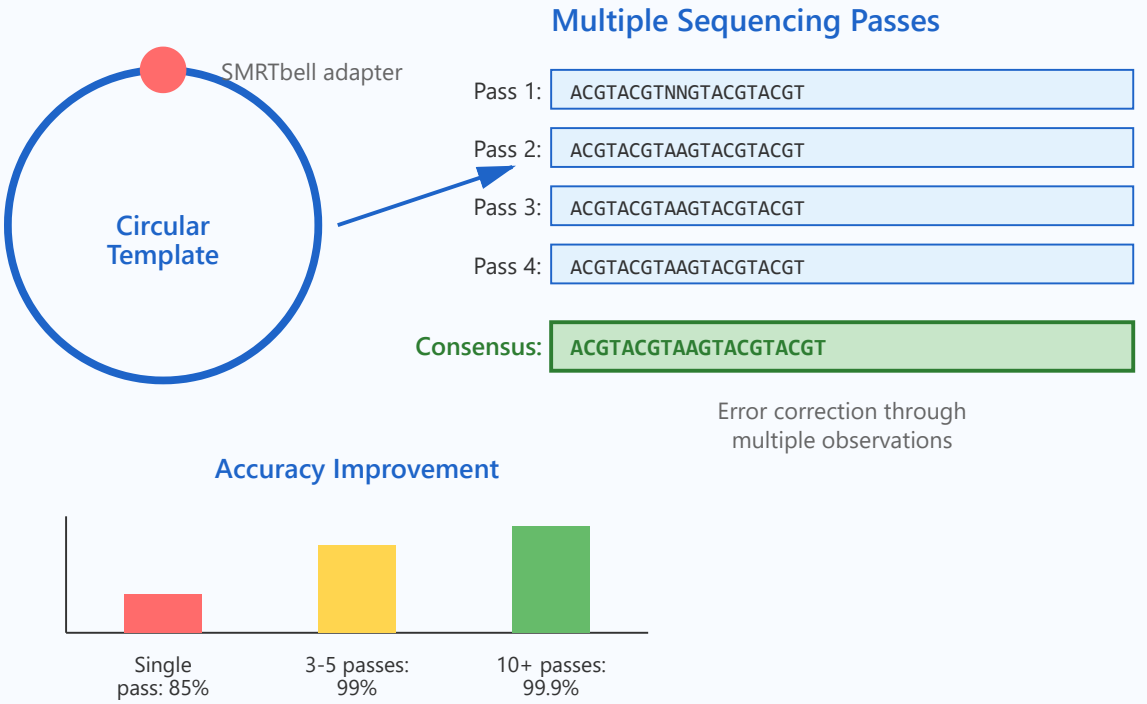
Long reads are particularly valuable for resolving complex genomic regions such as segmental duplications, tandem repeats, and other repetitive sequences that are typically fragmented or misassembled with short-read technologies. This capability significantly improves genome assembly contiguity and reduces the number of gaps in assembled genomes.

Example Application:

The human dystrophin gene (DMD) spans approximately 2.4 million base pairs with 79 exons. Short-read sequencing would require computational assembly of thousands of reads to reconstruct this gene, with potential for errors in repetitive regions. A single PacBio HiFi read can span multiple exons continuously, enabling direct detection of splicing patterns, structural variants, and disease-causing mutations without assembly artifacts.

3. High Accuracy (99.9% HiFi)

Circular Consensus Sequencing (CCS) for HiFi Reads



HiFi Technology: PacBio HiFi (High Fidelity) reads combine the advantages of long read lengths with high accuracy through Circular Consensus Sequencing (CCS). The DNA insert is ligated to hairpin adapters creating a circular template called a SMRTbell. The polymerase continuously sequences this circular template multiple times, generating multiple passes over the same DNA molecule.

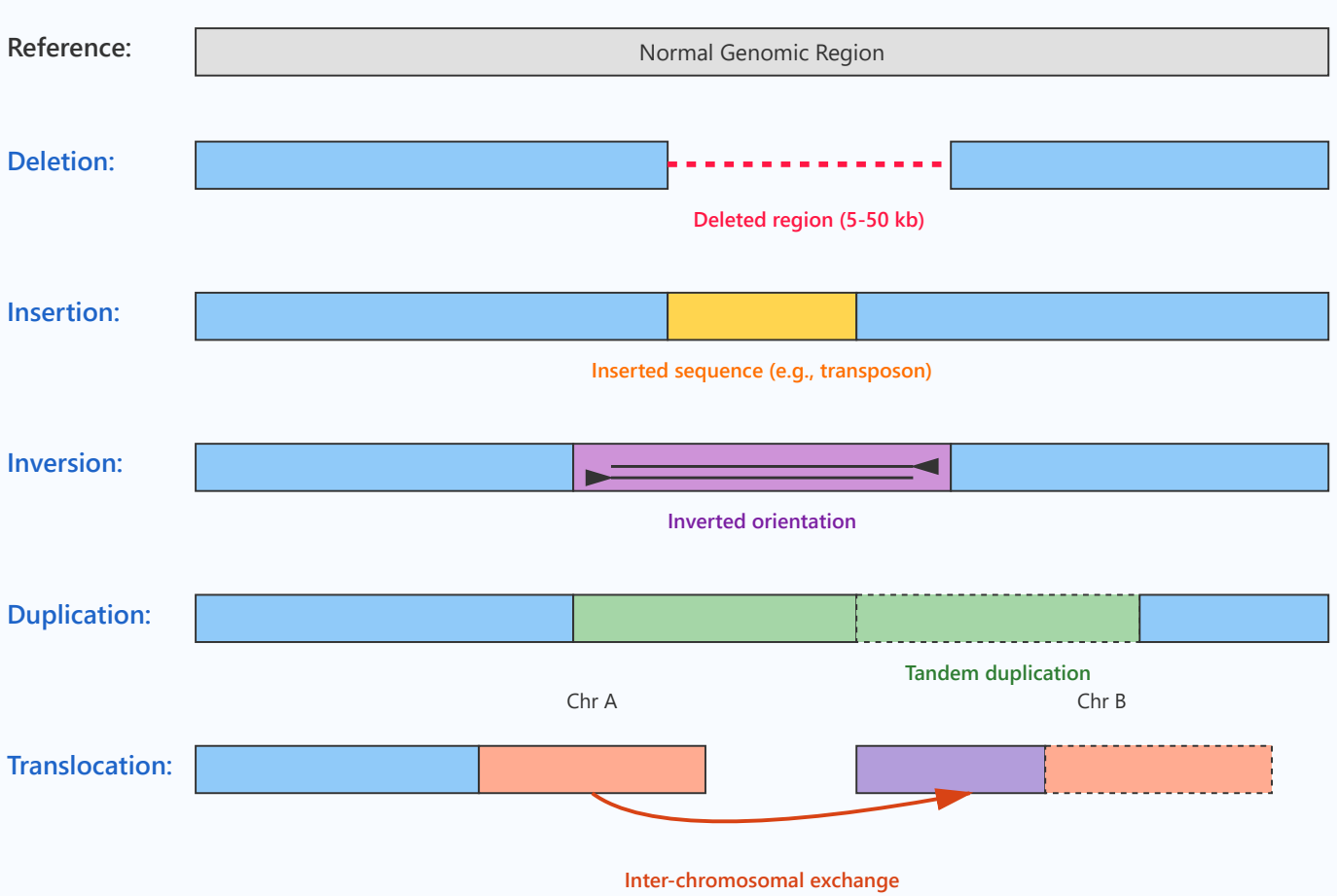
Each pass may contain random errors, but by sequencing the same molecule 10-20 times, these errors can be identified and corrected through consensus calling. The final HiFi read achieves greater than 99.9% (Q30) accuracy while maintaining read lengths of 10-25 kb, matching or exceeding the accuracy of short-read platforms while preserving long-range information.

Example Application:

In clinical diagnostics, HiFi sequencing enables accurate detection of single nucleotide variants (SNVs) and small insertions/deletions (indels) across complex genomic regions. For example, HiFi reads can accurately sequence through the highly polymorphic HLA genes (human leukocyte antigen) which are critical for transplant matching, resolving both allelic variations and structural differences with clinical-grade accuracy.

4. Structural Variant Detection

Types of Structural Variants Detected by Long Reads



Structural Variant Detection Capabilities: Long reads excel at detecting structural variants (SVs) that are difficult or impossible to identify with short reads. SVs include deletions, insertions, inversions, duplications, and translocations that affect segments typically larger than 50 base pairs. These variants play crucial roles in genome evolution, genetic disease, and cancer development.

PacBio reads can span entire structural variants, including their breakpoints, providing direct evidence of the variant structure. This is particularly valuable in repetitive regions where short reads cannot uniquely map.

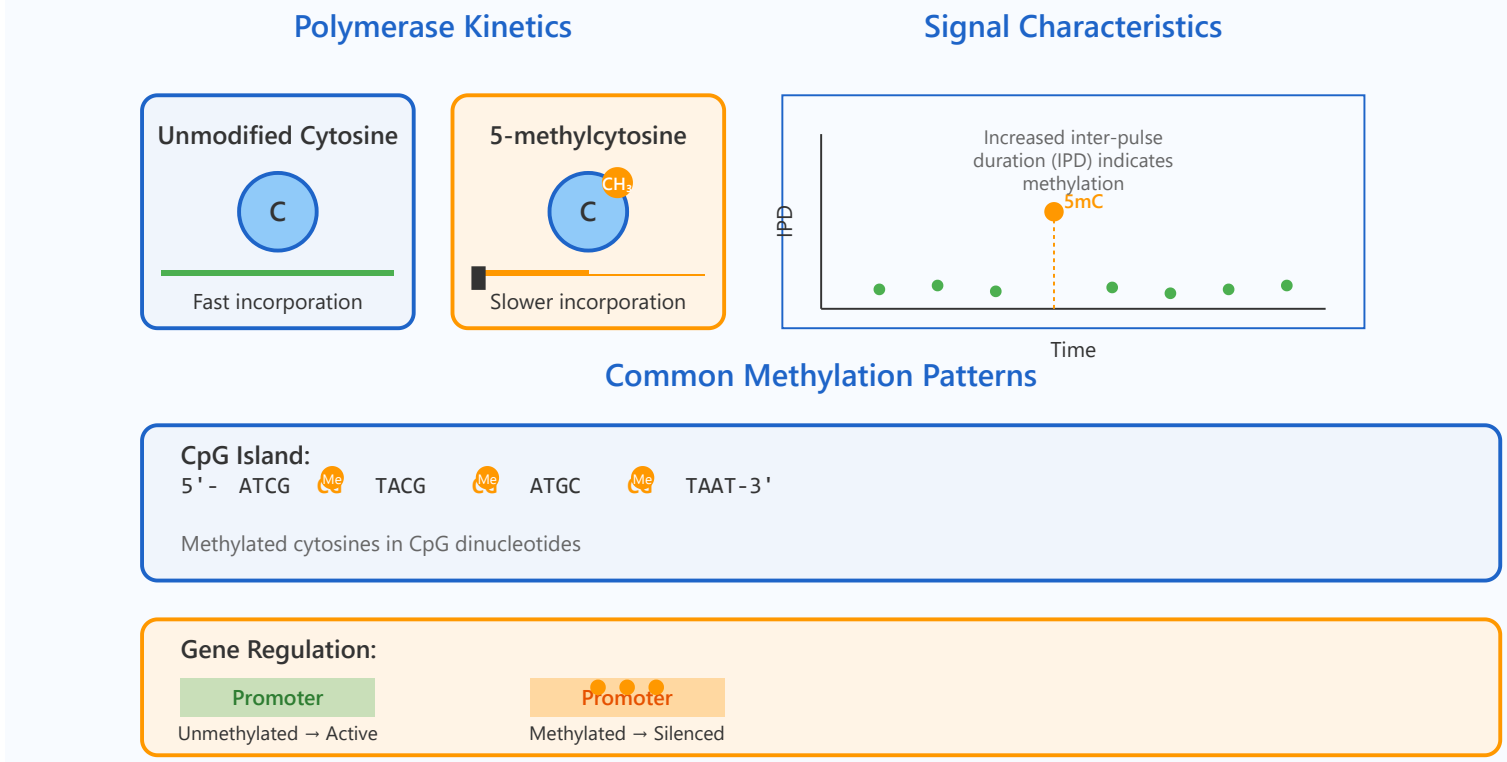
Long reads can also resolve complex rearrangements involving multiple events and determine their phase on individual chromosomes.

Example Application:

In cancer genomics, structural variants such as gene fusions drive oncogenesis. For example, the BCR-ABL1 fusion in chronic myeloid leukemia results from a translocation between chromosomes 9 and 22. PacBio long reads can span the fusion breakpoint, precisely determining the exact junction sequence and any additional complexity (e.g., insertions or deletions at the breakpoint) that may affect treatment response to tyrosine kinase inhibitors.

5. Native Base Modification Detection (Methylation)

Direct Detection of DNA Methylation



Direct Methylation Detection: Unlike other sequencing platforms that require chemical conversion (bisulfite sequencing) to detect methylation, PacBio directly detects DNA modifications during sequencing. Modified bases, such as 5-methylcytosine (5mC) and N6-methyladenine (6mA), affect the kinetics of DNA polymerase incorporation.

The SMRT system measures the inter-pulse duration (IPD), which is the time between successive nucleotide incorporations. When the polymerase encounters a modified base, it pauses slightly longer, resulting in an increased IPD. By analyzing these kinetic signatures across the genome, PacBio can simultaneously determine DNA sequence and map methylation patterns without additional sample preparation or loss of sequence information.

Example Application:

In cancer epigenetics research, aberrant DNA methylation patterns are hallmarks of tumorigenesis. PacBio sequencing can identify hypermethylation of tumor suppressor gene promoters (such as BRCA1 or MLH1) that leads to gene silencing without requiring separate methylation assays. This integrated approach enables researchers to correlate structural variants, sequence mutations, and epigenetic modifications in a single experiment, providing a comprehensive view of cancer genome architecture.

Feature	Bisulfite Sequencing	PacBio Native Detection
Sample preparation	Chemical conversion required	No conversion needed
DNA degradation	Significant (~90% loss)	No degradation
Read length	Reduced due to treatment	Full long-read length maintained
Modification types	5mC only	5mC, 6mA, and other modifications
Sequence context	C/T ambiguity	Original sequence preserved
Phasing information	Lost in short reads	Maintained across long reads

Summary: PacBio SMRT Sequencing Advantages

PacBio's Single Molecule Real-Time (SMRT) sequencing technology represents a paradigm shift in genomics by providing long, accurate reads with native modification detection. The key advantages include:

Long-Range Information

Spanning 10-30 kb enables resolution of complex genomic regions, complete gene structures, and haplotype phasing

High Accuracy

HiFi reads achieve >99.9% accuracy through multiple sequencing passes, suitable for clinical applications

Structural Variant Detection

Direct observation of large insertions, deletions, inversions, and complex rearrangements

Epigenetic Profiling

Simultaneous detection of DNA sequence and modifications without additional library preparation

PacBio technology is particularly valuable for de novo genome assembly, characterization of complex genetic diseases, cancer genomics, and microbiome research where comprehensive genomic information is essential.

Nanopore Sequencing Technology

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

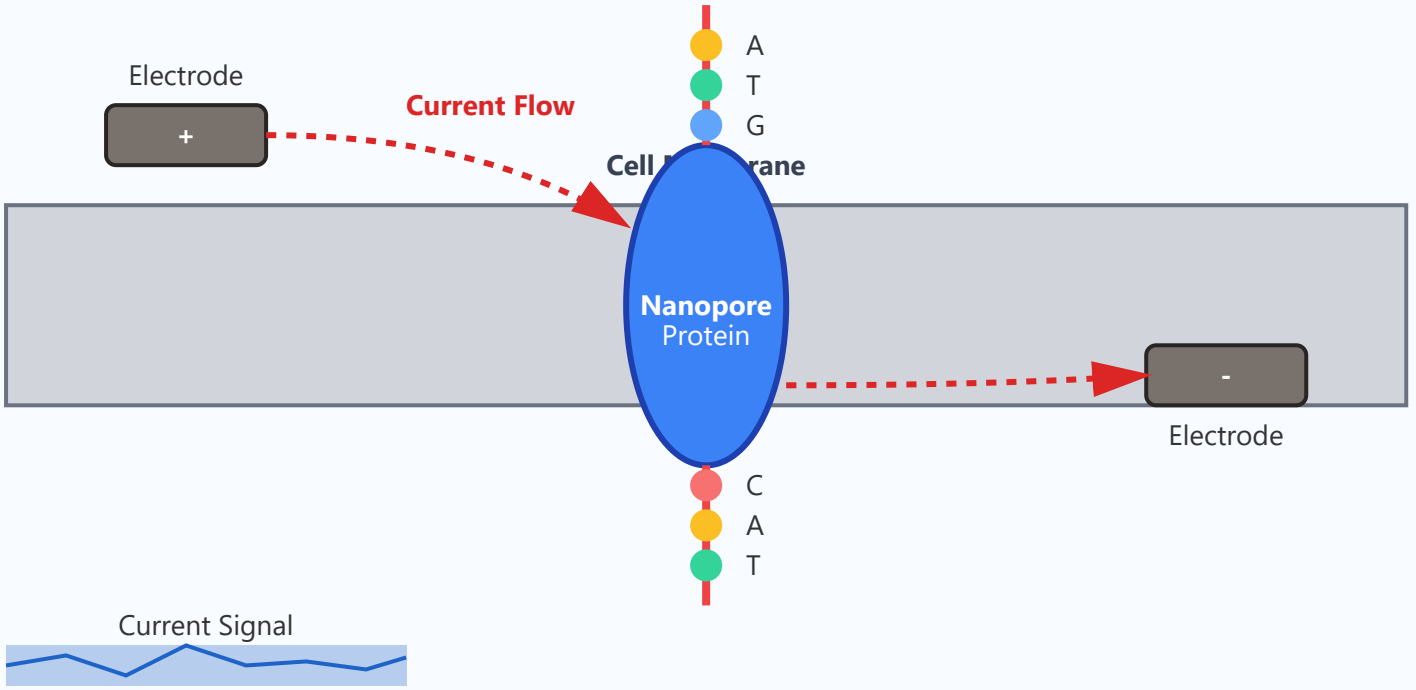
Accuracy

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

1. TECHNOLOGY PRINCIPLE - DETAILED EXPLANATION



How It Works

Nanopore sequencing is a revolutionary single-molecule sequencing technology that reads DNA or RNA sequences directly by measuring changes in electrical current as nucleic acids pass through a protein nanopore.

Step-by-Step Process:

1. A motor protein unwinds the double-stranded DNA and feeds single-stranded DNA through the nanopore at a controlled speed (approximately 450 bases per second).

2. An ionic current is applied across the membrane. As each nucleotide passes through the nanopore, it causes a characteristic disruption to the current.

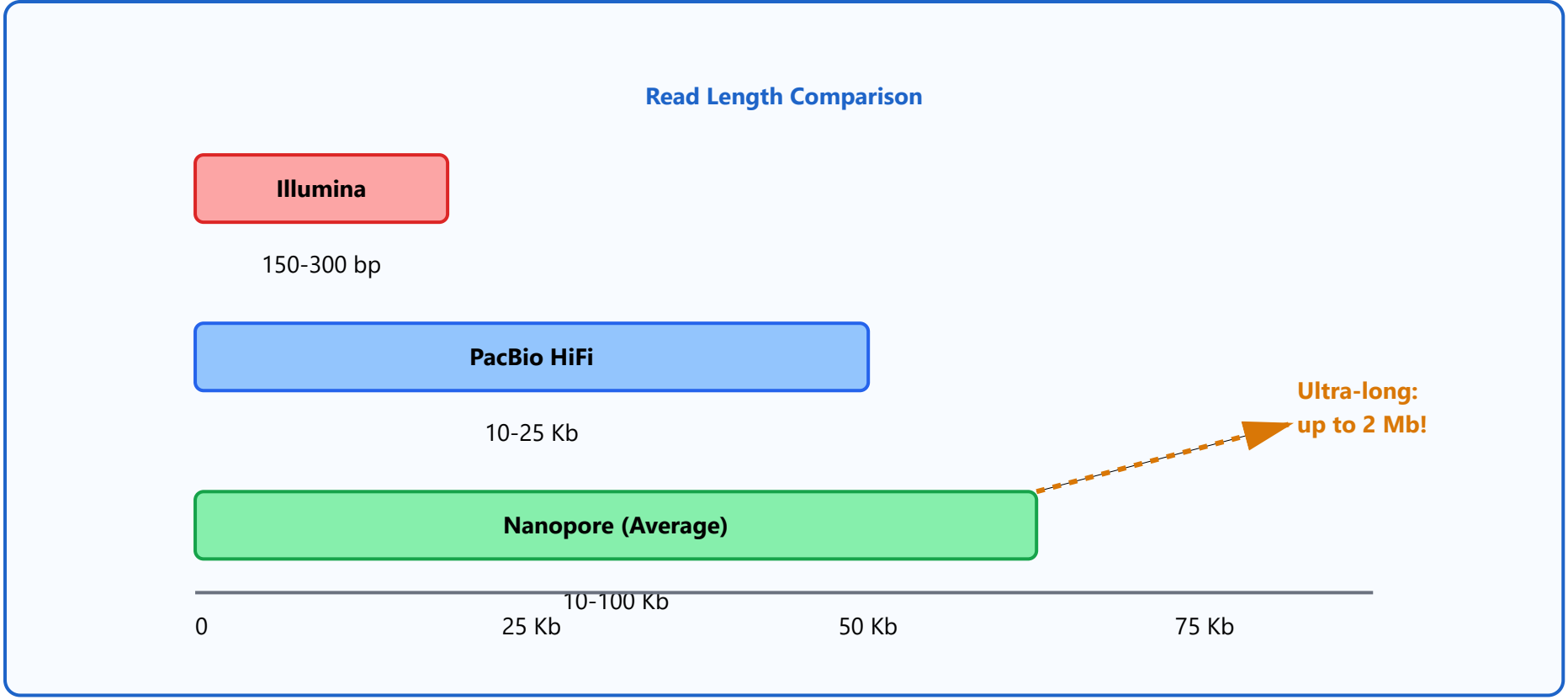
3. Each of the four DNA bases (A, T, G, C) has a unique chemical structure that blocks the current differently, creating a distinctive electrical signature.

4. Advanced algorithms analyze these current changes in real-time to identify the sequence of bases passing through the pore.

Key Advantages of the Principle

→ No optical detection or fluorescent labels required
→ No amplification step needed - sequences native DNA/RNA directly
→ No theoretical limit on read length - depends only on DNA fragment length
→ Can detect modified bases (methylation, etc.) in their native form
→ Real-time data streaming enables immediate analysis

2. READ LENGTH CAPABILITIES



Why Read Length Matters

Read length is one of the most critical parameters in DNA sequencing, and nanopore technology excels in this area with dramatically longer reads compared to traditional methods.

Applications Enabled by Long Reads:

Application	Why Long Reads Help	Nanopore Advantage
De novo Assembly	Span repetitive regions that are longer than short reads	Can assemble entire bacterial genomes in single contigs

Structural Variants	Detect large deletions, insertions, inversions	Reads can span entire variant, making detection unambiguous
Phasing	Connect distant variants on same chromosome	Ultra-long reads can phase across megabases
Isoform Detection	Sequence full-length transcripts	Direct RNA sequencing of complete mRNA molecules
Repeat Analysis	Resolve complex repetitive regions	Can sequence through centromeres and telomeres

Record-Breaking Achievements

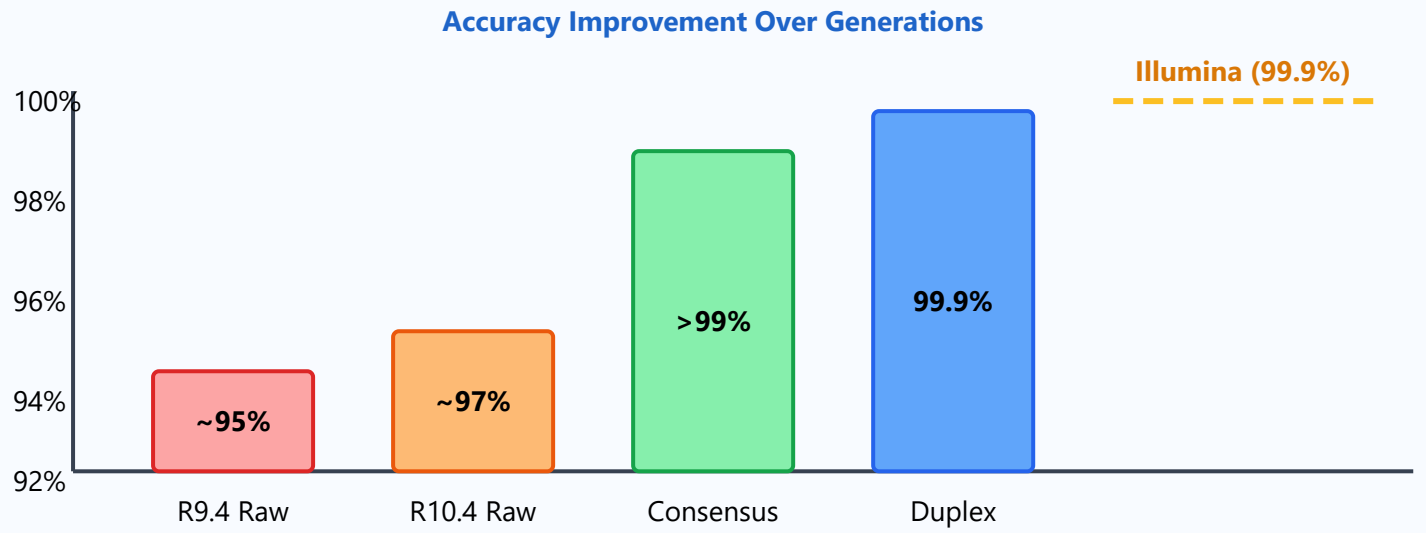
→ Longest single read: >4 Mb (megabases) reported

→ N50 values routinely exceed 50 Kb with proper DNA preparation

→ Ultra-long read protocols can achieve >100 Kb average read length

→ Enabled complete gapless assemblies of human chromosomes

3. SEQUENCING ACCURACY



Understanding Accuracy Metrics

Nanopore sequencing accuracy has dramatically improved over the years through better chemistry, improved nanopores, and advanced basecalling algorithms powered by deep learning.

Types of Accuracy:

1. Raw Read Accuracy (~95-97%): Single-pass sequencing of one DNA strand. The R10.4 chemistry with improved nanopores and basecallers like Dorado achieve ~97% raw accuracy.

2. Consensus Accuracy (>99%): Multiple reads of the same DNA molecule are combined using consensus algorithms. This requires higher coverage (typically 20-50×) but achieves accuracy comparable to short-read platforms.

3. Duplex Accuracy (99.9%): Both strands of the same DNA molecule are sequenced and compared. This provides the highest accuracy, matching Illumina's gold standard, while maintaining long read lengths.

Factors Affecting Accuracy

→ **Pore Chemistry:** R10.4 pores provide better resolution than earlier R9.4 versions

→ **Basecalling Algorithm:** Deep learning models (Guppy, Dorado) significantly improve accuracy

→ **Sequencing Speed:** Slower translocation gives more data points per base

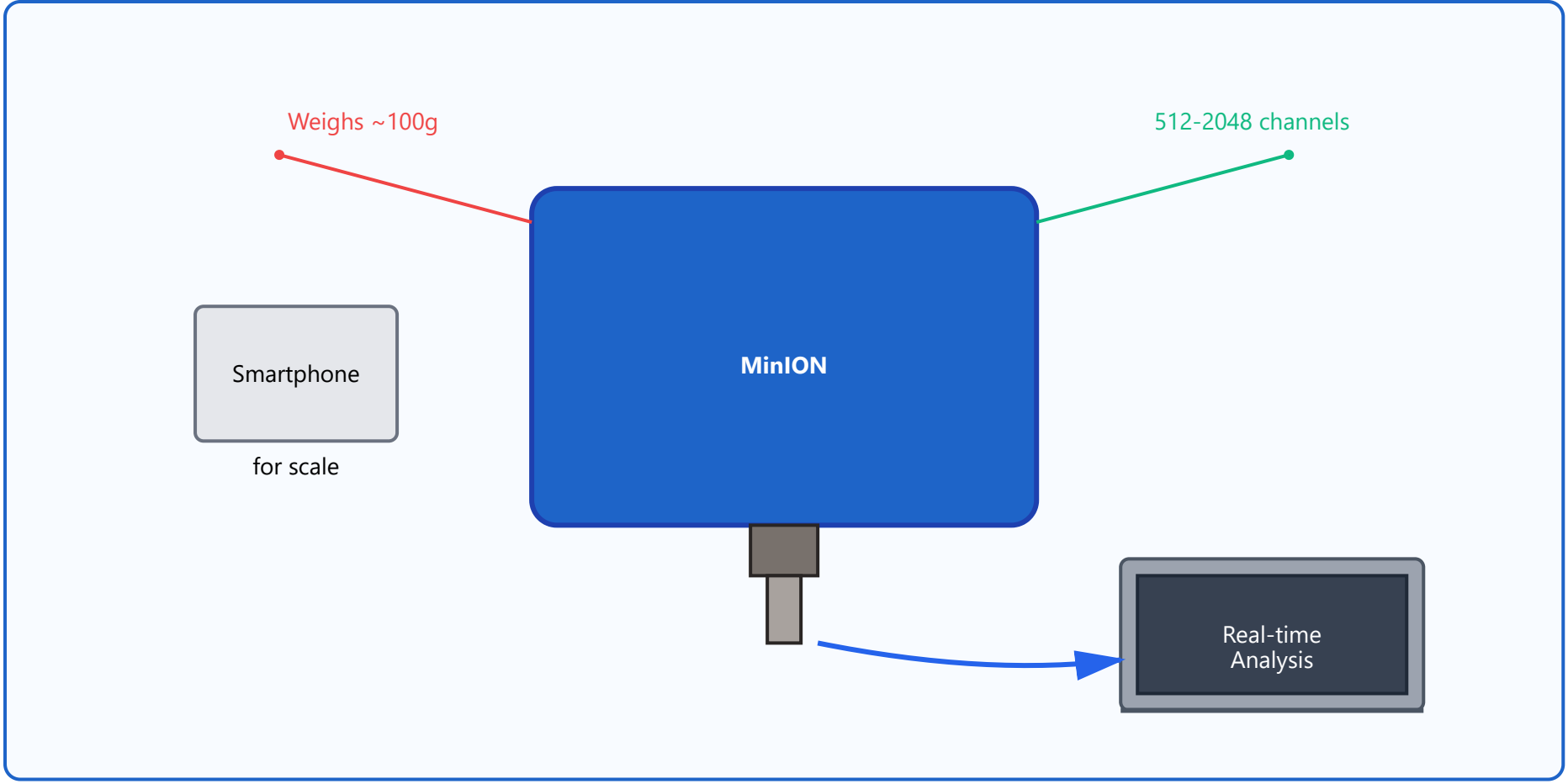
→ **DNA Quality:** High molecular weight, pure DNA yields better results

→ **Homopolymers:** Stretches of same base (e.g., AAAAA) remain challenging

Error Profile

Unlike Illumina sequencing which has systematic errors and GC bias, nanopore errors are more random and can be corrected through consensus. The error rate is higher in homopolymer regions but more uniform across different sequence contexts.

4. PORTABLE DEVICE - MINION USB SEQUENCER



Revolutionary Portability

The MinION is the world's first portable DNA sequencer, roughly the size of a USB thumb drive. This breakthrough in miniaturization has transformed where and how sequencing can be performed.

Device Specifications:

Feature	MinION	Traditional Sequencer
Size	10 cm × 3.2 cm × 2 cm	Desktop to room-sized

Weight	~100 grams	50-500+ kg
Cost	$1,000 device (flow cells $500-900)	$100,000 - $1,000,000+
Power	USB-powered (5W)	Dedicated power (1000W+)
Setup	Plug-and-play, <10 minutes	Specialized facility required
Throughput	Up to 50 Gb per flow cell	100 Gb - 6 Tb per run

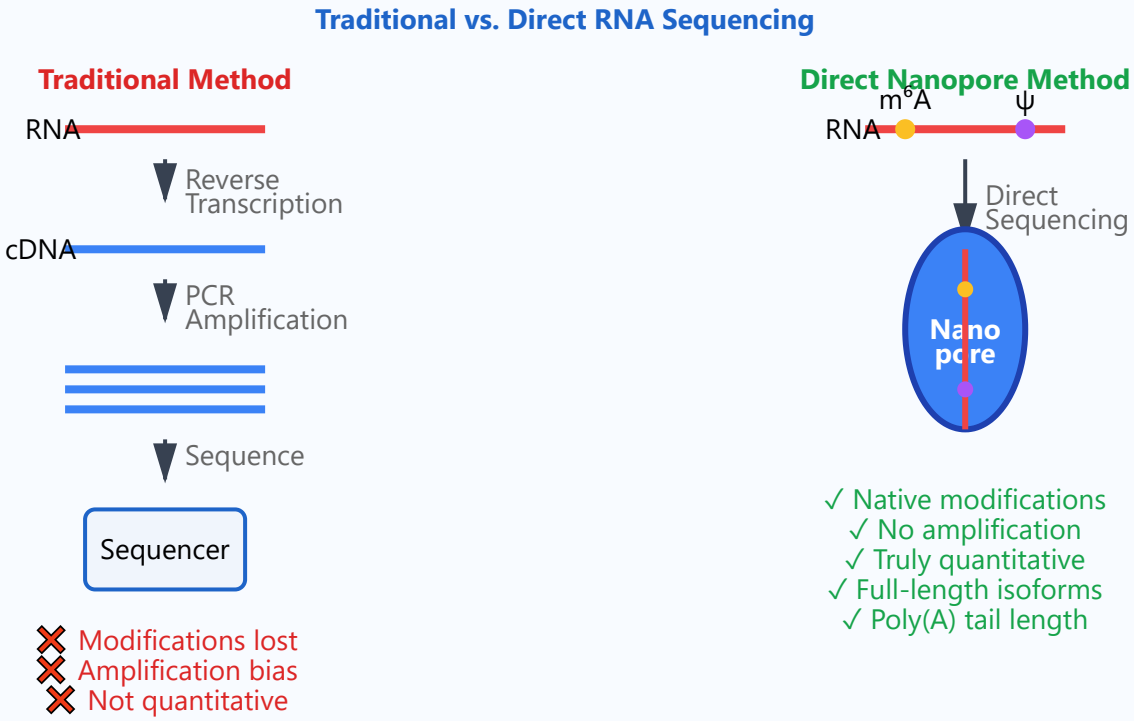
Game-Changing Applications

→ **Field Research:** Used in Antarctic, rainforests, and remote locations for real-time pathogen detection
→ **Outbreak Response:** Deployed during Ebola, Zika, and COVID-19 outbreaks for rapid viral genome sequencing
→ **Clinical Settings:** Bedside sequencing for rapid diagnosis in ICUs and emergency situations
→ **Space Research:** First DNA sequencer used on the International Space Station (2016)
→ **Educational Use:** Makes sequencing accessible to teaching labs and small research groups
→ **Point-of-Care:** Enables sequencing in resource-limited settings without specialized infrastructure

Oxford Nanopore Product Line

Flongle: Even smaller, single-use adapter with 126 channels for quick tests (~$90)
MinION: USB portable device with 512-2048 channels (varies by flow cell)
GridION: Benchtop device that runs 5 MinION flow cells simultaneously
PromethION: High-throughput platform with up to 48 flow cells, generating terabases of data

5. DIRECT RNA SEQUENCING & NATIVE BASE MODIFICATION DETECTION



Direct RNA Sequencing

Nanopore technology is unique in its ability to sequence RNA molecules directly without conversion to cDNA. This preserves critical information about RNA modifications and structure that is lost in traditional RNA-seq methods.

What Can Be Detected:

Feature	Information Preserved	Applications
Full-length Transcripts	Complete mRNA from 5' cap to poly(A) tail	Isoform identification, alternative splicing analysis
RNA Modifications	m^6A, m^5C, pseudouridine (ψ), inosine	Epitranscriptomics, gene regulation studies
Poly(A) Tail Length	Exact length of poly(A) tail (not just presence)	mRNA stability, translation regulation
Base Modifications	Direct detection without chemical treatment	RNA editing, post-transcriptional regulation
True Quantification	No PCR bias, direct molecule counting	Accurate gene expression levels

DNA Base Modification Detection

Beyond RNA, nanopore sequencing can detect modified bases in DNA, including methylation patterns, without bisulfite conversion or other chemical treatments that can damage DNA.

Detectable DNA Modifications:

- 5-methylcytosine (5mC) - most common DNA methylation
- 5-hydroxymethylcytosine (5hmC)
- N6-methyladenine (6mA)
- 4-methylcytosine (4mC) - bacterial restriction modification

Revolutionary Applications

→ **Cancer Epigenetics:** Detect methylation patterns across entire genomes simultaneously with sequence

→ **Epitranscriptomics:** Study of RNA modifications and their roles in gene regulation

→ **Bacterial Typing:** Identify bacteria by their unique methylation patterns

→ **Full Isoform Characterization:** Understand complete transcript structures including novel splice variants

→ **No Bias:** Eliminates PCR and RT biases that affect quantitative accuracy

→ **Developmental Biology:** Track changes in modifications during cell differentiation

The Future of Modification Detection

As machine learning algorithms improve, nanopore sequencing is becoming increasingly capable of detecting more types of modifications with higher accuracy. This opens new frontiers in understanding how chemical modifications regulate gene expression and cellular function beyond the genetic code itself.

CONCLUSION: THE FUTURE OF SEQUENCING

Transforming Genomics

Nanopore sequencing represents a paradigm shift in how we read genetic information. By eliminating the need for synthesis, amplification, and specialized laboratory infrastructure, it has democratized access to sequencing technology and enabled applications that were previously impossible.

The combination of ultra-long reads, portability, real-time analysis, and native modification detection makes nanopore sequencing uniquely suited for addressing some of the most challenging problems in genomics, from completing reference genomes to rapid pathogen surveillance during disease outbreaks.

As the technology continues to improve in accuracy and throughput, nanopore sequencing is positioned to become a primary tool for both research and clinical applications, bringing us closer to the vision of ubiquitous, real-time genomic information.

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%

Detailed Component Breakdown

Line 1: Sequence Identifier

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

Components:

- **@** - Indicates start of FASTQ record
- **HWUSI-EAS100R** - Instrument name
- **6** - Flow cell lane
- **73** - Tile number within lane
- **941** - X-coordinate on tile
- **1973** - Y-coordinate on tile
- **#0** - Index sequence (for multiplexing)
- **/1** - Read number (paired-end: /1 or /2)

Line 2: Nucleotide Sequence

```
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
```

Details:

- Raw base calls from sequencing instrument
- Standard nucleotides: **A** (Adenine), **T** (Thymine), **C** (Cytosine), **G** (Guanine)
- **N** represents ambiguous base call
- Length varies by sequencing platform (typically 50-300 bp)
- Read direction: 5' → 3'

Line 3: Separator Line

```
+
```

Purpose:

- Always begins with **+** symbol
- Separates sequence from quality scores
- May optionally repeat the identifier from Line 1
- Modern FASTQ files typically use just "+" for efficiency

Line 4: Quality Scores

```
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Encoding System:

- ASCII characters represent Phred quality scores
- Each character corresponds to one base in Line 2
- **Must be same length as sequence**

Quality Score Examples:

- **!** (ASCII 33) = Q0 = 0% accuracy
- ***** (ASCII 42) = Q9 = 87.4% accuracy
- **5** (ASCII 53) = Q20 = 99% accuracy
- **?** (ASCII 63) = Q30 = 99.9% accuracy
- **I** (ASCII 73) = Q40 = 99.99% accuracy

Calculation: Quality (Q) = ASCII value - 33

Phred Quality Score Visual Guide

ASCII Character to Quality Score Mapping

Low Quality (Q0-Q20)
Characters: ! " # $ % & ' ( ) * + , - . / 0 1 2 3 4

Medium Quality (Q20-Q30)
Characters: 5 6 7 8 9 : ; < = >

High Quality (Q30+)
Characters: ? @ A B C D E F G H I J K

Phred Quality Score Formula

$Q = -10 \times \log_{10}(P)$ | $P = 10^{(-Q/10)}$ | Where P = probability of incorrect base call

Complete FASTQ Record Example

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Interpretation:

- Sequence length: 36 bases
- Quality scores: Mostly "I" characters (Q40 = 99.99% accuracy)

• Last few bases show slightly lower quality: "9" (Q24 = 99.4%), "I" (Q40), "G" (Q38), "9" (Q24), "I" (Q40), "C" (Q34)

• Overall: High-quality read suitable for downstream analysis

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

Quality Score Principles

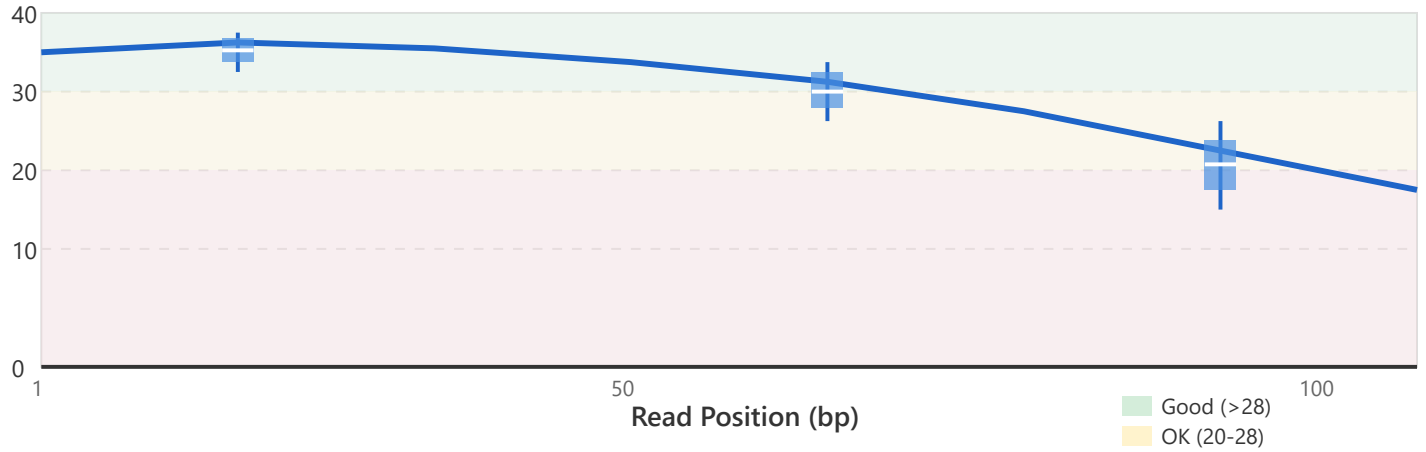
Phred Quality Score

$Q = -10 \times \log_{10}(P)$

where P is the probability of base calling error.

Phred 20	Phred 30	Phred 40
99% Accuracy	**99.9% Accuracy**	**99.99% Accuracy**
1/100 error rate	1/1,000 error rate	1/10,000 error rate

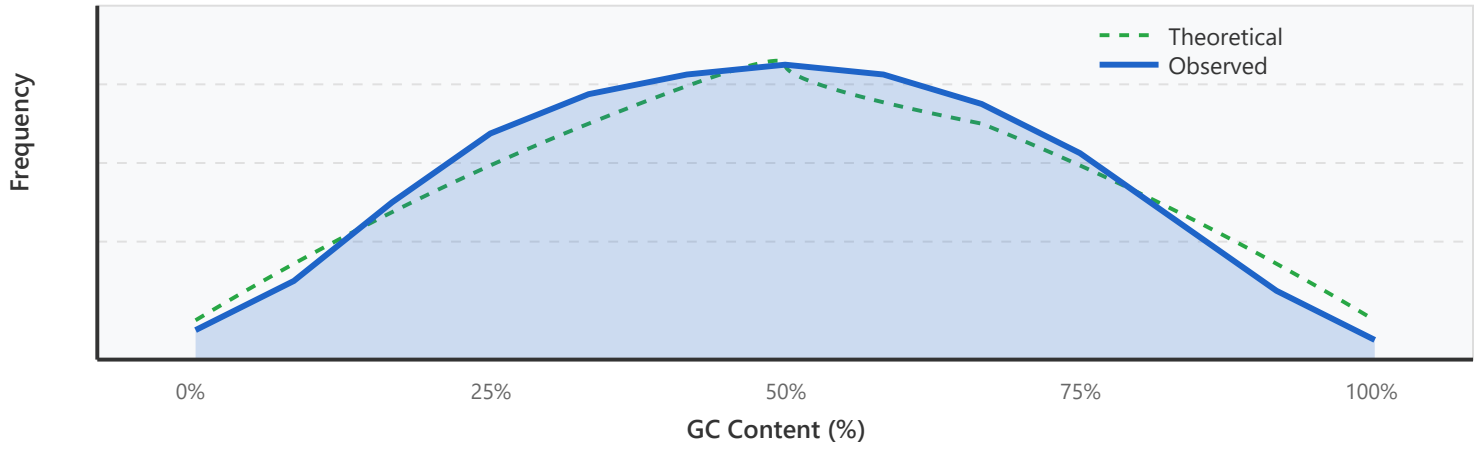
Quality Score Visualization Example



Typical pattern showing quality score decrease toward the end of reads

GC Content Analysis

GC Content Distribution



Normal samples show a normal distribution centered around the species' average GC content

Normal GC Distribution

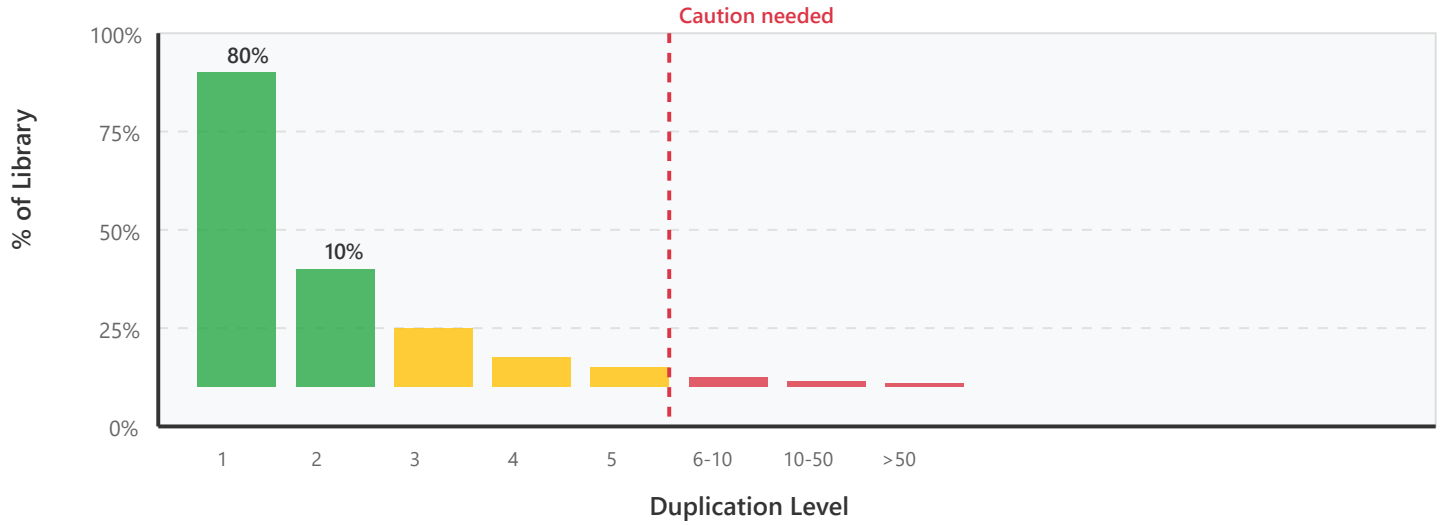
- Single peak
- Near expected GC%
- Narrow variance
- Matches theoretical distribution

Abnormal GC Distribution

- Multiple peaks (contamination)
- GC% bias
- Wide variance
- Deviates from theory

Sequence Duplication Analysis

Duplication Level Distribution



Good quality: most sequences are unique (single occurrence) / Poor quality: high duplication

- **PCR duplicates:**Caused by excessive PCR amplification during library preparation
- **Optical duplicates:**Caused by overly dense sequencing clusters
- **Biological duplicates:**Actually identical DNA fragments (highly expressed genes in RNA-seq)
- **Threshold:**Generally, >50% duplication indicates a problem

Adapter Contamination

Adapter Content Example

Read sequence:
ATCGATCGATCGATCGAGATCGGAAGAGC

Insert (original DNA):
ATCGATCGATCGATCG
+ Adapter sequence (needs removal)

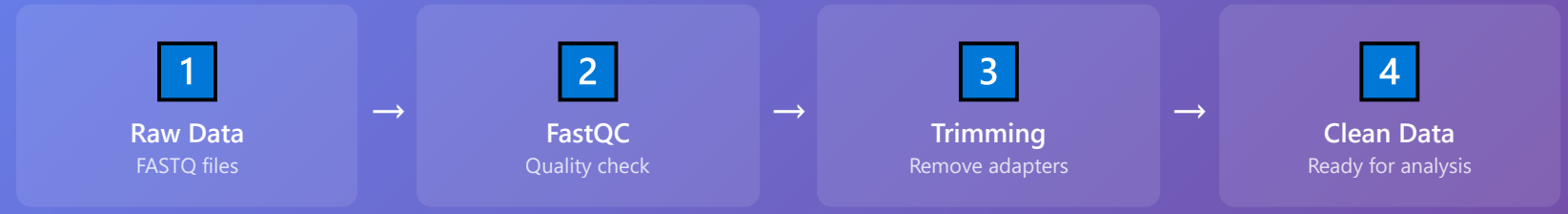
Why Adapters Remain

- When insert DNA is shorter than read length
- Sequencing reads beyond insert into adapter
- Common in small RNA-seq

Removal Methods

- Use Cutadapt, Trimmomatic
- Specify adapter sequences
- Set minimum length
- Perform quality trimming simultaneously

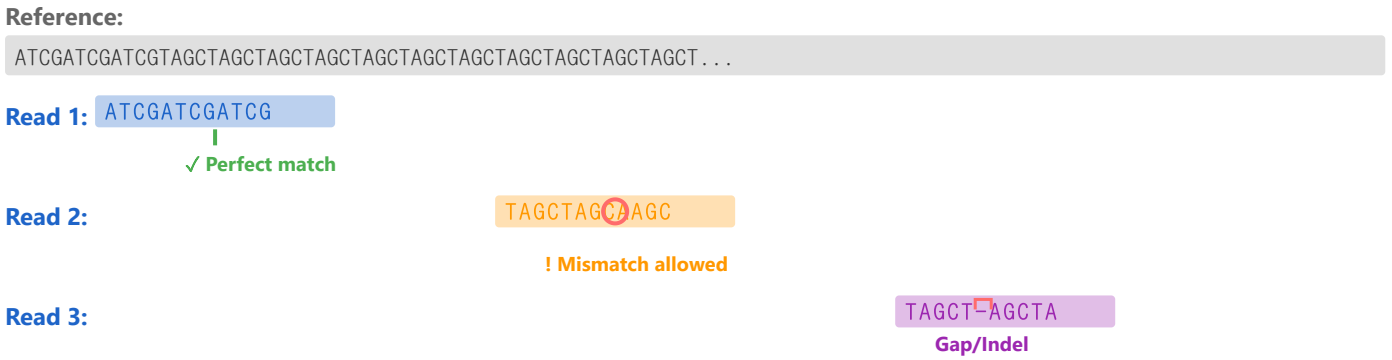
Quality Control Workflow



Essential first step for all NGS analysis - Quality control determines downstream analysis reliability

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)

📊 Detailed Examples and Explanations by Field

📌 1. QNAME (Query Name) - Read Name

QNAME: HWUSI-EAS100R:6:73:941:1973
QNAME: SRR123456.1
QNAME: READ_00001

Description: Unique identifier for each read generated by the sequencing instrument.

- Typically contains sequencer information and coordinate data
- Paired reads from the same DNA fragment share the same QNAME
- Format: InstrumentID:RunNumber:FlowCell:Lane:Tile:X-coord:Y-coord

📌 2. FLAG - Bitwise Flag

FLAG Value Calculation Example

```
FLAG = 99 (0x63)
= 1 (paired) + 2 (properly paired) + 32 (mate reverse strand) + 64 (first in pair)
```

- `0x1 (1) = paired`
- `0x2 (2) = properly paired`
- `0x4 (4) = unmapped`
- `0x8 (8) = mate unmapped`
- `0x10 (16) = reverse strand`
- `0x20 (32) = mate reverse`
- `0x40 (64) = first in pair`
- `0x80 (128) = second in pair`

Description: Integer value representing various read attributes in binary format.

• Efficiently stores multiple attributes in a single number

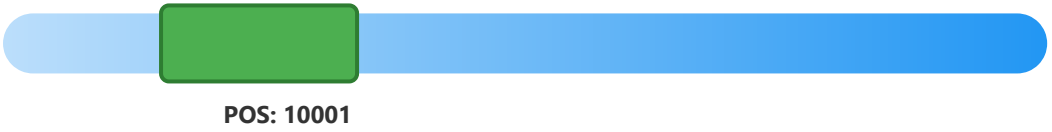
• Each bit has independent meaning (verified using AND operation)

• Example: FLAG=99 → Paired read, properly mapped, first read, mate is reverse strand

📌 3. RNAME & POS - Reference Sequence Name and Position

```
RNAME: chr1, chr2, chrX, chrY, chrM
POS: 10001 (1-based coordinate)
```

Chromosome Position Visualization



← chr1 start (1) ····· Read mapping position ····· chr1 end (248,956,422) →

Description:

• **RNAME:** Chromosome or contig name in reference genome (e.g., chr1, chr2, chrX)

• **POS:** 1-based coordinate where the read is mapped (chromosome starts at 1)

• '*' or '0' indicates an unmapped read

• BAM file indexing (.bai) enables rapid retrieval of specific regions

📌 4. MAPQ - Mapping Quality Score

MAPQ Value	Meaning	Error Probability	Confidence
60	Very high quality	0.0001% (1/1,000,000)	✓✓✓✓✓
40	High quality	0.01% (1/10,000)	✓✓✓✓
20	Medium quality	1% (1/100)	✓✓✓
10	Low quality	10% (1/10)	✓✓
0	No mapping confidence	-	✗

Description: Phred-scaled mapping quality score ($-10 \times \log_{10}(P)$)

- Represents the probability that the read is mapped to an incorrect position
- MAPQ ≥ 30: Generally considered reliable mapping
- MAPQ 0: Multi-mapping (maps equally well to multiple locations)
- Frequently used as filtering criteria in variant calling

📌 5. CIGAR - Alignment String

CIGAR: 50M2I48M1D25M

50M 2I 48M 1D 25M

Visual Representation:

💡 SAM/BAM formats are the standard formats for NGS data analysis, serving as the foundation for all sequencing analyses including variant analysis, RNA-seq, and ChIP-seq.

```
Ref:  ACGTACGT-ACGTACGTACGTACGT
Read: ACGTACGTTTACGTACGT-CGTACGT
CIGAR: 8M2I8M1D7M
```

Operator	Meaning	Description
M	Match/Mismatch	Aligned to reference (includes matches/mismatches)
I	Insertion	Base inserted in read (not in reference)
D	Deletion	Base deleted in read (present in reference)
S	Soft Clipping	Unaligned read ends (sequence retained)
H	Hard Clipping	Unaligned read ends (sequence removed)
N	Skipped Region	Skipped reference region (RNA-seq introns)

Description: Compact representation of alignment between read and reference sequence

- Each operation: number (length) + character (operation type)
- Critical information for structural variant and splice junction detection
- In RNA-seq, 'N' operation represents splicing junctions

📌 6. Complete SAM Line Example

```
READ1 99 chr1 10001 60 76M = 10052 127
ACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGT
```

```
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
NM:i:0 MD:Z:76 AS:i:76
```

Field	Value	Description
QNAME	READ1	Read identifier
FLAG	99	Paired, properly mapped, first read, mate reverse strand
RNAME	chr1	Chromosome 1
POS	10001	Start position
MAPQ	60	Very high mapping quality
CIGAR	76M	76 bases perfectly aligned
RNEXT	=	Mate also mapped to same chromosome
PNEXT	10052	Mate start position
TLEN	127	Template length (insert size)
SEQ	ACGT...	Read sequence
QUAL	IIII...	Per-base quality scores (Phred+33)

📌 7. BAM File Operations Example

```
# Convert SAM to BAM
samtools view -bS input.sam > output.bam

# Sort BAM
```

```
samtools sort output.bam -o sorted.bam

# Index BAM (creates .bai file)
samtools index sorted.bam

# Extract specific region
samtools view sorted.bam chr1:10000-20000

# View statistics
samtools flagstat sorted.bam
```

Description: BAM files are essential for efficient data processing.

- Only sorted BAM files can be indexed
- .bai index file enables rapid retrieval of specific regions
- Visualization tools like IGV require BAM + index
- Typical compression ratio is 1/3 to 1/5 compared to SAM

etailed Examples

otype Determination

t Calling

G)] / P(D)

P(D | G): Likelihood of data given genotype

10177 (A→G)

Genotype: G/G

Observed: 1A, 5G reads

P(D|G/G) = 0.413

P(G/G|D) ≈ 0.81 (81%) ✓

nd selects the one with the highest
read (A), the model correctly identifies G/G
equencies. The low probability for A/A

3. Read Depth & Allele Frequency Analysis

Interpreting Coverage and Allele Balance

Example 1: Heterozygous SNV (A/G)

Aligned Reads (DP=20):

Allele A: (9 reads, 45%)

Allele G: (11 reads, 55%)

Total Depth (DP): 20
Allele Frequency (AF): 0.55 (55%)
Expected for Het: ~0.5 (50%)
✓ **Consistent with A/G genotype**

Example 2: Homozygous Variant (G/G)

Aligned Reads (DP=25):

Allele A: (1 read, 4%)

Allele G: (24 reads, 96%)

Total Depth (DP): 25
Allele Frequency (AF): 0.96 (96%)
Expected for Hom: ~1.0 (100%)
✓ **Consistent with G/G genotype**

Example 3: Low Coverage (Unreliable)

Aligned Reads (DP=3):

Allele A: (1 read, 33%)

Allele G: (2 reads, 67%)

Total Depth (DP): 3
Allele Frequency (AF): 0.67 (67%)
Minimum recommended: DP ≥ 10
✗ **Insufficient coverage - unreliable call**

Example 4: Strand Bias (Artifact)

Aligned Reads (DP=20):

Forward (+): 10 A reads 0 G reads

Reverse (-): 0 A reads 10 G reads

Fisher Strand Bias (FS): 85.2
Threshold: FS < 60 for SNVs
Variant only on one strand
✗ **Likely sequencing artifact**

Coverage & Allele Balance Interpretation

Read Depth (DP): Higher coverage provides more statistical power. Minimum 10x recommended, 30x or higher ideal for clinical applications. **Allele Frequency (AF):** For diploid organisms, heterozygous variants should have AF ≈ 0.5, homozygous variants AF ≈ 1.0. Deviations may indicate sequencing bias, copy number variations, or sample contamination. **Strand Bias:** True variants should appear on both DNA strands equally; systematic bias suggests technical artifacts.

4. Haplotype-Based Variant Callin

Local A

Step 1:

Reference: ...ATCGATCG AAAA TCGATCGATCG

Active Region (H

S

Reads in active region:

Read1: ...ATCG AAA TCGATC G GATCG...

Read2: ...ATCG AAAA TCGATC C GATCG...

...and 13 more reads supporting these patterns...

Assembled Haplotypes:

H1: ...ATCG-AAA-TCGATC-G-GATCG... (1bp deletion +

Step 3:

H2/H2 (Ref/Ref)

Both chromosomes match reference

Likelihood: 0.001

P(H2/H2|reads) = 2%

Advantages of Haplotype-Based Calling

Traditional variant callers evaluate each posi
variants. Haplotype-based methods like GAT
haplotypes in regions of variation. This appr
phases nearby variants that occur on the sa

n Variant Calling

rmula

rror)

l or variant call is incorrect

Table

Accuracy	Interpretation
90%	Low quality
99%	Moderate
99.9%	Good (Standard)
99.99%	High quality
99.9999%	Very high quality

10-fold decrease in error probability. A
neaning we accept only variants with
are combined with mapping quality and

Better handles complex variants like MNPs (
than multiple independent SNVs.

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

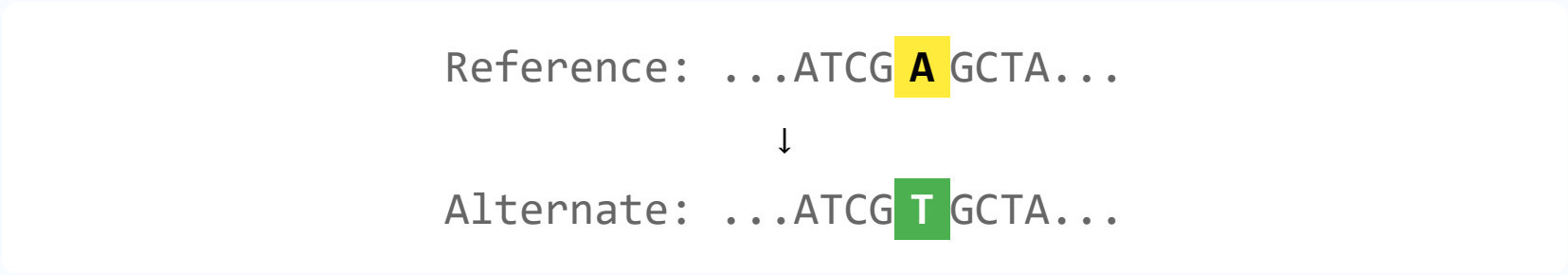
AF = Allele frequency

AC = Allele count

AN = Total alleles

Variant Types with Examples

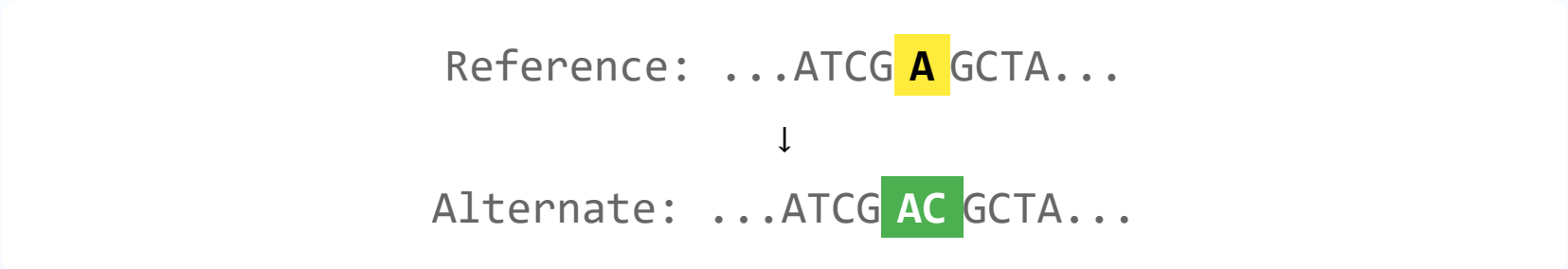
1. SNP (Single Nucleotide Polymorphism)




```
chr1 10352 rs123 A T 100 PASS
```

The most common type of variant, where a single nucleotide is substituted with another. In this example, A at position 10352 is changed to T.

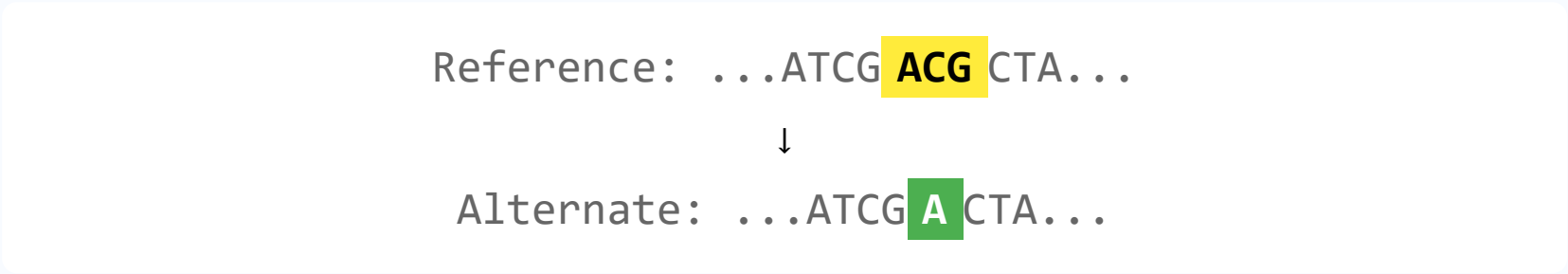
2. Insertion




```
chr1 10177 . A AC 50 PASS
```

Nucleotides are added to the reference sequence. REF is the base at the insertion position (A), and ALT includes REF + the inserted base (AC).

3. Deletion




```
chr1 20000 . ACG A 80 PASS
```

The sequence is removed. REF includes the deleted bases (ACG), and ALT shows only the remaining base (A).

Genotype Interpretation

0/0

Homozygous Reference

Allele 1: A
Allele 2: A

Both alleles match the reference sequence

0/1

Heterozygous

Allele 1: A
Allele 2: T

One reference allele, one alternate allele

1/1

Homozygous Alternate

Allele 1: T
Allele 2: T

Both alleles are alternate variants

Quality Metrics Explained

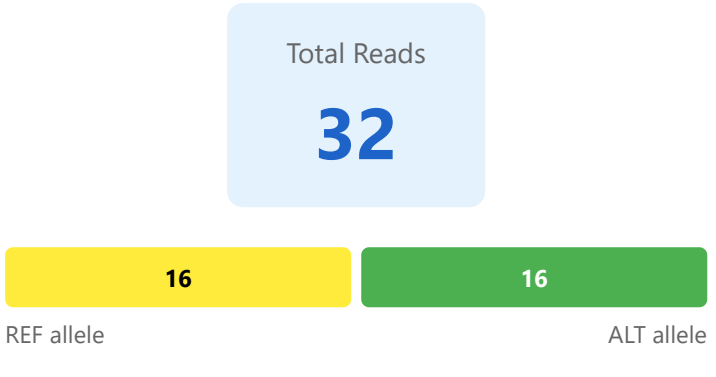
QUAL (Quality Score)

Phred-scaled quality score

```
QUAL = 10 → 90% accuracy
QUAL = 20 → 99% accuracy
QUAL = 30 → 99.9% accuracy
QUAL = 40 → 99.99% accuracy
```

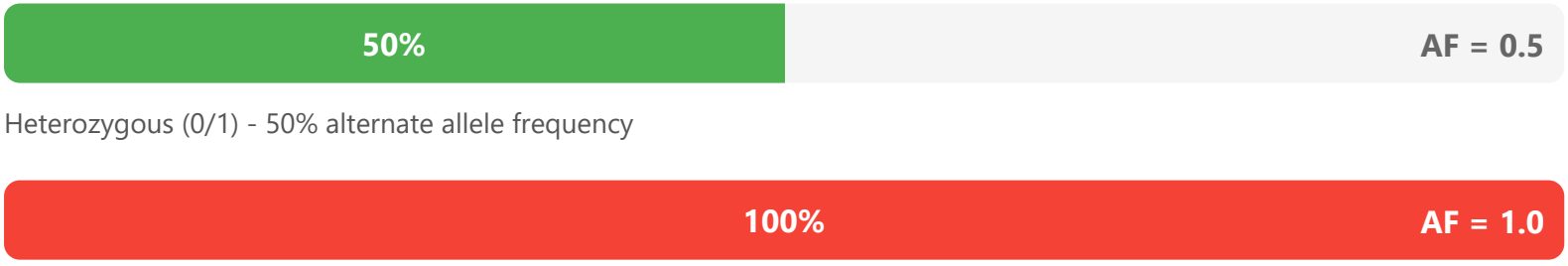
Higher values indicate greater confidence in the variant call. Generally, values of 30 or above are considered reliable variants.

DP (Depth)



The total number of reads covering this position. Higher depth increases confidence in variant calling.

AF (Allele Frequency)



Heterozygous (0/1) - 50% alternate allele frequency

Homozygous alternate (1/1) - 100% alternate allele frequency

Complete Example Analysis

```
chr1 10352 rs123 T A 100 PASS DP=45;AF=1.0 GT:DP:GQ 1/1:45:99
```

Location Information

- Chromosome: chr1
- Position: 10,352
- dbSNP ID: rs123

Variant Information

- Reference base: T
- Alternate base: A
- Variant type: SNP

✓ Quality Information

- Quality Score: 100 (very high)
- Filter: PASS
- Depth: 45 reads

Sample Information

- Genotype: 1/1 (homozygous)
- Allele Frequency: 100%
- Genotype Quality: 99

Interpretation

This sample shows a homozygous variant at position 10,352 on chromosome 1, where both alleles have changed from T to A. Confirmed by 45 reads, with high quality score (100) and genotype quality (99), indicating a highly reliable variant call.

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)

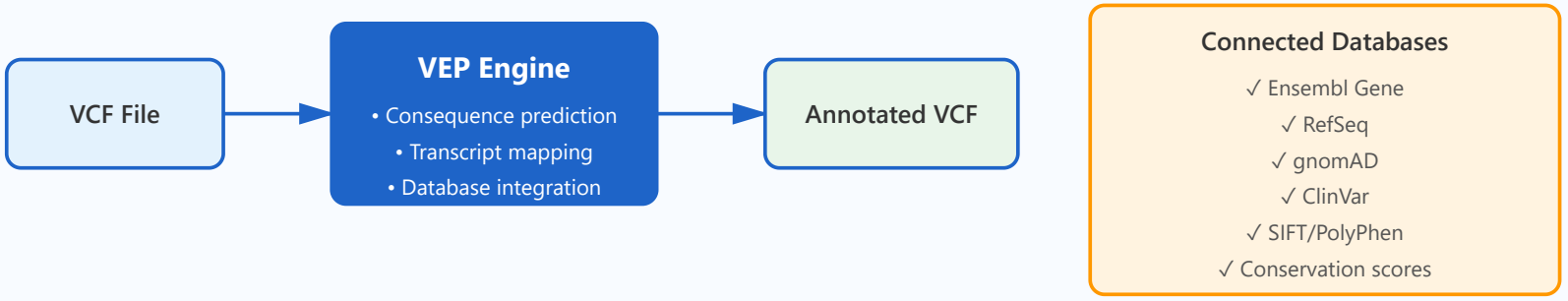
Prediction Tools

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

Conservation

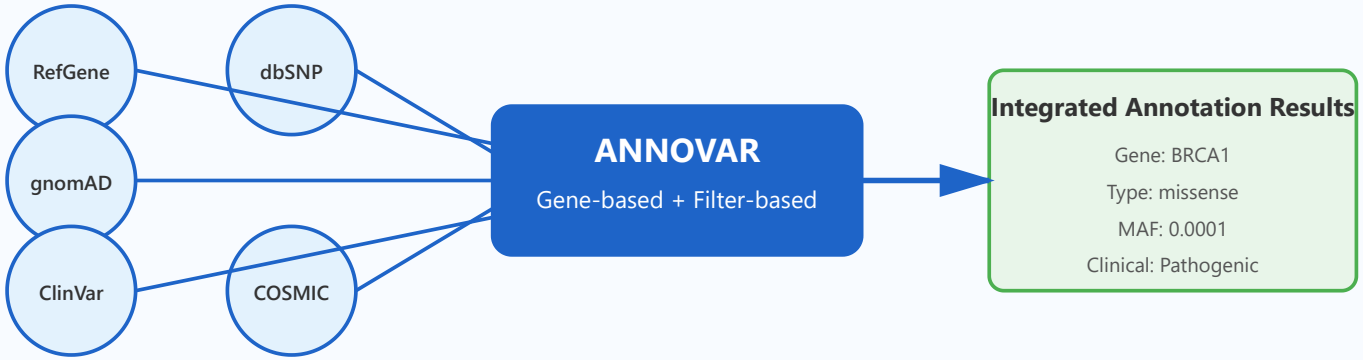
- PhyloP
- GERP++
- PhastCons

VEP (Variant Effect Predictor) Details



Key Features: VEP is a tool developed by Ensembl that predicts the functional effects of genetic variants. It automatically classifies consequence types such as missense, nonsense, and frameshift for each variant, and simultaneously analyzes effects across multiple transcripts.

ANNOVAR Workflow



Features: ANNOVAR has a modular architecture that allows sequential application of various databases. Gene-based annotation identifies gene functions, while filter-based annotation adds frequency and clinical information. Command-line based, it is efficient for large-scale data processing.

SnpEff Annotation Process

Advantages: SnpEff is Java-based and boasts very fast processing speeds, automatically generating variant effect statistics in HTML format along with annotation results. It is particularly efficient for large-scale WGS data or population studies, storing all information in a structured format in the ANN field.

Population Database Usage Example

gnomAD (Genome Aggregation Database)

```
Variant: chr17-43044295-G-A (BRCA1)
Global AF: 0.00015 (15/100,000)
East Asian AF: 0.00008 (8/100,000)
European AF: 0.00022 (22/100,000)
```

> 💡 **Interpretation:** Very low frequency → Consider pathogenicity possibility as rare variant

Clinical Database Example

ClinVar - Clinical Significance

Pathogenic ★★★★

Review Status: **4 stars**
Condition: Breast cancer
Submitters: 12 labs

OMIM - Disease Association

Gene: BRCA1

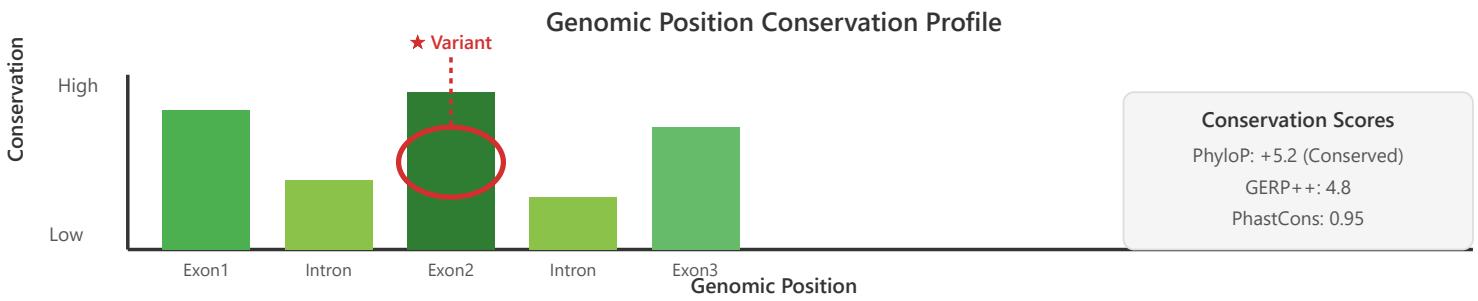
MIM: 113705
Associated:
- Breast-ovarian cancer
- Fanconi anemia

Pathogenicity Prediction Comparison

Tool	Score/Result	Interpretation	Method
SIFT	0.01 (Deleterious)	Score < 0.05	Sequence homology
PolyPhen-2	0.98 (Damaging)	Score > 0.85	Structure + conservation
CADD	28.5 (Pathogenic)	Phred > 20	Machine learning

✓ **Overall Assessment:** All three tools predict pathogenicity with concordant results. Concordance across multiple prediction tools increases confidence in the functional impact of the variant.

Conservation Score Visualization



💡 **Interpretation Guide:** High conservation scores indicate that the position is evolutionarily conserved, and variants at these positions are likely to have significant effects on protein function. In particular, high

conservation in exon regions suggests functional importance.

Practical Annotation Workflow Example

📋 Recommended Filtering Criteria

✓ Population AF < 0.01 (1%)

✓ CADD score > 15

✓ Loss-of-function variants (frameshift, nonsense, splice site)

✓ ClinVar: Pathogenic/Likely pathogenic

✓ Gene: Disease-related gene panel

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

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

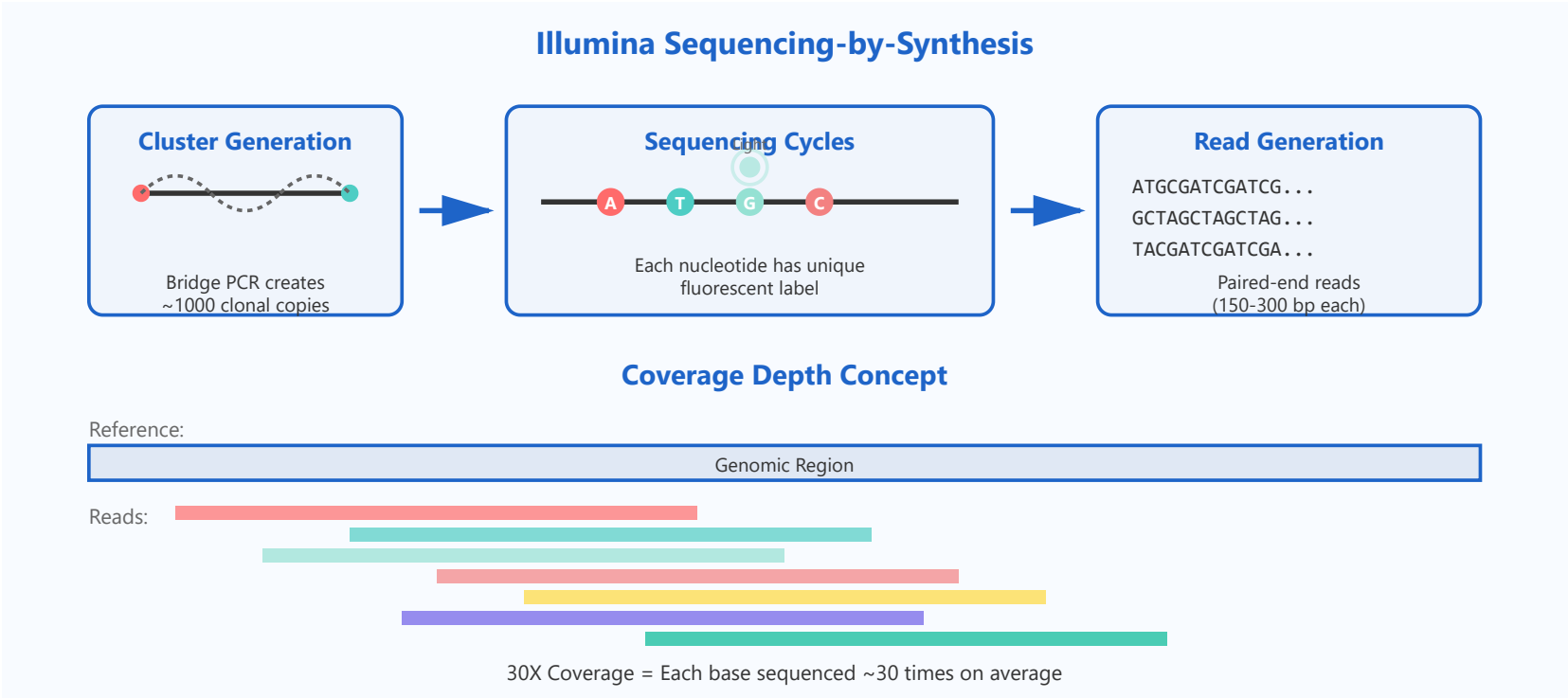
↓

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

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

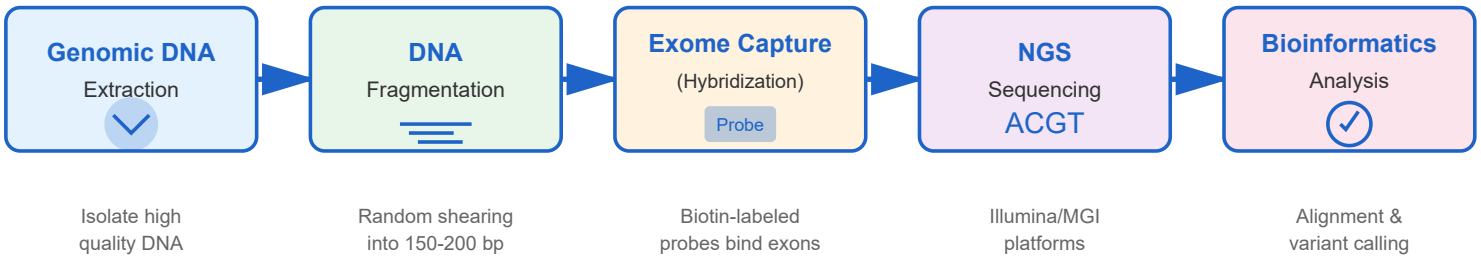
WES Limitations

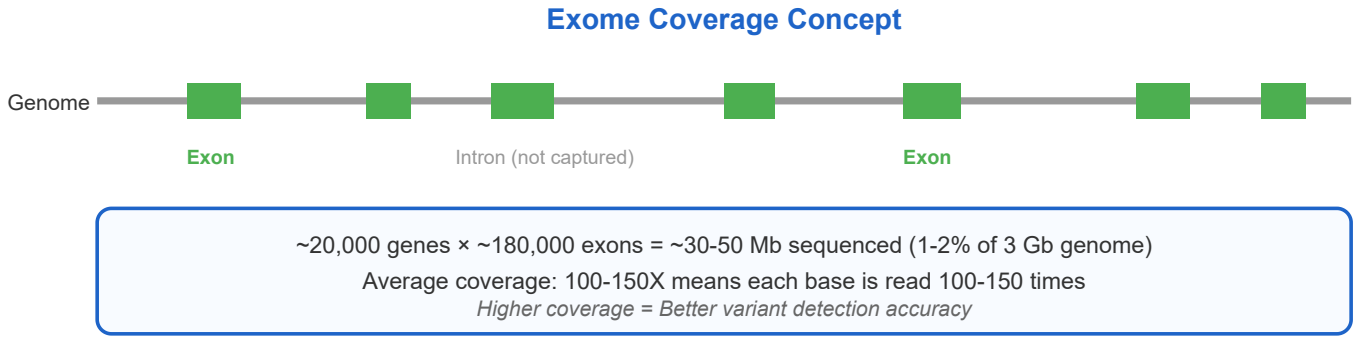
- Misses regulatory variants
- Limited structural variant detection
- Capture bias
- Non-coding regions excluded

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

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	Cardio	Neuro
50-500 genes	**50-200 genes**	**100-300 genes**
Oncology hotspots	Heart conditions	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

Principle: Selective Enrichment

- Target enrichment focuses sequencing on specific genomic regions of interest
- Reduces sequencing cost by 10-1000x compared to whole genome sequencing
- Increases depth of coverage for better variant detection
- Enables detection of low-frequency somatic variants (as low as 1-5%)

Capture Methods

1. Hybridization Capture (Solution-based)

Custom oligonucleotide probes (baits) complementary to target regions are mixed with fragmented DNA library. Target fragments hybridize to biotinylated probes and are captured using streptavidin-coated magnetic beads. Non-target DNA is washed away.

Examples: Agilent SureSelect, IDT xGen, Twist Bioscience
Best for: Larger panels (>100 genes), exome sequencing

2. Amplicon-based Sequencing (PCR)

Multiple primer pairs designed to amplify specific target regions simultaneously in a single multiplex PCR reaction. Amplified products are pooled and sequenced directly.

Examples: Illumina AmpliSeq, Ion Torrent AmpliSeq
Best for: Small-medium panels (10-200 genes), hotspot regions

3. Molecular Inversion Probes (MIPs)

Single-stranded DNA probes with sequences complementary to regions flanking the target. After hybridization, the probe circularizes around the target sequence, which is then amplified.

Best for: SNP genotyping, copy number variation detection

Targeted Panel Sequencing Workflow

1

DNA Extraction & QC

Extract high-quality genomic DNA from sample (blood, tissue, saliva). Assess quantity (10-500 ng typically required) and quality (DIN/RIN score).

2

Library Preparation

Fragment DNA to optimal size (150-300 bp). Attach adapters and unique molecular identifiers (UMIs) to enable sequencing and reduce PCR duplicates.

3

Target Enrichment

Apply hybridization capture or amplicon-based enrichment to isolate genomic regions of interest. Wash away non-target DNA sequences.

4

PCR Amplification

Amplify enriched library to generate sufficient material for sequencing. Typical: 8-12 PCR cycles to minimize amplification bias.

5

Next-Generation Sequencing

Sequence enriched library on NGS platform (Illumina, Ion Torrent, MGI). Generate paired-end reads (typically 2×150 bp) with high depth (>500X average coverage).

6

Bioinformatics Analysis

Align reads to reference genome, call variants (SNVs, indels, CNVs), filter artifacts, annotate variants, and interpret clinical significance using databases (ClinVar, COSMIC, gnomAD).

Method Comparison

Feature	Hybridization Capture	Amplicon-based
Input DNA	50-500 ng	10-50 ng
Uniformity	Excellent across targets	Variable (primer efficiency)
Target Size	Best for large panels (>1 Mb)	Best for small panels (<500 kb)
Workflow Time	2-3 days	1 day

Feature	Hybridization Capture	Amplicon-based
Sensitivity	5-10% allele frequency	1-5% allele frequency
Cost per Sample	$150-400	$100-250
Best Application	Hereditary disease panels, exomes	Cancer hotspots, pharmacogenetics

Key Performance Metrics

Coverage Metrics

- **Mean coverage depth:** >500X
- **Target coverage:** >95% at 100X
- **Uniformity:** >80% bases within 0.2x mean
- **On-target rate:** >50% reads

Variant Detection

- **SNV sensitivity:** >99%
- **Indel sensitivity:** >95%
- **CNV detection:** Exon-level resolution
- **Somatic VAF:** As low as 1-5%

High depth sequencing enables confident detection of both germline and somatic variants with clinical-grade accuracy

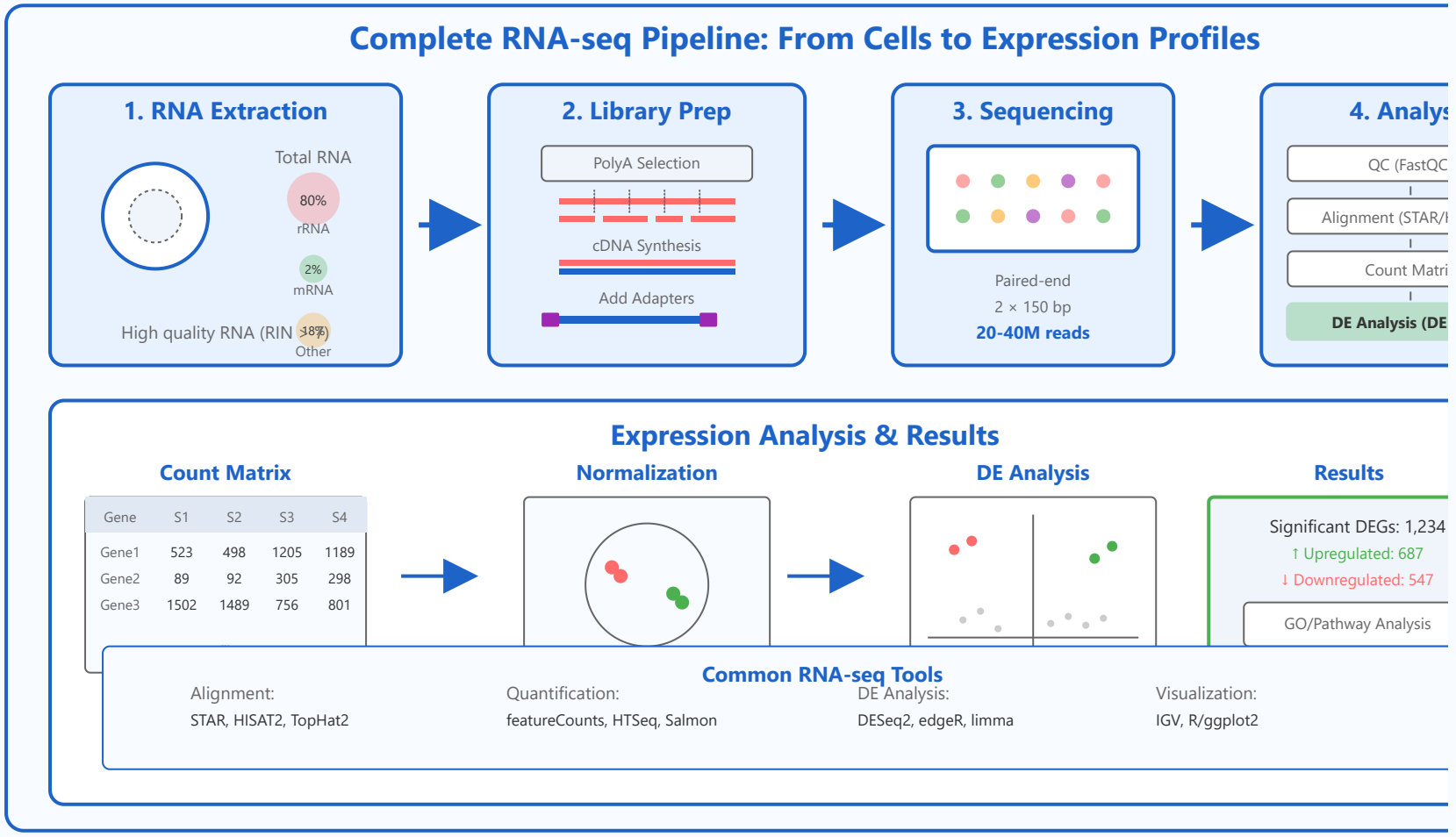
Clinical Applications

- **Oncology:**Somatic mutation profiling for targeted therapy selection (e.g., EGFR, KRAS, BRAF in solid tumors)
- **Hereditary Cancer:**BRCA1/2, Lynch syndrome genes (MLH1, MSH2, MSH6, PMS2), Li-Fraumeni (TP53)
- **Cardiovascular:**Cardiomyopathy genes (MYH7, MYBPC3, TTN), arrhythmia panels (SCN5A, KCNQ1)
- **Neurology:**Epilepsy genes (SCN1A, KCNQ2), intellectual disability panels, muscular dystrophy genes

- **Pharmacogenomics:** Drug metabolism genes (CYP2D6, CYP2C19, TPMT, SLCO1B1) for personalized medication

RNA-seq Overview

RNA-seq Workflow & Mechanism



Applications

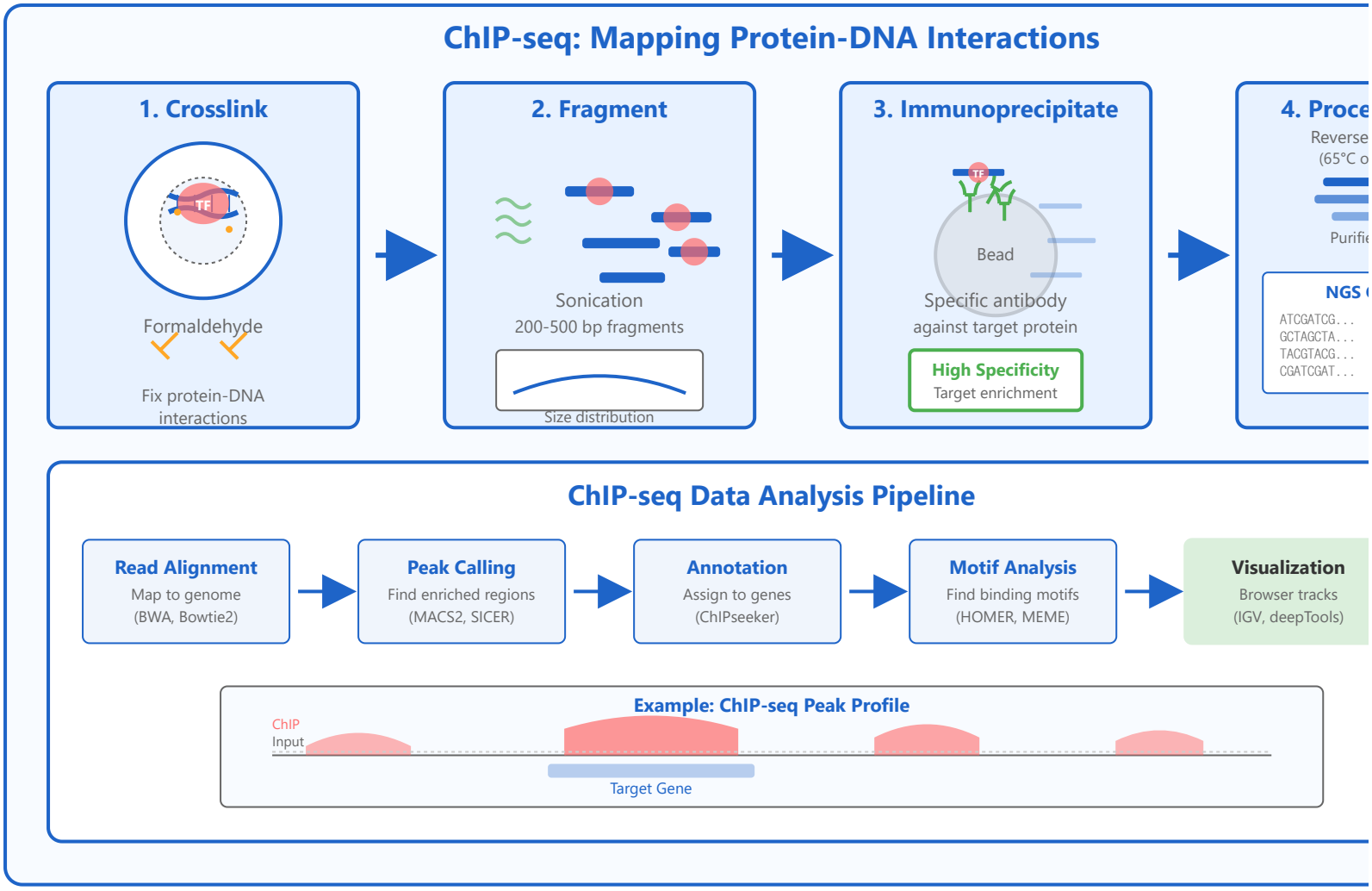
- Differential gene expression
- Alternative splicing analysis
- Novel transcript discovery
- Allele-specific expression

Key Considerations

- Biological replicates (≥3)
- Read depth (20-40M reads)
- Strand-specific protocols
- Batch effect correction

ChIP-seq (Chromatin Immunoprecipitation Sequencing)

ChIP-seq Mechanism and Workflow



Common Targets

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

Critical Controls

- Input DNA (no IP)
- IgG control (non-specific)
- Biological replicates
- Technical validation

Requires high-quality antibodies and proper controls for reliable results

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

Detailed Workflow: Step-by-Step Protocol

1 Cell Preparation & Lysis

Fresh or frozen cells are gently lysed using cold lysis buffer to isolate intact nuclei. The nuclear membrane is permeabilized while maintaining chromatin integrity. Critical for preserving native chromatin structure.

2 Transposition Reaction

Hyperactive Tn5 transposase loaded with sequencing adapters simultaneously fragments and tags accessible DNA regions. The reaction occurs at 37°C for 30 minutes. Tn5 preferentially inserts into open chromatin while nucleosome-bound DNA remains protected.

3 DNA Purification

Tagmented DNA is purified using column-based or bead-based methods to remove proteins, enzymes, and debris. This yields adapter-tagged DNA fragments ready for amplification.

4 PCR Amplification

Limited-cycle PCR (typically 5-12 cycles) amplifies tagmented fragments and adds indexing barcodes. The number of cycles is optimized based on input cell number to minimize PCR bias.

5

Library Quality Control & Sequencing

Final libraries are assessed for size distribution (typically 200-600 bp) and concentration. High-throughput paired-end sequencing (50-100 bp reads) generates 50-100 million read pairs per sample.

Chromatin Accessibility: Open vs. Closed Regions

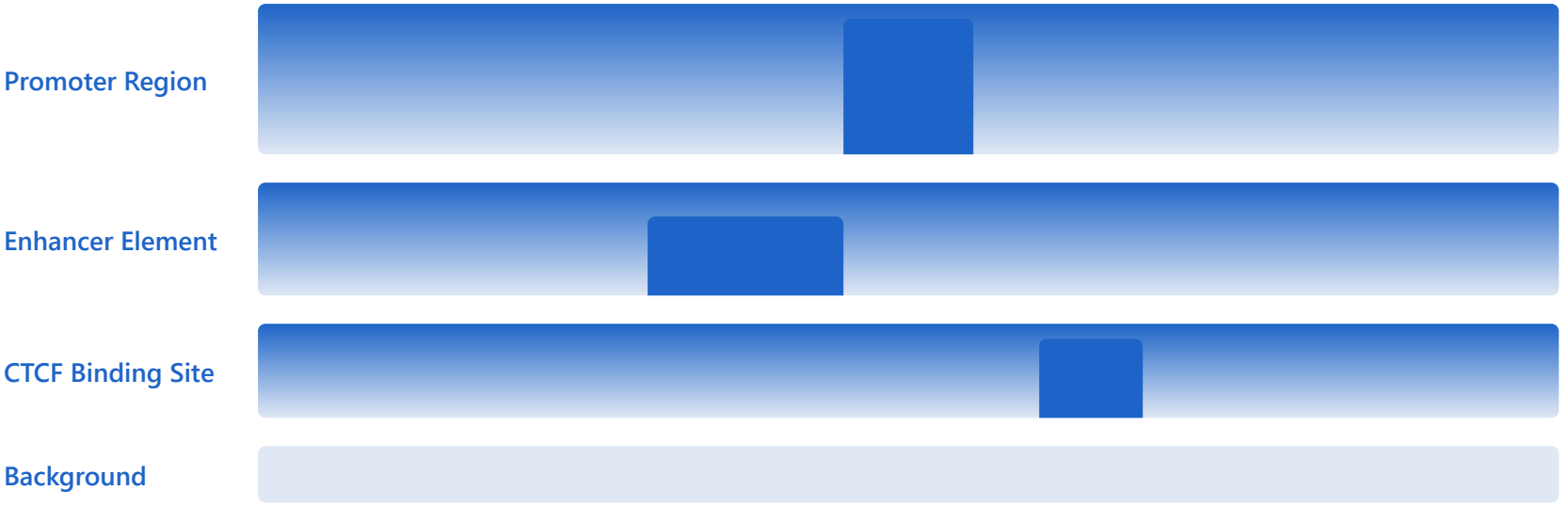
Tightly packed nucleosomes
Tn5 cannot access
Transcriptionally inactive

Nucleosome-depleted regions
Tn5 inserts here
Active regulatory elements

Key Concept: ATAC-seq exploits the differential accessibility of chromatin. The Tn5 transposase enzyme can only insert sequencing adapters into DNA regions that are not wrapped around histones. This creates a map of regulatory regions including promoters, enhancers, silencers, and insulators.

Data Analysis & Peak Interpretation

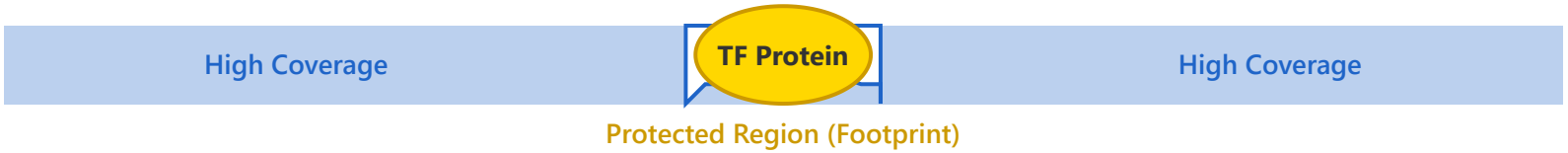
Typical ATAC-seq Signal Patterns



Peak Calling: Computational algorithms identify regions with significantly higher read coverage compared to background. Promoters typically show sharp, narrow peaks, while enhancers show broader peaks. Peak width and height correlate with regulatory element type and activity level.

Transcription Factor Footprinting

TF Binding Protection from Tn5



Footprinting Principle: When transcription factors bind to their cognate DNA sequences, they protect that region from Tn5 insertion. This creates a characteristic "dip" in coverage within an accessible region.

Applications: By analyzing footprint patterns and matching them to known TF binding motifs, researchers can infer which transcription factors are actively bound in a particular cell type or condition, enabling cell type identification and regulatory network reconstruction.

ATAC-seq vs. Other Chromatin Profiling Methods

Method	Target	Cell Input	Time Required	Key Advantage	Limitation
ATAC-seq	Open chromatin	500-50,000	~3 hours	Fast, low input, no antibodies	Cannot identify specific proteins
ChIP-seq	Specific proteins/modifications	1-10 million	2-3 days	Protein-specific information	Requires antibodies, high input
DNase-seq	Open chromatin	1-5 million	1-2 days	Gold standard for accessibility	Higher cell input, more complex

Method	Target	Cell Input	Time Required	Key Advantage	Limitation
FAIRE-seq	Nucleosome-depleted regions	~10 million	1 day	No specialized enzymes	Lower resolution, high input
MNase-seq	Nucleosome positioning	1-10 million	1-2 days	Precise nucleosome mapping	Doesn't directly measure accessibility

Why ATAC-seq is popular: The combination of low cell input requirements, rapid protocol, and high data quality has made ATAC-seq the method of choice for chromatin accessibility profiling, especially in rare cell populations and clinical samples where cell numbers are limited.

Applications & Research Examples

🧬 Development & Differentiation

Track chromatin remodeling during cell fate transitions and embryonic development. Identify lineage-specific regulatory elements.

- Hematopoiesis progression mapping
- Neural differentiation studies
- Stem cell characterization

🔬 Cancer Research

Discover cancer-specific regulatory alterations and identify driver mutations in non-coding regulatory regions.

- Tumor heterogeneity analysis
- Oncogenic enhancer identification
- Drug resistance mechanisms

🧪 Single-Cell Epigenomics

Profile chromatin accessibility in thousands of individual cells to reveal cellular heterogeneity and rare cell populations.

🧠 Disease Mechanisms

Link genetic variants to regulatory dysfunction in complex diseases through integration with GWAS data.

- Autoimmune disease studies

- Cell type identification
- Trajectory inference
- Regulatory variation mapping

- Neurological disorders
- Metabolic disease research

Quality Control Metrics

Library Complexity

TSS Enrichment: Signal enrichment at transcription start sites should be >7

FRiP Score: Fraction of reads in peaks should be >0.3 for good libraries

Fragment Size Distribution

Nucleosomal Pattern: Should see clear periodicity at ~200bp intervals

NFR Fragments: Peak at <100bp representing nucleosome-free regions

Mapping Statistics

Alignment Rate: >95% of reads should align to reference genome

Duplicate Rate: Should be <20% for sufficient library complexity

Peak Characteristics

Number of Peaks: Typically 50,000-150,000 peaks in mammalian cells

Peak Width: Median width usually 200-600bp

Computational Analysis Pipeline

1

Read Alignment

Paired-end reads are aligned to reference genome using Bowtie2 or BWA. Duplicates are removed, and only properly paired, uniquely mapped reads are retained.

2

Peak Calling

MACS2, Genrich, or HMMRATAC identify significant peaks representing accessible regions. FDR threshold typically set at 0.05 or 0.01.

3

Peak Annotation

Peaks are annotated with genomic features (promoter, enhancer, intergenic, etc.) using tools like ChIPseeker or HOMER. Nearest genes are assigned.

4

Motif Analysis

Known transcription factor binding motifs are identified within peaks using HOMER, MEME, or Regulatory Genomics Toolbox to infer active regulatory networks.

5

Differential Accessibility

Compare chromatin accessibility between conditions using DESeq2 or edgeR. Identify regions with significant changes in accessibility associated with phenotypes.

ATAC-seq has revolutionized chromatin biology by making genome-wide accessibility profiling accessible to nearly any lab, enabling discoveries from developmental biology to precision medicine.

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

Metagenomic Analysis Workflow

1. **Sample Collection**
   Collect environmental sample (soil, water, gut, etc.)

↓

2. **DNA Extraction**
   Extract total DNA from all organisms in the sample

↓

3. **Library Preparation**
   16S amplification OR shotgun library construction

↓

4. **Sequencing**
   High-throughput sequencing (Illumina, PacBio, Oxford Nanopore)

↓

5 **Quality Control**
Remove adapters, filter low-quality reads, remove host contamination

↓

6 **Bioinformatic Analysis**
Taxonomic classification and/or functional annotation

↓

7 **Data Interpretation**
Statistical analysis, visualization, biological insights

Detailed Approach Comparison

16S rRNA Sequencing

Target Region:
16S ribosomal RNA gene (~1.5 kb)

Coverage:
Bacteria and Archaea only

Resolution:
Genus level (sometimes species)

Cost:
$50-150 per sample

Shotgun Metagenomics

Target Region:
Entire genome (all DNA)

Coverage:
All domains (Bacteria, Archaea, Eukarya, viruses)

Resolution:
Species and strain level

Cost:
$300-1000+ per sample

Data Size:
10,000-50,000 reads per sample

Advantages:
Cost-effective, well-established databases, rapid analysis

Limitations:
No functional information, limited taxonomic resolution, PCR bias

Data Size:
10-100 million reads per sample

Advantages:
Functional profiling, no PCR bias, novel gene discovery, higher resolution

Limitations:
Expensive, requires more computational resources, complex analysis

Detailed Application Examples

1. Clinical Microbiome Studies

Metagenomics enables comprehensive analysis of the human microbiome and its relationship to health and disease. By sequencing microbial communities from various body sites (gut, skin, oral cavity), researchers can identify dysbiosis patterns associated with conditions such as inflammatory bowel disease, obesity, diabetes, and mental health disorders.

> **Example:**
> A study analyzing gut microbiomes of Crohn's disease patients revealed decreased diversity and reduced abundance of beneficial Faecalibacterium prausnitzii, while pathogenic Escherichia coli was enriched. This information guides probiotic therapy development and disease monitoring.

2. Environmental Ecology

Environmental metagenomics assesses microbial diversity and function in natural ecosystems including soil, oceans, freshwater, and extreme environments. This approach reveals how microbial communities drive biogeochemical

cycles (carbon, nitrogen, sulfur), respond to environmental changes, and contribute to ecosystem resilience.

> **Example:**
> Ocean metagenomics discovered the abundant marine bacterium Pelagibacter ubique and revealed novel photosynthetic proteins (proteorhodopsins) that contribute significantly to global carbon cycling. Soil metagenomics identified thousands of antibiotic resistance genes in pristine environments.

3. Industrial Biotechnology

Metagenomics serves as a powerful tool for discovering novel enzymes and metabolic pathways with industrial applications. By screening uncultured microbial communities from diverse environments, researchers identify biocatalysts for chemical synthesis, biodegradation, biofuel production, and other biotechnological processes.

> **Example:**
> Metagenomic screening of hot spring samples yielded thermostable DNA polymerases superior to traditional Taq polymerase. Compost metagenomics discovered cellulases and laccases for biofuel production and textile processing. These discoveries bypass the need to culture organisms in the laboratory.

4. Pathogen Detection & Surveillance

Metagenomic approaches enable culture-independent detection of pathogens in clinical samples, food products, and environmental sources. This is particularly valuable for identifying unknown or emerging infectious agents, monitoring antimicrobial resistance, and investigating disease outbreaks.

> **Example:**

Metagenomic sequencing identified the novel coronavirus SARS-CoV-2 in early 2020. Wastewater metagenomics now tracks COVID-19 variants in communities. Food metagenomics detects Salmonella and E. coli contamination without time-consuming culturing steps.

5. Agriculture & Food Science

Agricultural metagenomics examines soil microbiomes to optimize crop productivity, identifies plant-beneficial microbes for biofertilizers, and characterizes fermented food microbiomes. Understanding these microbial communities helps develop sustainable farming practices and improve food quality.

Example:

Rhizosphere metagenomics identified nitrogen-fixing bacteria and mycorrhizal fungi that enhance plant nutrient uptake. Cheese and wine metagenomics characterized microbial communities responsible for flavor development, leading to better quality control and product consistency.

Bioinformatic Tools in Detail

Kraken2

Ultra-fast taxonomic classifier using exact k-mer matches. Assigns taxonomic labels to DNA sequences by comparing k-mers against a reference database. Processes millions of reads in minutes with high accuracy.

MetaPhlAn4

Computational tool for profiling microbial communities using clade-specific marker genes. Provides species-level resolution with high precision, useful for tracking specific organisms across samples and studies.

QIIME2

Comprehensive platform for microbiome analysis supporting quality control, taxonomic classification, diversity analysis, and

HUMAnN3

Functional profiling tool that characterizes metabolic pathways and gene families in metagenomic samples. Maps reads to

statistical testing. Particularly popular for 16S rRNA data with extensive visualization capabilities.

reference databases (UniRef, KEGG) to determine community functional potential.

MEGAHIT / metaSPAdes

De novo assemblers for reconstructing longer contigs and genomes from short metagenomic reads. Essential for discovering novel organisms and genes not present in reference databases.

CheckM2

Quality assessment tool for evaluating completeness and contamination of metagenome-assembled genomes (MAGs). Uses machine learning to estimate genome quality across diverse taxonomic groups.

DIAMOND

High-performance sequence aligner for protein and translated DNA searches. Up to 20,000x faster than BLASTX, making it essential for functional annotation of large metagenomic datasets.

Kaiju

Protein-level taxonomic classifier that translates DNA reads and compares them to protein databases. Particularly useful for detecting divergent or poorly characterized organisms.

Key Concepts & Terminology

- **Alpha Diversity:**Diversity within a single sample (richness and evenness of species)
- **Beta Diversity:**Diversity between samples (compositional differences)
- **Operational Taxonomic Unit (OTU):**Cluster of similar sequences (typically 97% identity for 16S)
- **Amplicon Sequence Variant (ASV):** Unique exact sequence from amplicon data (higher resolution than OTUs)
- **Metagenome-Assembled Genome (MAG):** Reconstructed genome from metagenomic assembly and binning
- **Read Depth:**Number of sequencing reads covering a genomic position
- **Taxonomic Profiling:**Identifying "who is there" in the community

- **Functional Profiling:**Identifying "what they can do" (metabolic potential)

Current Challenges in Metagenomics

Computational Challenges

- Massive data volumes (100+ GB per sample)
- High memory and processing requirements
- Long analysis pipelines
- Need for specialized infrastructure

Biological Challenges

- Unknown/unculturable organisms
- Incomplete reference databases
- Horizontal gene transfer complexity
- Strain-level variation

Technical Challenges

- DNA extraction bias
- PCR amplification bias (16S)
- Short read limitations
- Host DNA contamination

Analytical Challenges

- Distinguishing contamination
- Batch effects between studies
- Causation vs correlation
- Standardization across labs

Future Directions in Metagenomics

- Long-read sequencing (PacBio HiFi, Oxford Nanopore) for complete genome assembly
- Single-cell genomics combined with metagenomics
- Meta-transcriptomics and meta-proteomics for active function assessment
- Machine learning for pattern recognition and prediction

- Real-time metagenomic monitoring (portable sequencers)
- Multi-omics integration (metagenomics + metabolomics + metatranscriptomics)

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

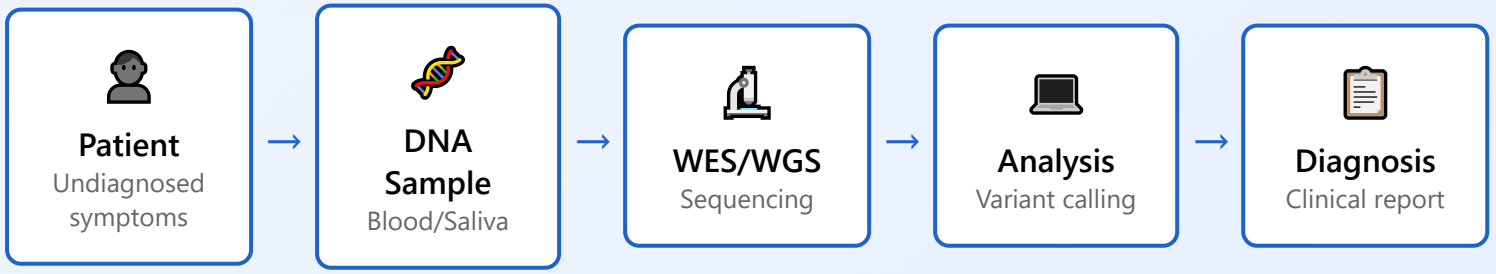
- Medical necessity
- Prior authorization

Requires multidisciplinary team: clinicians, geneticists, bioinformaticians, counselors

Detailed Clinical Applications & Examples

1 Diagnosis of Rare Genetic Diseases

Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS) have revolutionized the diagnosis of rare genetic disorders. These technologies enable comprehensive analysis of all protein-coding genes or the entire genome, identifying pathogenic variants that cause disease.



Clinical Case Example

Patient: 6-year-old boy with developmental delay, seizures, and intellectual disability
Previous testing: Karyotype, microarray - negative
WES Result: De novo pathogenic variant in SCN1A gene
Diagnosis: Dravet syndrome
Impact: Changed treatment plan, avoided ineffective/harmful medications, genetic counseling for family

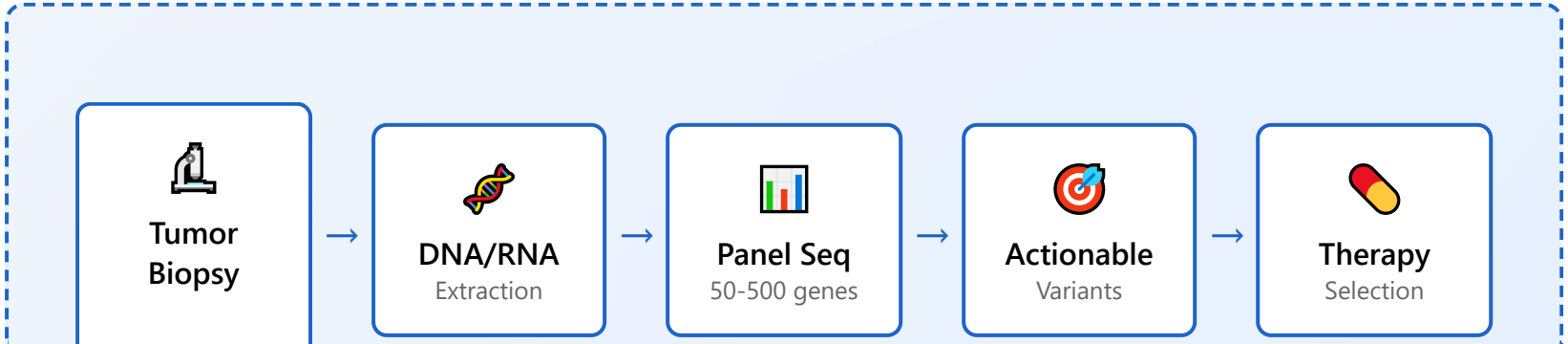
Common Genes Analyzed in Rare Disease Panels

SCN1A | MECP2 | SMN1 | CFTR | DMD | FMR1 | COL4A5 | BRCA1/2

25-50%	~7,000	3-6 weeks
Diagnostic Yield for WES	Known Rare Diseases	Typical TAT

2 Cancer Precision Medicine & Treatment Selection

Tumor sequencing identifies somatic mutations, copy number variations, and gene fusions that drive cancer growth. This information guides targeted therapy selection, predicts treatment response, and monitors disease progression through liquid biopsies.



Tissue/Blood

📌 Clinical Case Example

Patient: 58-year-old woman with stage IV non-small cell lung cancer (NSCLC)
Tumor sequencing: Comprehensive cancer panel (468 genes)
Key findings: EGFR exon 19 deletion, TMB-high (15 mutations/Mb)
Treatment: First-line EGFR TKI (osimertinib) → significant tumor reduction
Monitoring: ctDNA liquid biopsy for resistance mutations (T790M)

```
> Actionable Variants Detected:
EGFR: c.2235_2249del15 (p.Leu747_Thr751del) - PATHOGENIC
↳ FDA-approved therapy: Osimertinib, Erlotinib, Gefitinib
↳ Evidence level: 1A (NCCN Guidelines)

PD-L1 expression: 60% TPS
↳ Eligible for pembrolizumab combination therapy

TMB: 15.2 mutations/Mb (TMB-High)
↳ Potential benefit from immunotherapy
```

Common Targetable Alterations by Cancer Type

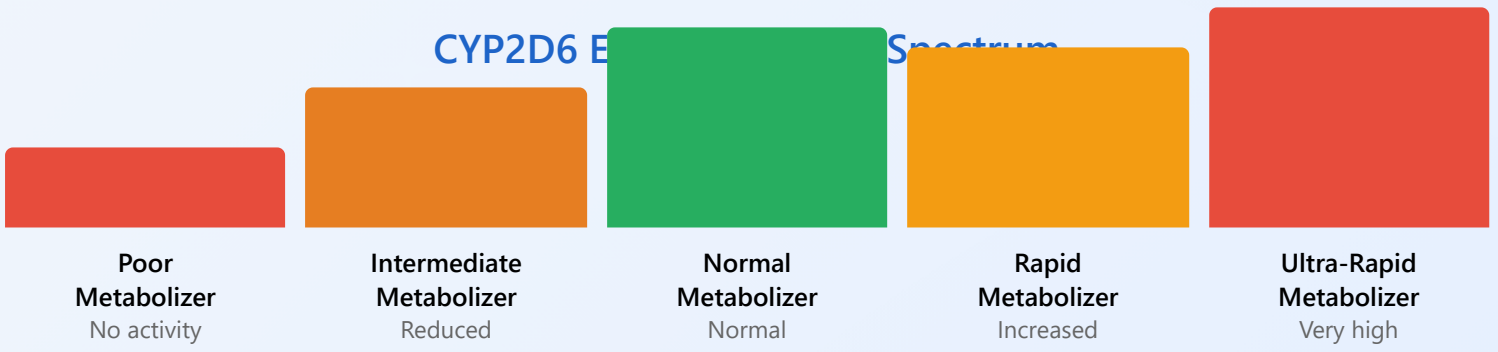
Lung Cancer: EGFR, ALK, ROS1, BRAF, MET, KRAS G12C
Breast Cancer: HER2, PIK3CA, ESR1, BRCA1/2, PALB2
Colorectal: KRAS, NRAS, BRAF V600E, MSI-H, TMB-H
Melanoma: BRAF V600E/K, NRAS, KIT, NF1
Ovarian: BRCA1/2, HRD score, CCNE1 amplification

30-40%	**50-75%**	**7-10 days**
Patients with Actionable Mutations		Typical TAT

Improved Response with Targeted Therapy

3 Pharmacogenomics - Drug Response Prediction

Pharmacogenomic testing analyzes genetic variants in genes encoding drug-metabolizing enzymes, transporters, and drug targets. This enables personalized medication selection and dosing to maximize efficacy and minimize adverse reactions.



📌 Clinical Case Example

Patient: 45-year-old woman starting antidepressant therapy
PGx Testing: CYP2D6, CYP2C19, CYP3A4/5, SLCO1B1
Results:
• CYP2D6: *4/*4 (Poor metabolizer)
• CYP2C19: *1/*17 (Rapid metabolizer)
Interpretation:
• Avoid codeine (no therapeutic effect), tramadol

• Reduce dose of metoprolol by 75%
• Standard dose clopidogrel appropriate

Drug Selection: Venlafaxine selected over paroxetine (CYP2D6 substrate)

Key Pharmacogenes and Associated Drugs

CYP2D6:	Codeine, tramadol, metoprolol, paroxetine, tamoxifen
CYP2C19:	Clopidogrel, omeprazole, escitalopram, voriconazole
CYP2C9:	Warfarin, phenytoin, NSAIDs, losartan
TPMT:	Azathioprine, mercaptopurine, thioguanine
SLCO1B1:	Statins (simvastatin, atorvastatin)
DPYD:	5-fluorouracil, capecitabine (cancer therapy)
VKORC1:	Warfarin dosing
HLA-B*57:01:	Abacavir hypersensitivity

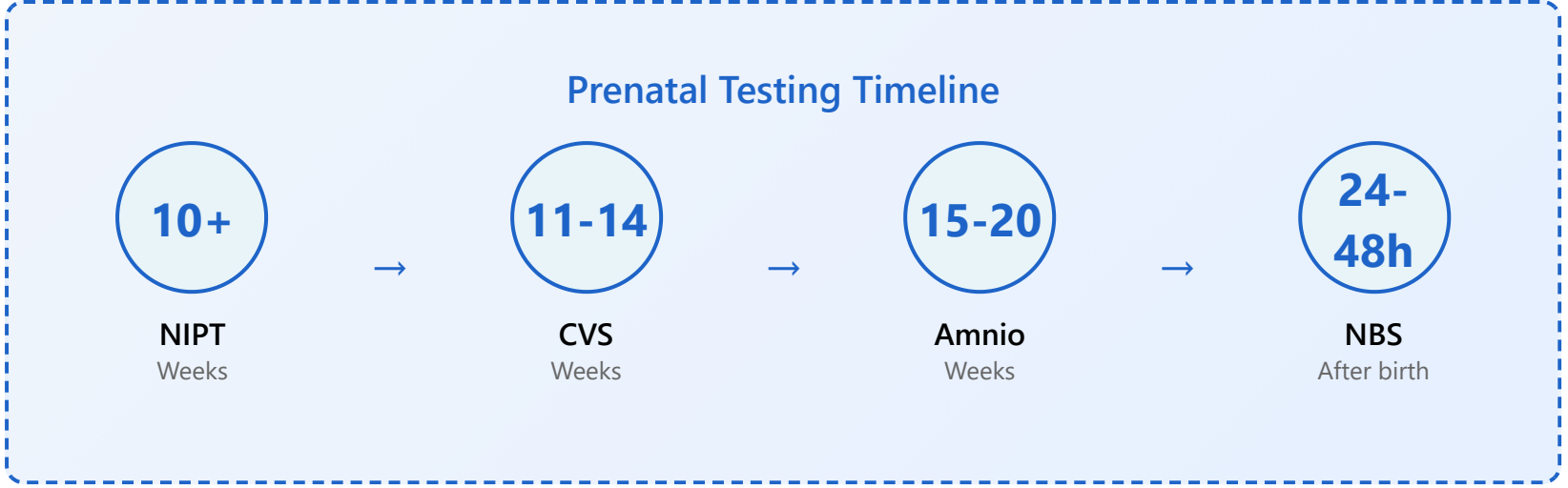
95%
Population with Actionable PGx Variant

30%
ADR Prevention Rate

270+
FDA PGx Drug Labels

4 Prenatal and Newborn Screening

Non-invasive prenatal testing (NIPT) analyzes cell-free fetal DNA in maternal blood to screen for chromosomal abnormalities. Newborn sequencing enables early detection and treatment of genetic disorders before symptoms appear.



📌 Clinical Case Example - NIPT

Patient: 38-year-old pregnant woman, 12 weeks gestation
Test: cfDNA NIPT from maternal blood
Results: Elevated risk for Trisomy 21 (Down syndrome)
Follow-up: Diagnostic amniocentesis confirmed T21
Outcome: Genetic counseling, preparation for specialized care

Detection rates: T21 (99%), T18 (97%), T13 (91%)

📌 Clinical Case Example - NBS

Patient: 48-hour-old newborn
Screen: State mandated newborn screening panel
Positive result: Elevated C8 acylcarnitine
Diagnosis: Medium-chain acyl-CoA dehydrogenase deficiency (MCADD)
Treatment initiated: Avoid fasting, high-carb diet, emergency protocol
Impact: Prevention of potentially fatal metabolic crisis

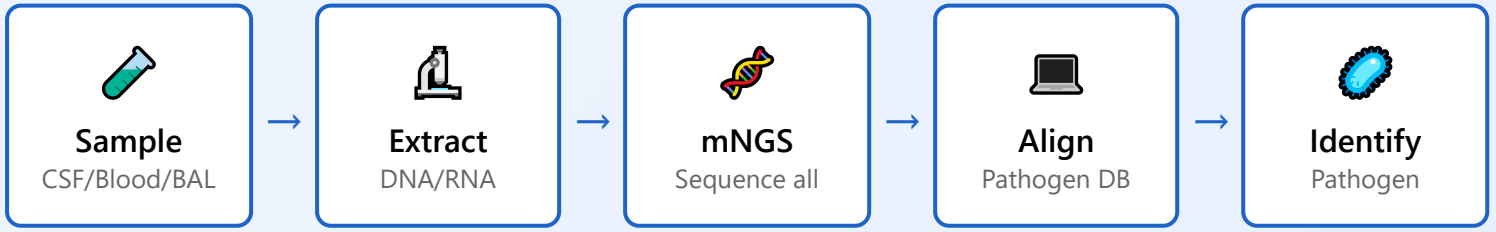
Conditions Screened in Newborns (US Recommended Core Panel)

Organic Acid Disorders: Propionic acidemia, Methylmalonic acidemia, Isovaleric acidemia
Fatty Acid Oxidation: MCADD, VLCADD, LCHAD deficiency
Amino Acid Disorders: PKU, Maple syrup urine disease, Homocystinuria
Hemoglobinopathies: Sickle cell disease, Beta-thalassemia
Endocrine: Congenital hypothyroidism, CAH
Other: Biotinidase deficiency, Galactosemia, Cystic fibrosis, SCID

99%	35+	1 in 300
NIPT Sensitivity for T21	NBS Core Conditions	NBS Detection Rate

5 Infectious Disease Identification

Metagenomic Next-Generation Sequencing (mNGS) enables unbiased detection of all microbial DNA/RNA in clinical samples. This approach identifies pathogens without prior knowledge or culture, crucial for diagnosing unusual or fastidious infections.



Clinical Case Example

Patient: 14-year-old boy with encephalitis, seizures, altered consciousness
Initial testing: Bacterial culture, viral PCR panel - all negative
mNGS (CSF): Performed after 1 week of empiric treatment
Results: *Balamuthia mandrillaris* detected (1,237 reads mapped)
Diagnosis: Rare amoebic encephalitis
Treatment: Switched to appropriate anti-parasitic therapy
TAT: 48 hours from sample to result

```
> mNGS Report Summary:
Total reads: 28,456,891
Human reads: 28,442,108 (99.95%)
Non-human reads: 14,783 (0.05%)

PATHOGEN DETECTED:
Organism: Balamuthia mandrillaris
Reads mapped: 1,237
Genome coverage: 12.4%
Confidence: HIGH
Clinical significance: PATHOGENIC - causes GAE
```

Clinical Applications of Infectious Disease Sequencing

Meningitis/Encephalitis: Rapid pathogen ID when culture-negative
Sepsis: Blood culture-independent pathogen detection
Pneumonia: Identification of atypical/fastidious organisms
Immunocompromised: Detection of opportunistic infections
Outbreak investigation: Strain typing and transmission tracking
Antimicrobial resistance: Detection of resistance genes
HIV/HCV: Viral load monitoring, resistance mutations
TB: *M. tuberculosis* detection and drug resistance profiling

Advantages over Traditional Methods

✓ Culture-independent (detects non-cultivable organisms)
✓ Unbiased (no prior hypothesis needed)
✓ Rapid results (24-48 hours vs days/weeks for culture)
✓ Detects co-infections

✓ Identifies novel/unexpected pathogens
✓ Simultaneous resistance gene detection
✓ Works with small sample volumes

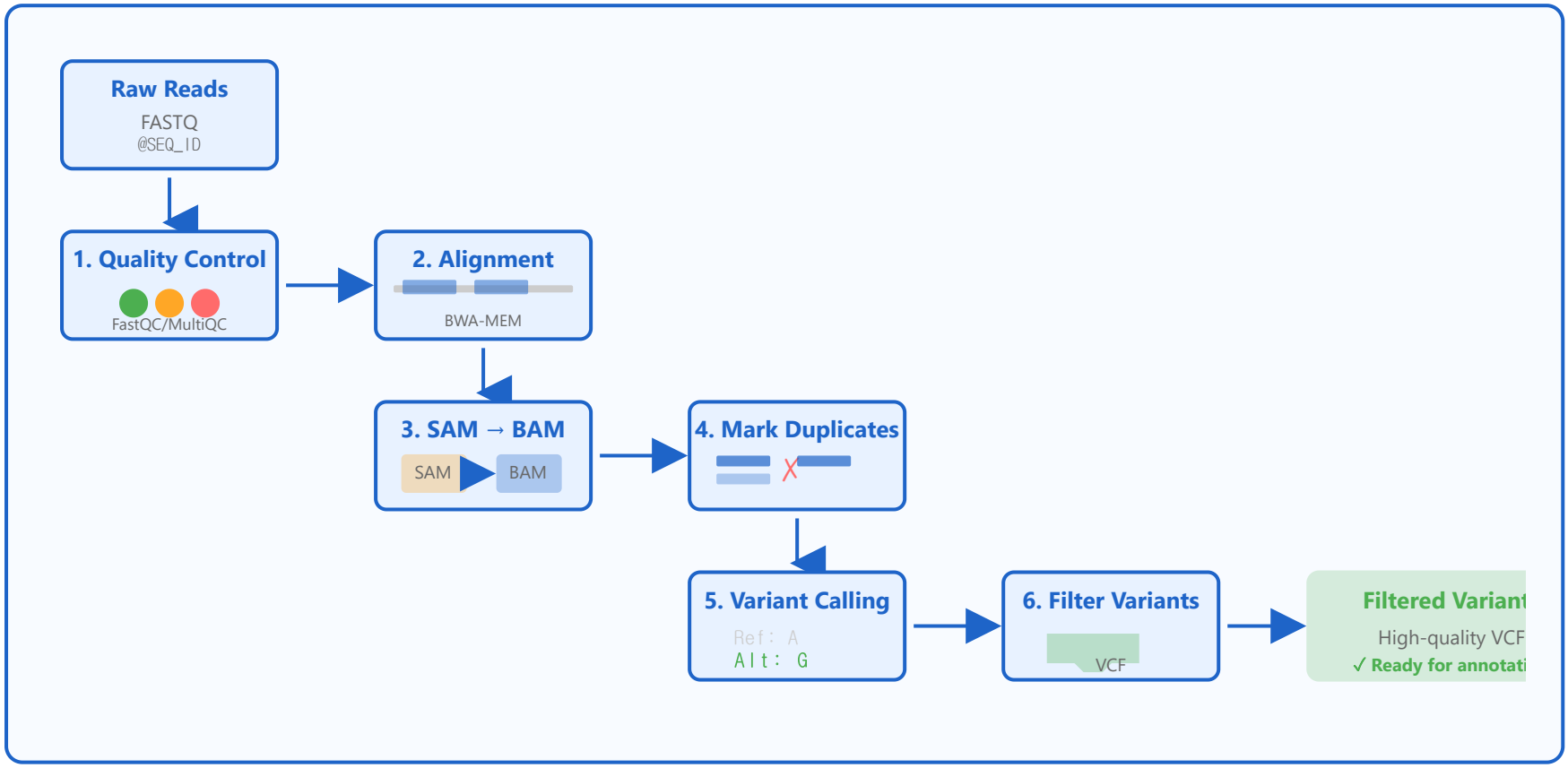
24-48h
mNGS Turnaround Time

40-70%
Diagnostic Yield in CNS Infections

1000s
Detectable Pathogens

Clinical sequencing has transformed medicine by enabling precision diagnosis and treatment. Each application requires rigorous validation, quality control, and multidisciplinary interpretation to translate genomic data into actionable clinical decisions.

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