The RNA-seq Protocol

Erin Osborne Nishimura
DSCI 512:
RNA-seq data analysis
November 5, 2024

The evolving face of High Throughput Sequencing

- **Short reads** – Illumina (aka **nextgen** sequencing)

- **Third gen sequencing** -- Long reads -- Pac Bio & minION (Oxford Nanopore Technologies)

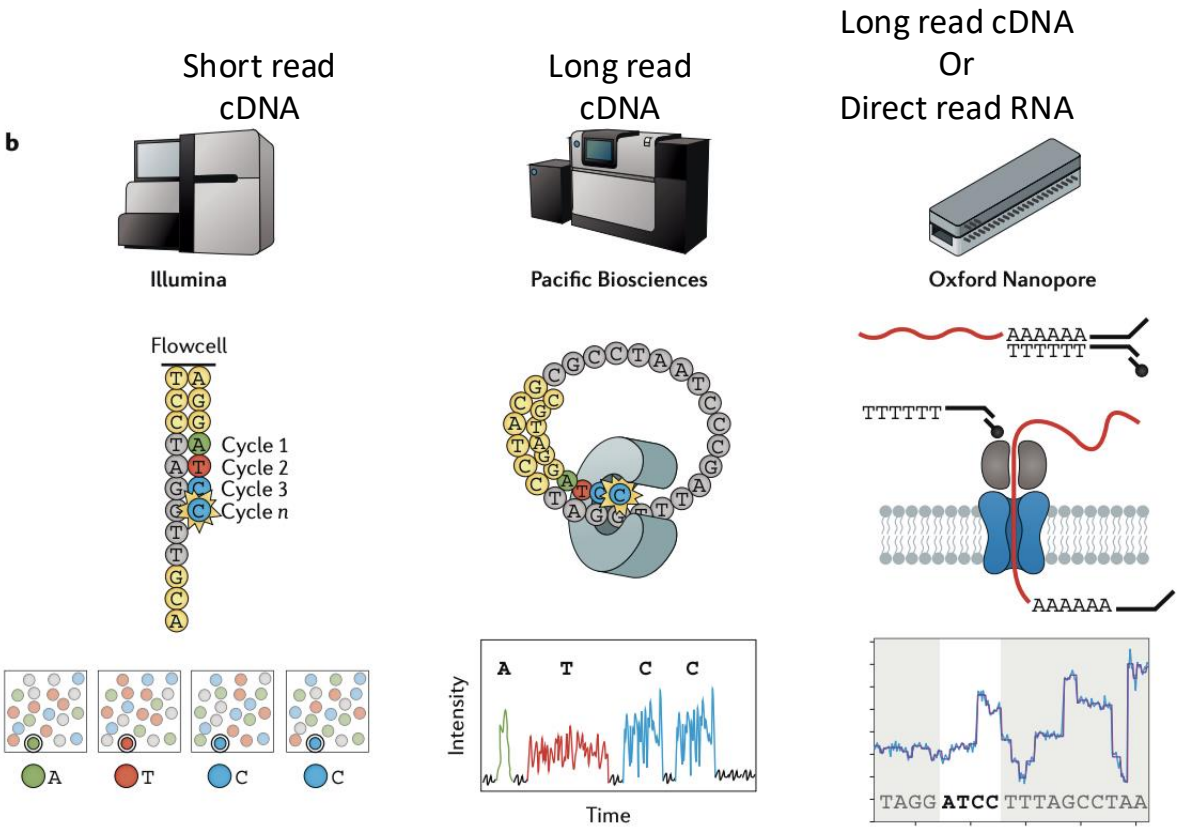
Genomics

Three flavors of high-throughput sequencing

Short read cDNA

Long read cDNA

Long read cDNA
Or
Direct read RNA



illumina®

PACBIO®

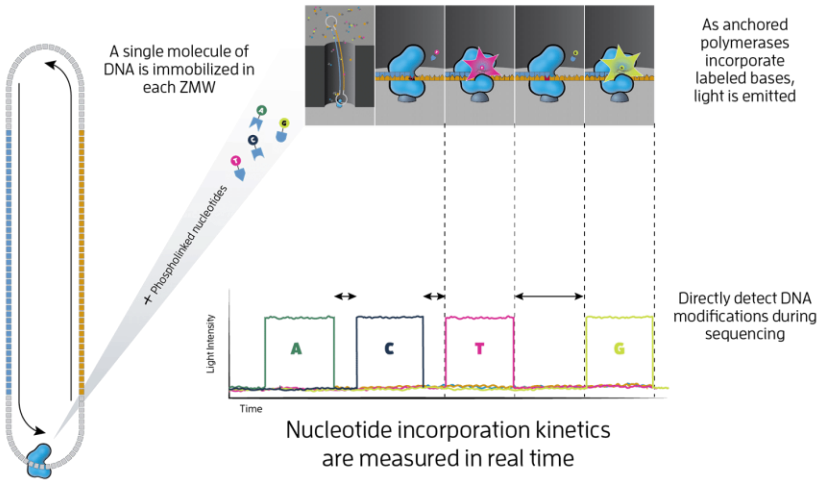
Oxford NANOPORE Technologies®

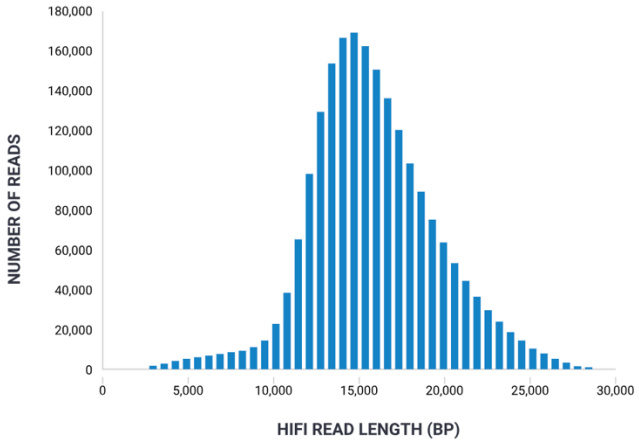
https://www.nature.com/articles/s41576-019-0150-2.pdf

Pac Bio – circular long-read sequencing

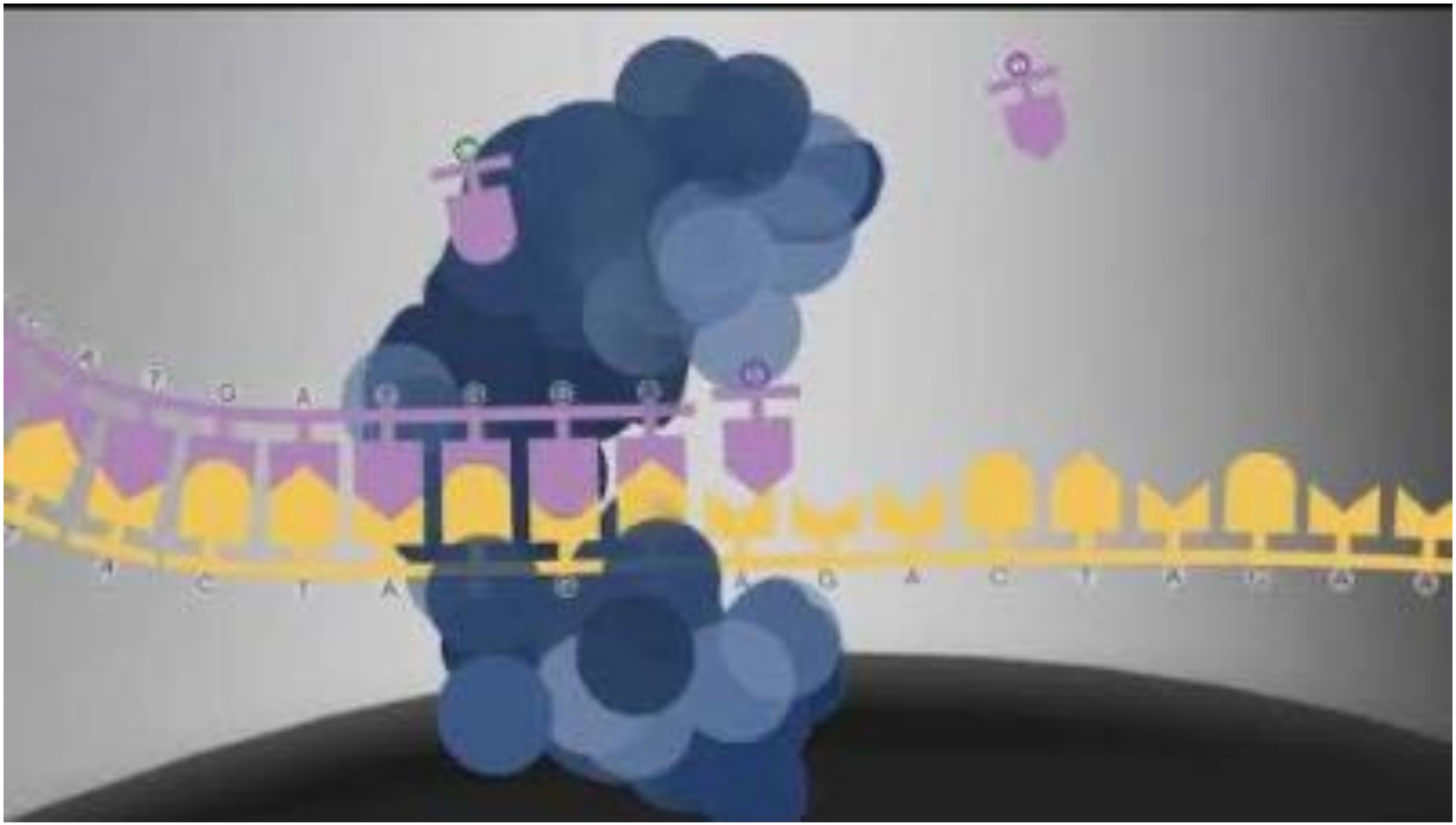
- Immobilized DNA Polymerases at the bottom of each well amplify a single circularized DNA molecule.
- Colored light is emitted for each nucleotide

- Long reads are produced
- Errors are minimized by going around each circle multiple times





Explore Pac Bio Sequencing on your own



https://youtu.be/_lD8JyAbwEo?si=-4pI3dLq5AoMlmeZ

https://www.pacb.com/blog/long-read-sequencing/

Oxford Nanopore – MinION

Long reads

Discover and phase hidden variation — from repetitive regions and structural variants to novel, full-length transcript isoforms

High yields

As much as 30 Gb* data suitable for all applications — from whole genomes and transcriptomes to high-throughput targeted analyses

Real time

Immediate access to actionable results — from pathogen and antimicrobial resistance identification to fusion transcripts

Portable

Sequence samples at source — combine with VolTRAX™ and MinIT™ for portable sample prep and analysis

Accessible

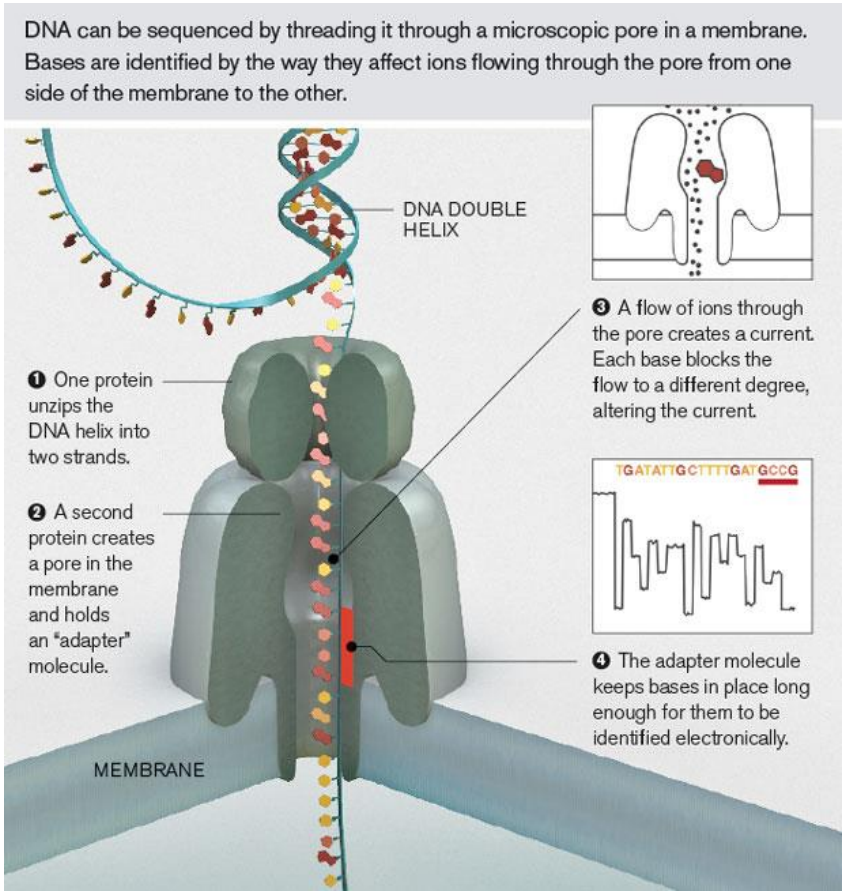
Starter Packs from just $1,000 — with no capital investment or complex IT infrastructure required

Direct

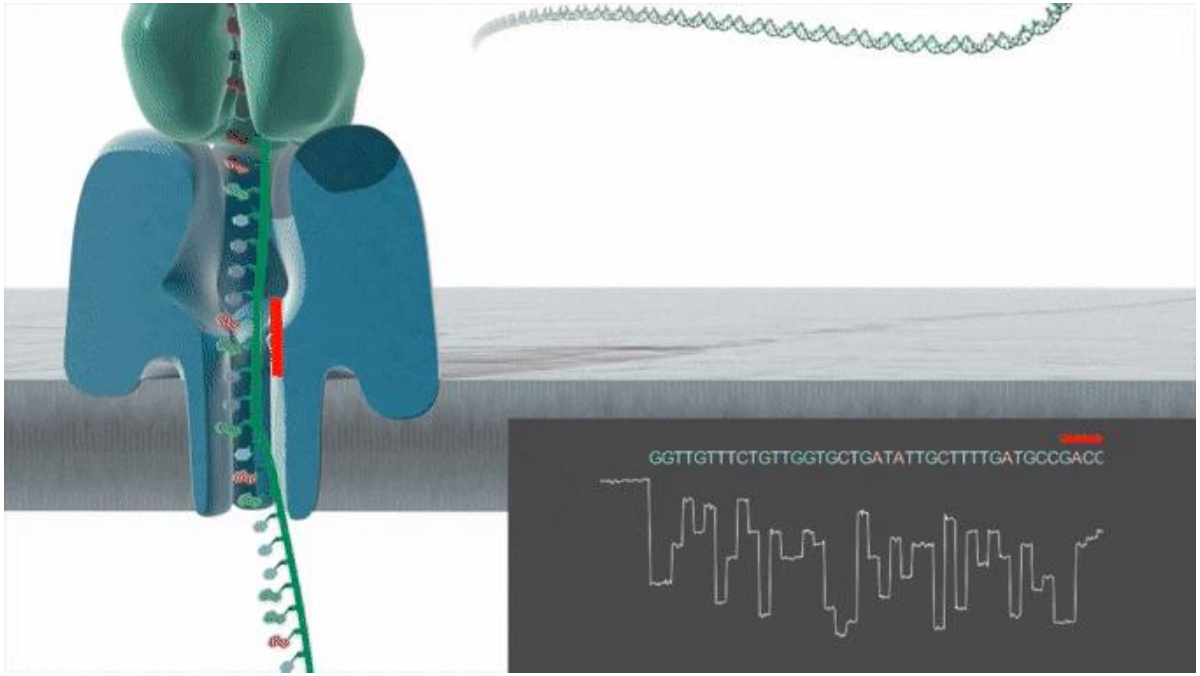
Study native DNA and RNA, not a copy — eliminate amplification bias and detect base modifications

Oxford Nanopore sequencing technology, a third-generation sequencing technology

- minION
- A single, long, unamplified DNA molecule is dissociated into a single strand and fed through a small protein pore in a membrane.
- As the DNA passes through the membrane, a small current is created that is distinct for each nucleotide.
- The ionic flow signature of each nucleotide can be read through an extremely sensitive detector

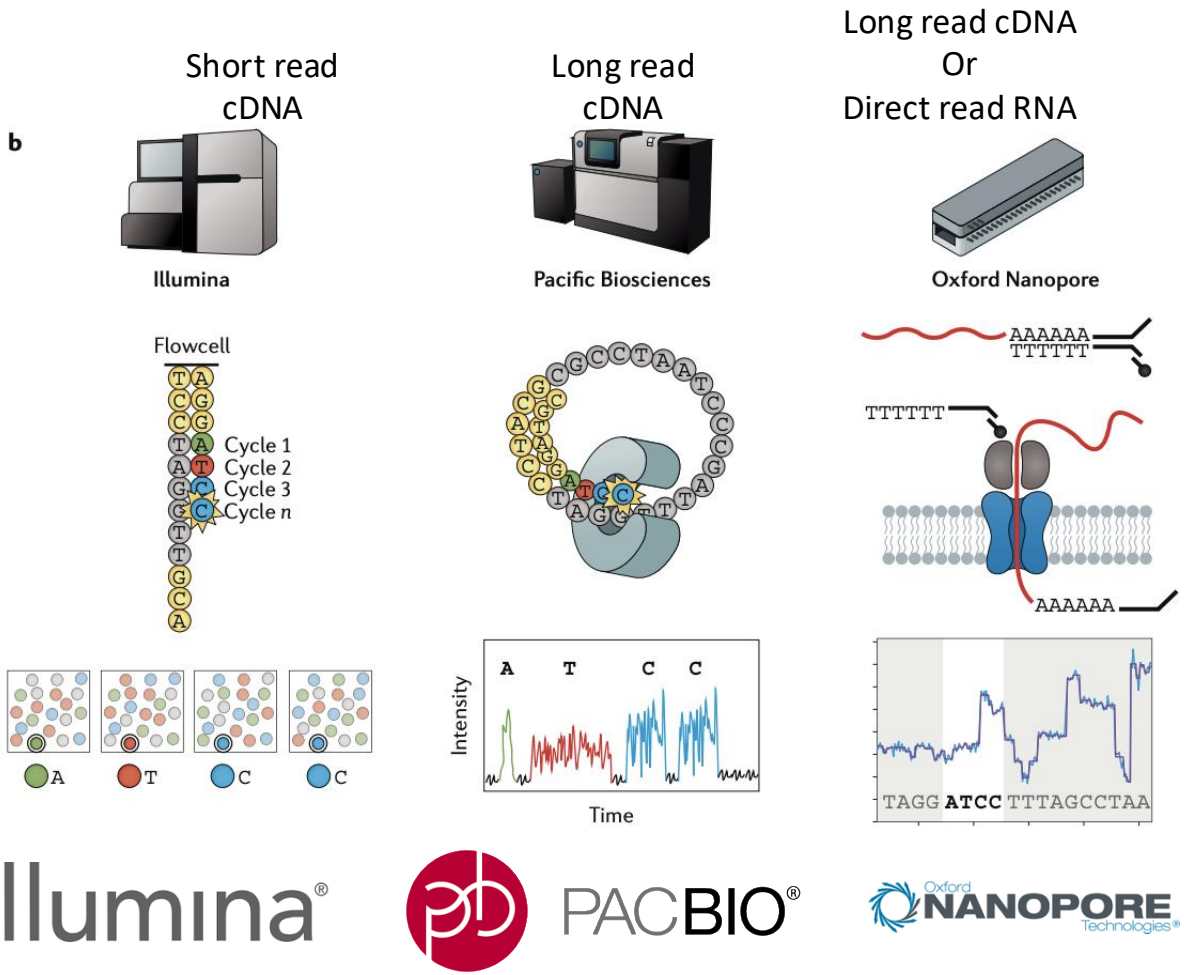


Nanopore sequencing technology



Genomics

Three flavors of high-throughput sequencing



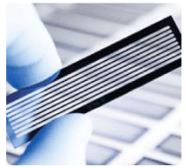
illumina®

PACBIO®

Oxford NANOPORE Technologies®

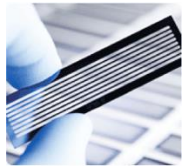
https://www.nature.com/articles/s41576-019-0150-2.pdf

Illumina Sequencing



	Benchtop Sequencers			P -Scale Sequencers	
	 iSeq 100	 MiniSeq	 MiSeq Series ⊕	 NextSeq 550 Series ⊕	 NextSeq 1000 & 2000
Popular Applications & Methods	Key Application	Key Application	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)					
Small Whole-Genome Sequencing (microbe, virus)	●	●	●	●	●
Exome & Large Panel Sequencing (enrichment-based)				●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●	●	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)				●	●
Run Time	9.5–19 hrs	4–24 hours	4–55 hours	12–30 hours	11-48 hours
Maximum Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	360 Gb*
Maximum Reads Per Run	4 million	25 million	25 million †	400 million	1.2 billion*
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp

Illumina Sequencing



School of Medicine
UNIVERSITY OF COLORADO
ANSCHUTZ MEDICAL CAMPUS

	Benchtop Sequencers	Production-Scale Sequencers	
	 NextSeq 550 Series	 NextSeq 1000 & 2000	 NovaSeq 6000
Popular Applications & Methods	Key Application	Key Application	Key Application
Large Whole-Genome Sequencing (human, plant, animal)			●
Small Whole-Genome Sequencing (microbe, virus)	●	●	●
Exome & Large Panel Sequencing (enrichment-based)	●	●	●
Targeted Gene Sequencing (amplicon-based, gene panel)	●	●	●
Single-Cell Profiling (scRNA-Seq, scDNA-Seq, oligo tagging assays)	●	●	●
Transcriptome Sequencing (total RNA-Seq, mRNA-Seq, gene expression profiling)	●	●	●
Run Time	12–30 hours	11-48 hours	~13 - 38 hours (dual SP flow cells) ~13–25 hours (dual S1 flow cells) ~16–36 hours (dual S2 flow cells) ~44 hours (dual S4 flow cells)
Maximum Output	120 Gb	360 Gb*	6000 Gb
Maximum Reads Per Run	400 million	1.2 billion*	20 billion
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 x 250**

Illumina short read sequencing: The basic protocol

RNA-seq library prep

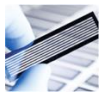
NEXTGEN sequencing

1. total RNA

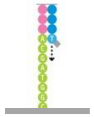
1. mRNA

1. cDNA

1. Library prep



2. Cluster generation



3. Sequencing by Synthesis

4. Data analysis

- Total RNA extracted
- mRNA (2% of total) is enriched
- RNA converted to cDNA using Reverse Transcriptase
- NextGen sequenced
- Sequences are aligned to the genome. Exons will be represented.
- RNA processing and library prep are typically done with kits (NEB Next RNA-seq kit)
- NEXTGEN sequencing is often done at a core facility

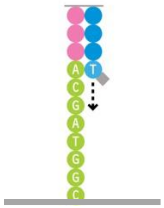
Basic steps of Illumina sequencing

1. Library preparation

- fragment the DNA
- ligate adapters onto fragments

2. Cluster generation

- adhere fragments to a flow cell
- amplify them



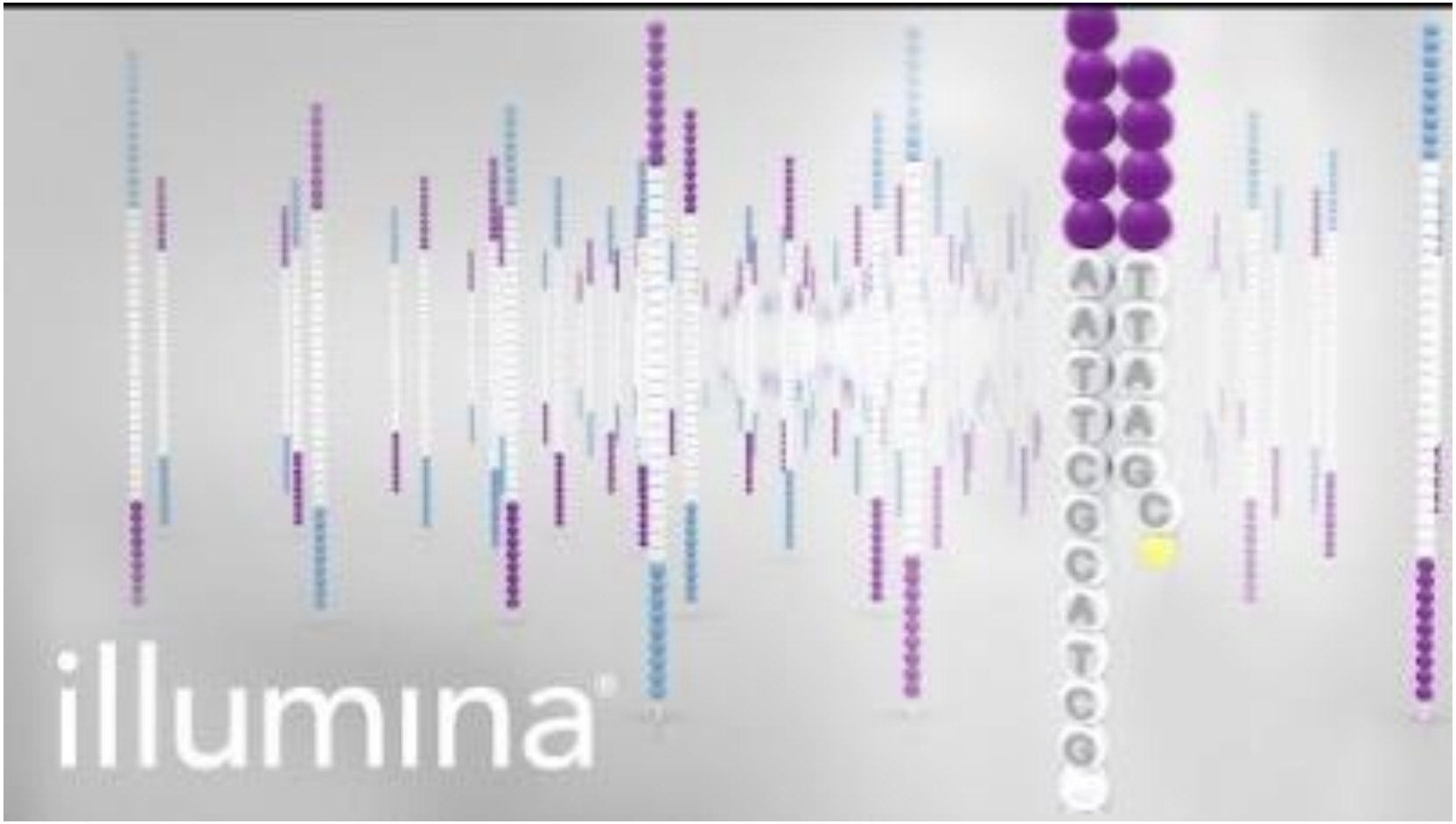
3. Sequencing by synthesis

- directly image a specialized polymerization reaction that allows visualization of each nucleotide incorporated
- Tens of millions of fragments are sequenced

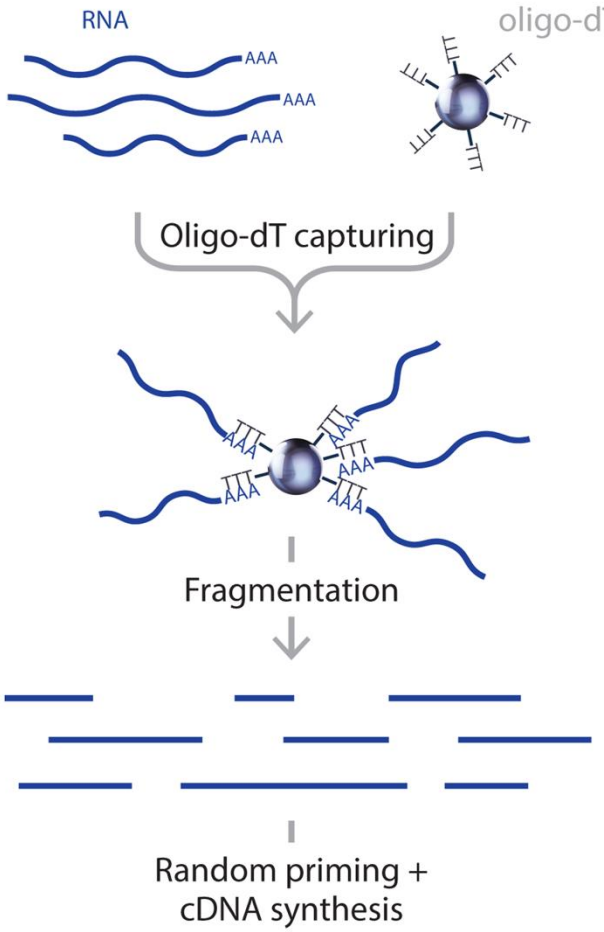
ATGCGGTTATAGTATA
TGCCACCTAATTTCCC
CAGGTCCCCTTTGTTA
ATGCGGTTATAGTATA
TGCCACCTAATTTCCC

4. Data analysis

- Use computer algorithms to align sequences to the genome


illumina

Step 1: Library preparation

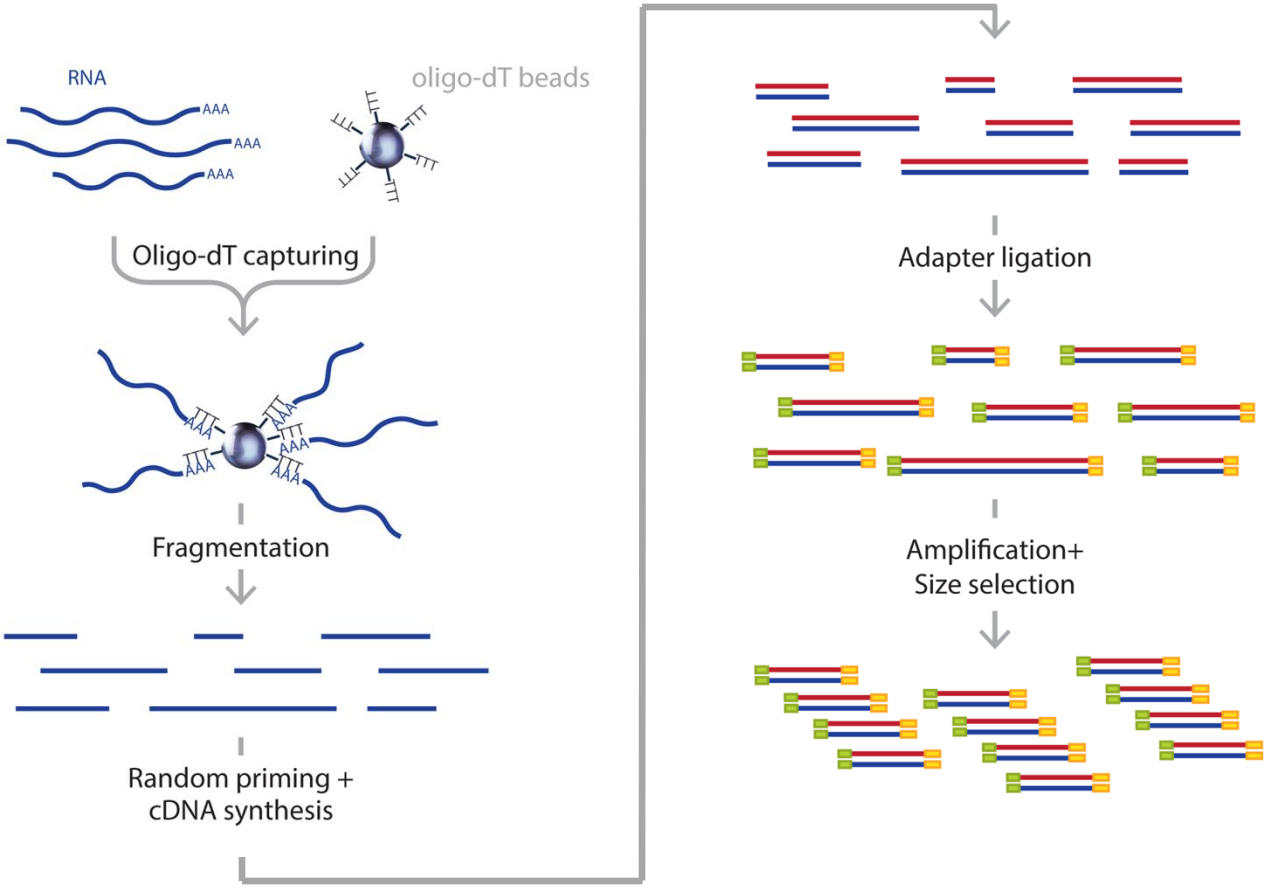


- Total RNA is extracted
- mRNA is enriched either by mRNA enrichment or rRNA subtraction
  - Oligo dT beads
  - Ribozero, Ribominus, Ribogone
- mRNA is fragmented through heat + ions
- cDNA is produced by random oligo or oligo dT priming

Zhernakova et al. (2009)

Step 1 – Adding adapters

Adapters are ligated to the cDNA pieces. These are like handles that allow downstream instrumentation to interact with each molecule.

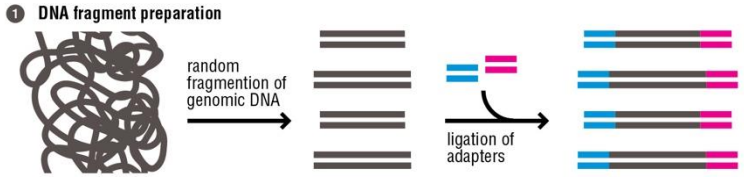


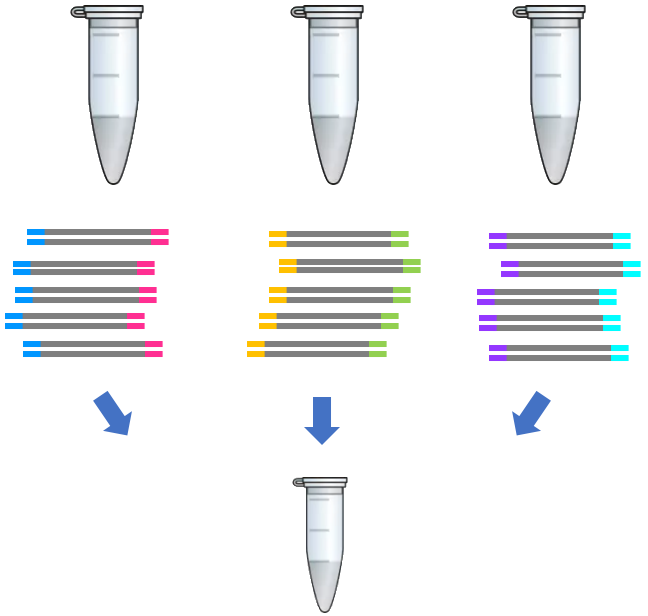
Zhernakova et al. (2009)

Please watch this video on your own for homework:

- https://www.neb.com/en-us/tools-and-resources/video-library?device=modal&videoid=%7Bd824c8c5-7942-437c-9086-e93ff3c94a12%7D
- A link is on your homework assignment page

Step 1 - Multiplexing: Multiple samples can be sequenced on the same flow cell





- Each sample is ligated to a specific adapter pair with its own sequence
- Samples are then are merged and run on the same flow cell
- Samples are split computationally using unique adapter sequences

Adapters: A closer look

Flow cell binding sequence: Platform-specific sequences for library binding to instrument

Sequencing primer sites: Binding sites for general sequencing primers

Sample indexes: Short sequences specific to a given sample library

Insert: Target DNA or RNA fragment from a given sample library

The aim of the sample prep step is to obtain nucleic acid fragments with adapters attached on both ends

RNA-seq

RNA-seq library prep

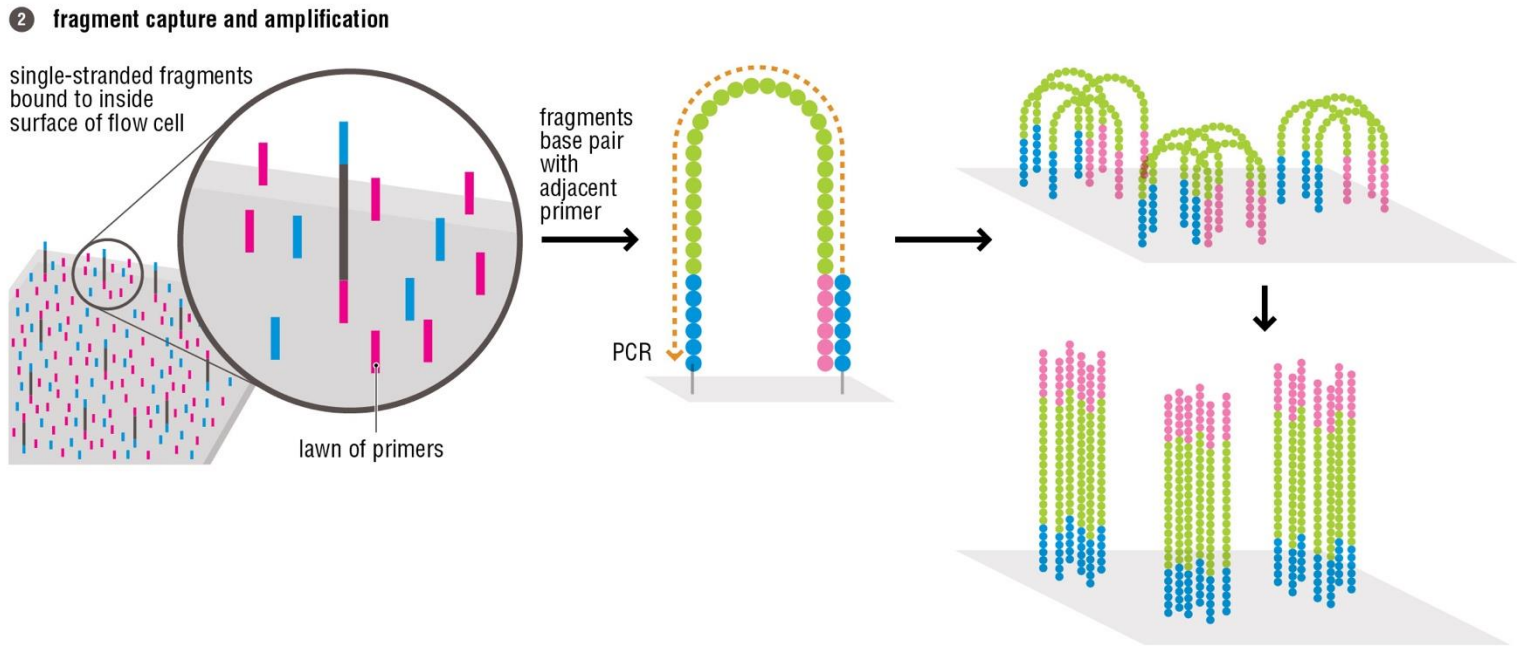
NEXTGEN sequencing

- Total RNA extracted
- mRNA (2% of total) is enriched
- RNA converted to cDNA using Reverse Transcriptase
- NextGen sequenced
- Sequences are aligned to the genome. Exons will be represented.
- RNA processing and library prep are typically done with kits (NEB Next RNA-seq kit)
- NEXTGEN sequencing is often done at a core facility

Illumina Sequencing: Step 2

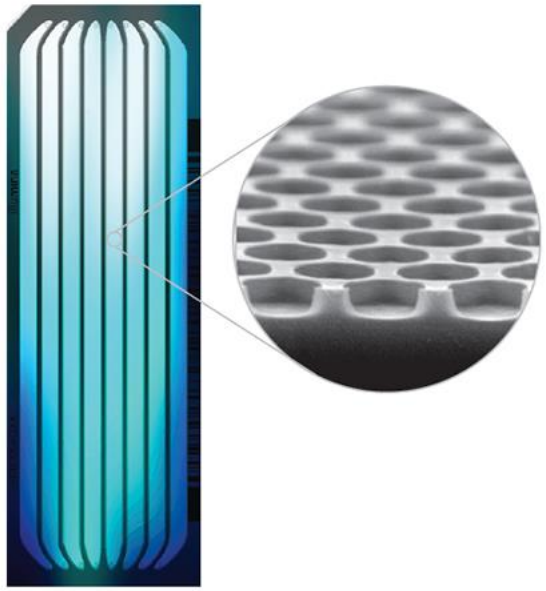
2. Cluster generation

- adhere fragments to a flow cell
- amplify them

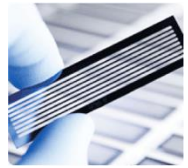


Craig et al., Molecular Biology; Figure 19.94

Illumina Sequencing: Step 2



Illumina Sequencing: Step 2



iSeq 100

MiniSeq

MiSeq Series ⊕

NextSeq 550 Series ⊕

NextSeq 1000 & 2000

NextSeq 550 Series ⊕

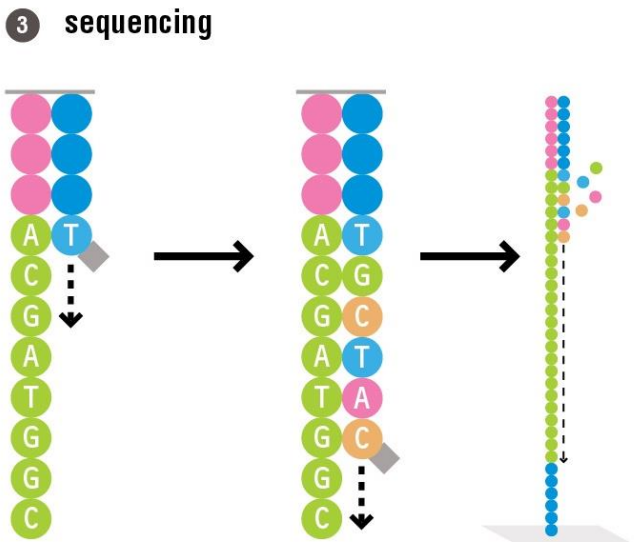
NextSeq 1000 & 2000

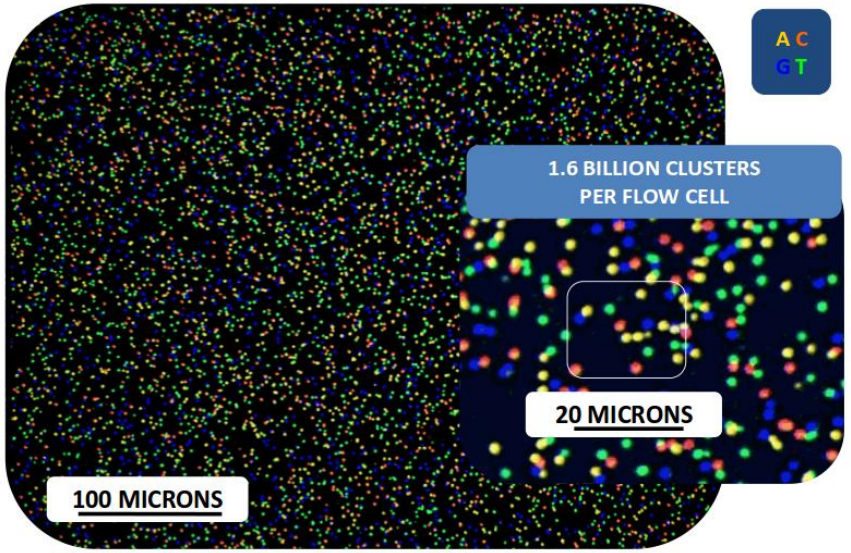
NovaSeq 6000

Illumina Sequencing: Step 3

3. Sequencing by synthesis

- Sequence using reversible dye termination.
- directly image the sequencing reaction
- Tens of millions of fragments are sequenced

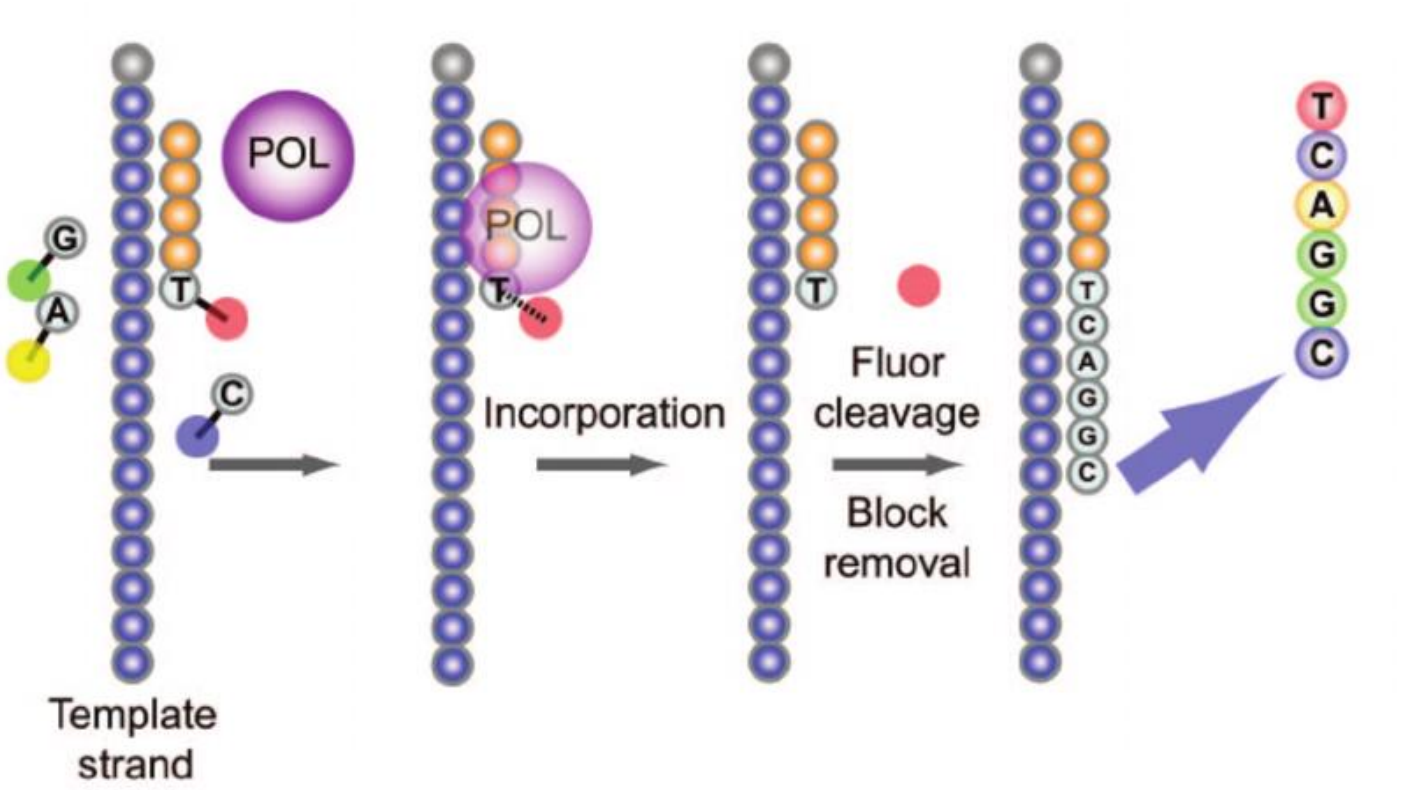




Craig et al., Molecular Biology; Figure 19.94

Illumina Sequencing: Step 3

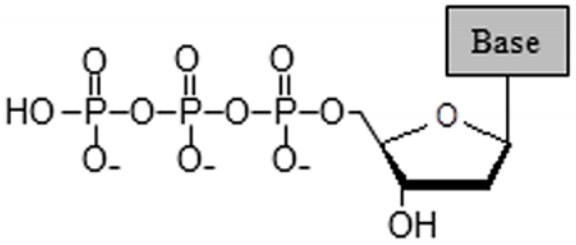
Reversible dye termination incorporates one fluorescently labeled nucleotide at a time



Illumina Sequencing: Step 3

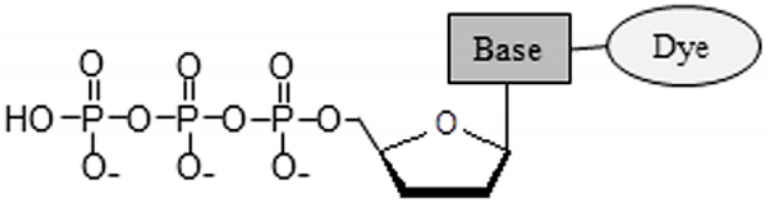
Reversible dye terminators

A



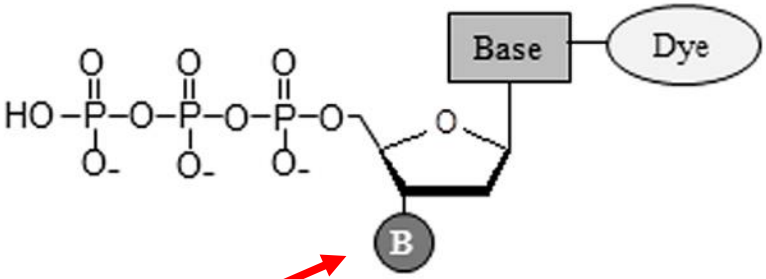
A regular deoxynucleotide (dNTP)

C



A dye-labeled dideoxy nucleotide (ddNTP) used in Sanger sequencing

D



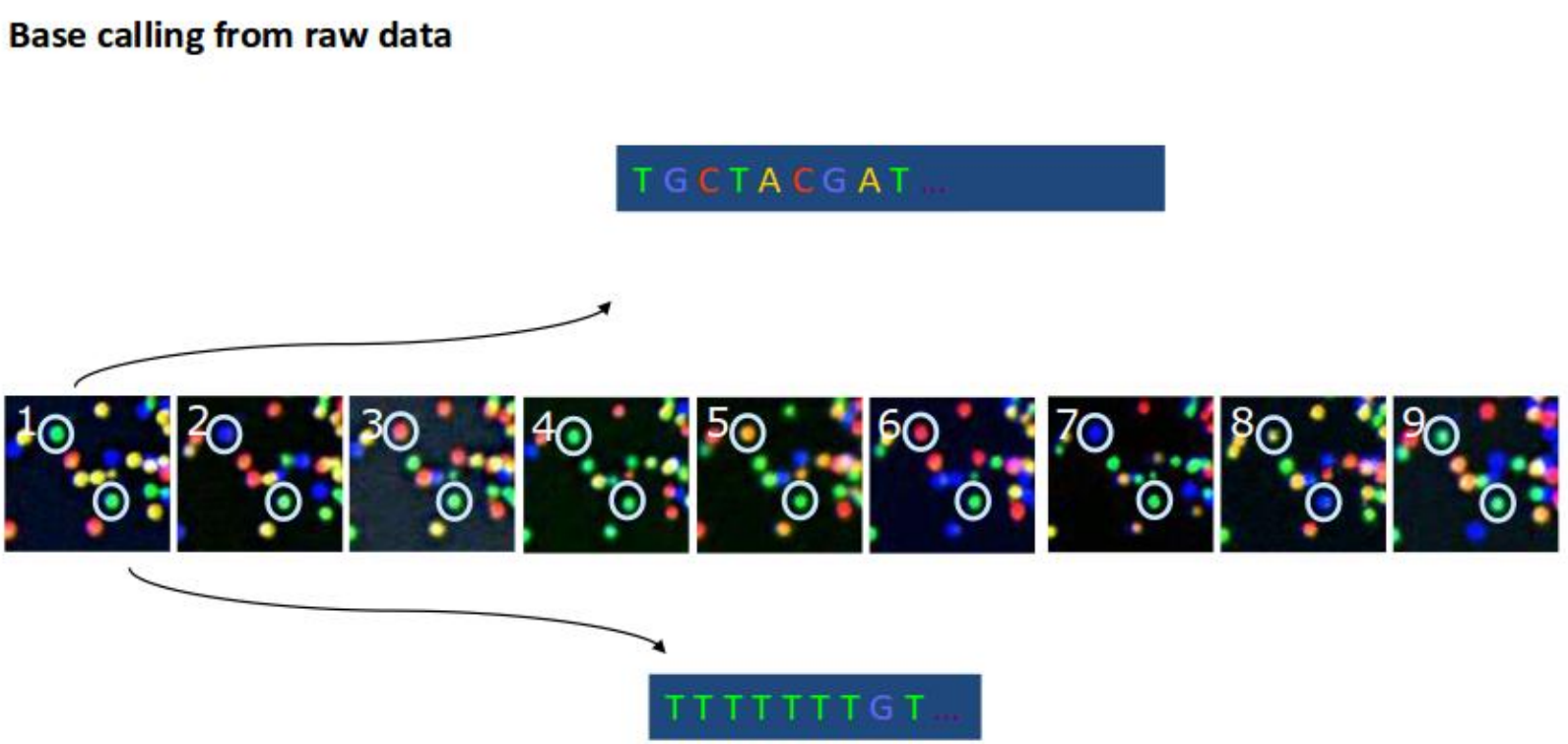
removable dye

A reversible dye terminator used in Sequencing By Synthesis

removable blocking group

Chen, CY. 2014. Frontiers in Microbiol. Vol 5. Article 305.

Illumina Sequencing: Step 3
A picture of the flow cell is taken after the incorporation of each nucleotide



J Fass. UCD Genome Center Bioinformatics Core

What do we get out?

Single-end sequencing

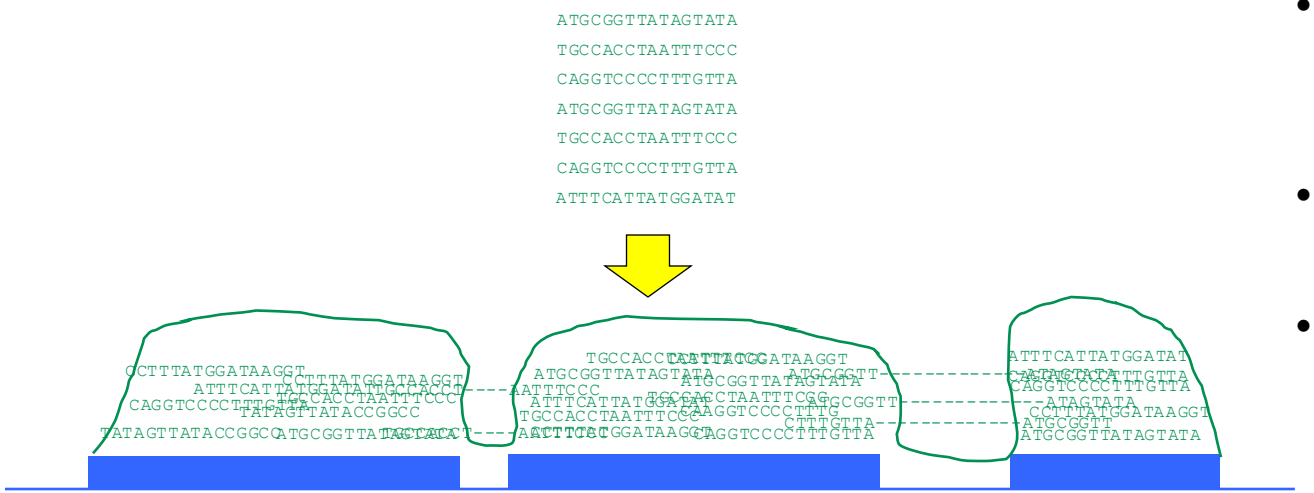
- 150 bp sequenced on one side of the insert
- Depending on the organism & project, 10,000,000 – 200,000,000 inserts sequenced
- 100 MB – 1 GB text file

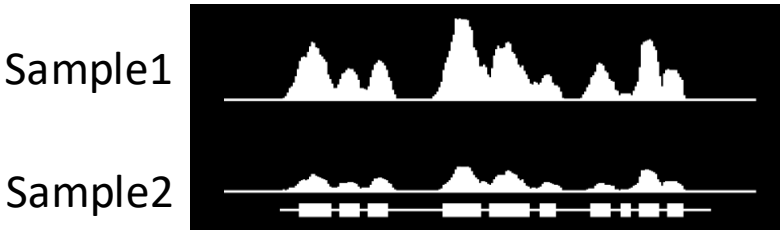
Paired-end sequencing

- 150 bp sequence on each side of the insert
- Depending on the organism & the project, 10,000,000 – 200,000,000 inserts sequenced
- Two 100 MB – 1 GB text files

4. Data analysis

Use computer algorithms to align sequences to the genome

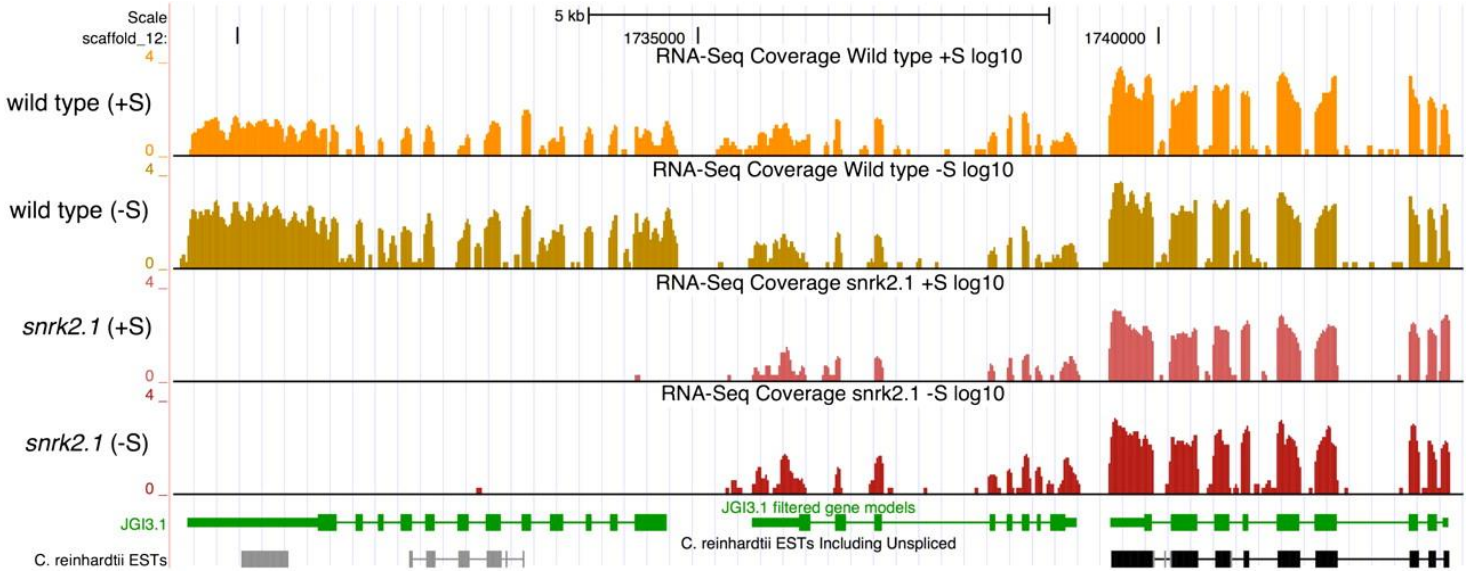




	Sample1	Sample2 …
Gene1	0	0
Gene2	0	0
Gene3	112	230
Gene4	60	1003
Gene5	0	0
Gene6	3	0
Gene7	234	360
…		

- Sequenced fragments (called reads) will align to the exons
- There will be gaps at introns
- The number of reads mapped to a gene is proportional to its expression level (provided that some normalization has been performed)

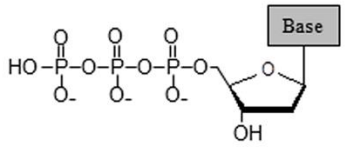
RNA-seq data is visualized on a browser



- Genome is below. Gene models are below (3 genes are shown).
- Samples are rows (two wild-types and two mutants are shown).
- Track heights represent the number of aligned reads for that region.
- Click around on a web-interface for the whole genome

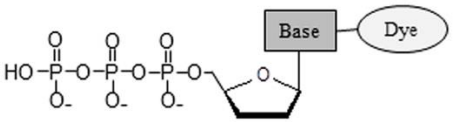
Question

A



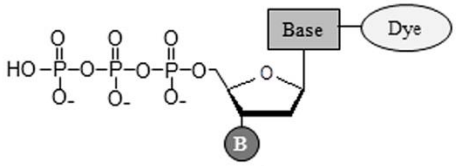
A regular deoxynucleotide (dNTP)

C



A dye-labeled dideoxy nucleotide (ddNTP) used in Sanger sequencing

D



A reversible dye terminator used in Sequencing By Synthesis

- What would happen if the sequencing tech accidentally substituted dNTPs for reversible dye terminators?
- What would happen if Sanger sequencing dye-labeled ddNTPs were used?

Decisions, Decisions

- Amount of input RNA?
- Method of mRNA enrichment?
- Strand specific (directional)?
- Type of library prep kit?
- Type of sequencing
  - Single-end or paired-end?
  - Read length, aka **cycles** (2x150)?
  - How many reads (> 20 million)?
  - What platform (novaSEQ or nextSEQ)?

Videos

- NEB Next Ultra Directional Library Prep
  - https://www.neb.com/tools-and-resources/video-library?device=modal&videoid=%7Bd824c8c5-7942-437c-9086-e93ff3c94a12%7D
- Illumina Sequencing By Synthesis
  - https://www.youtube.com/watch?v=fCd6B5HRaZ8
- If you're interested - 10x Genomics single-cell RNA-seq:
  - https://www.youtube.com/watch?v=6UVOdCc1Q7I