

# An overview of genomics and sequencing terminology and practices

Mark Stenglein, GDW



Math  
undergrad

7 years as a  
software engineer

PhD in  
virology /  
mol. biol.

Postdoc using  
microarrays, NGS,  
and bioinformatics

Assoc.  
Professor at  
CSU

1999, Bangkok, Thai Airways test facility



CENTER FOR VECTOR-BORNE  
INFECTIOUS DISEASES



University of California  
San Francisco  
*advancing health worldwide*

**Mark Stenglein, PhD**

Associate Professor

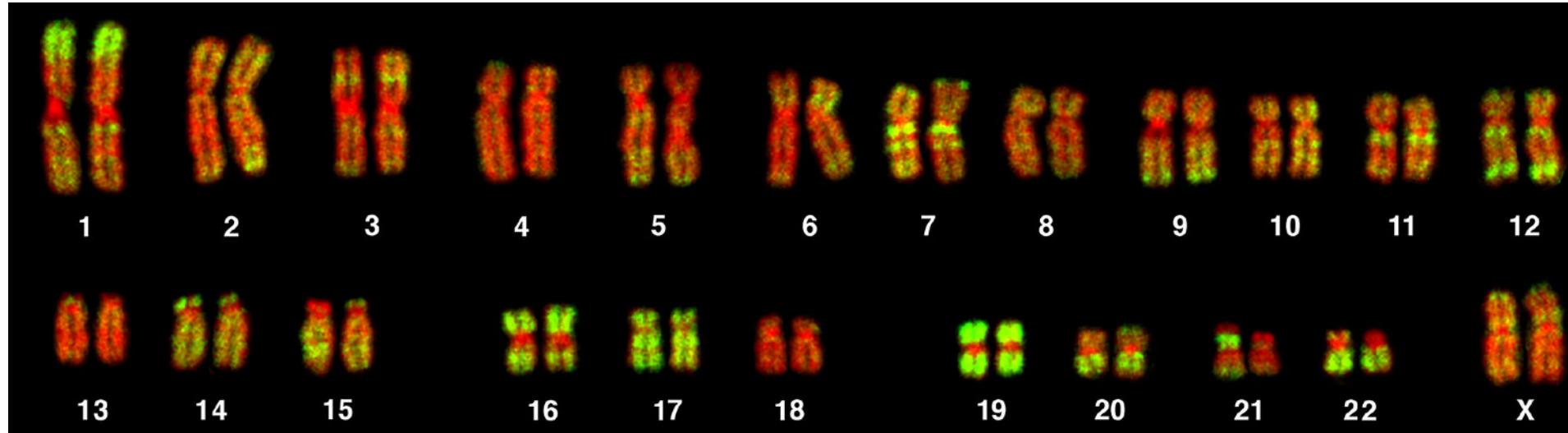
Department of Microbiology, Immunology, and Pathology  
College of Veterinary Medicine and Biomedical Sciences  
Colorado State University

**Mark.Stenglein@colostate.edu**

StengleinLab.org

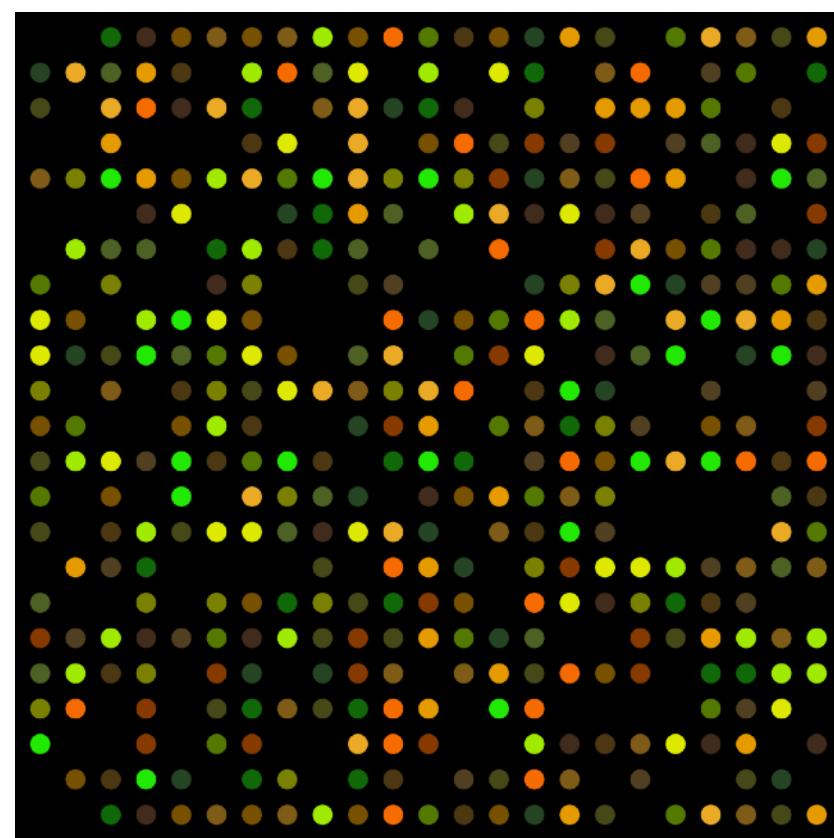
# Non-sequencing genomic techniques

FISH



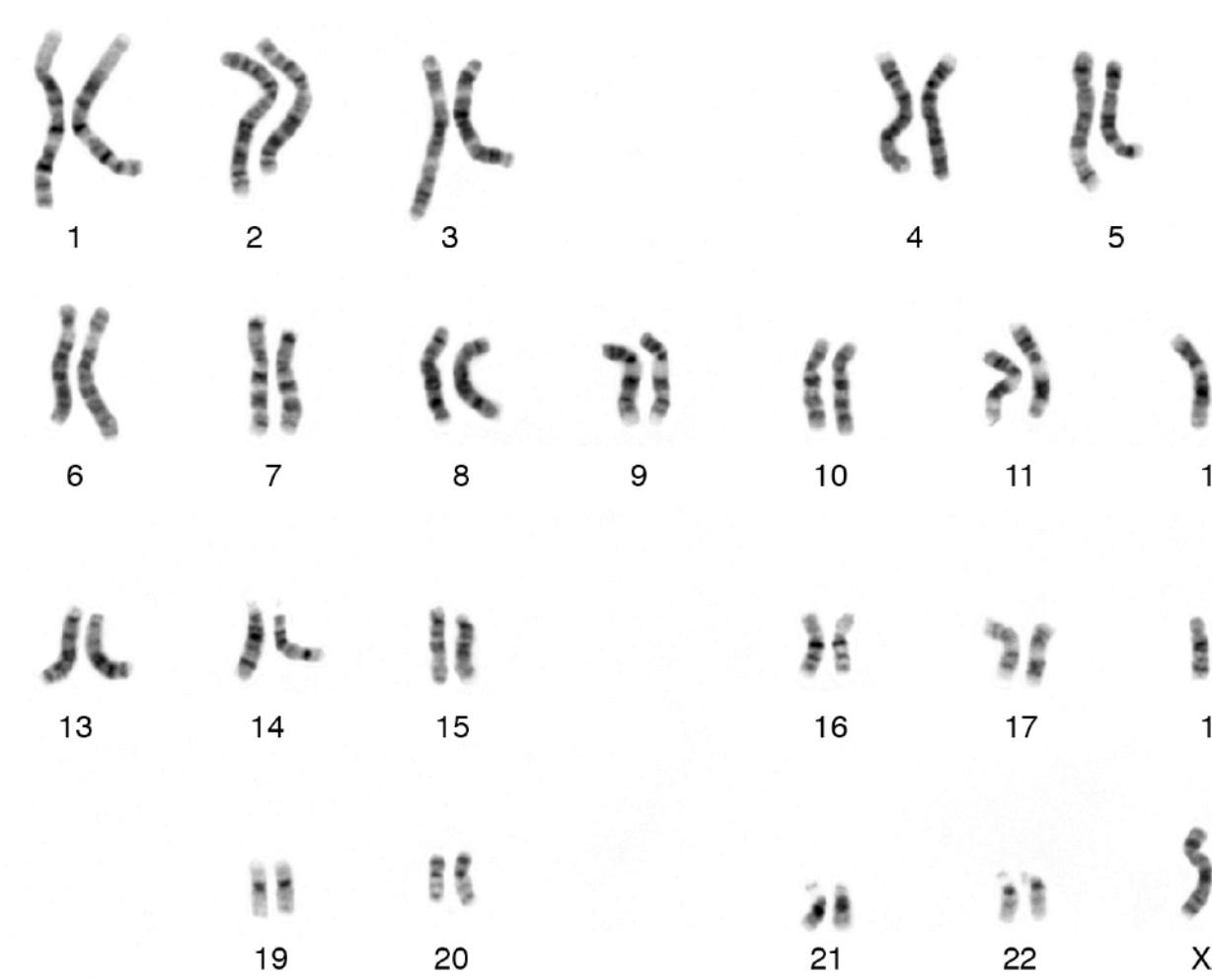
Bolzer et al (2005) PLoS Biol

Microarray



Wikimedia commons

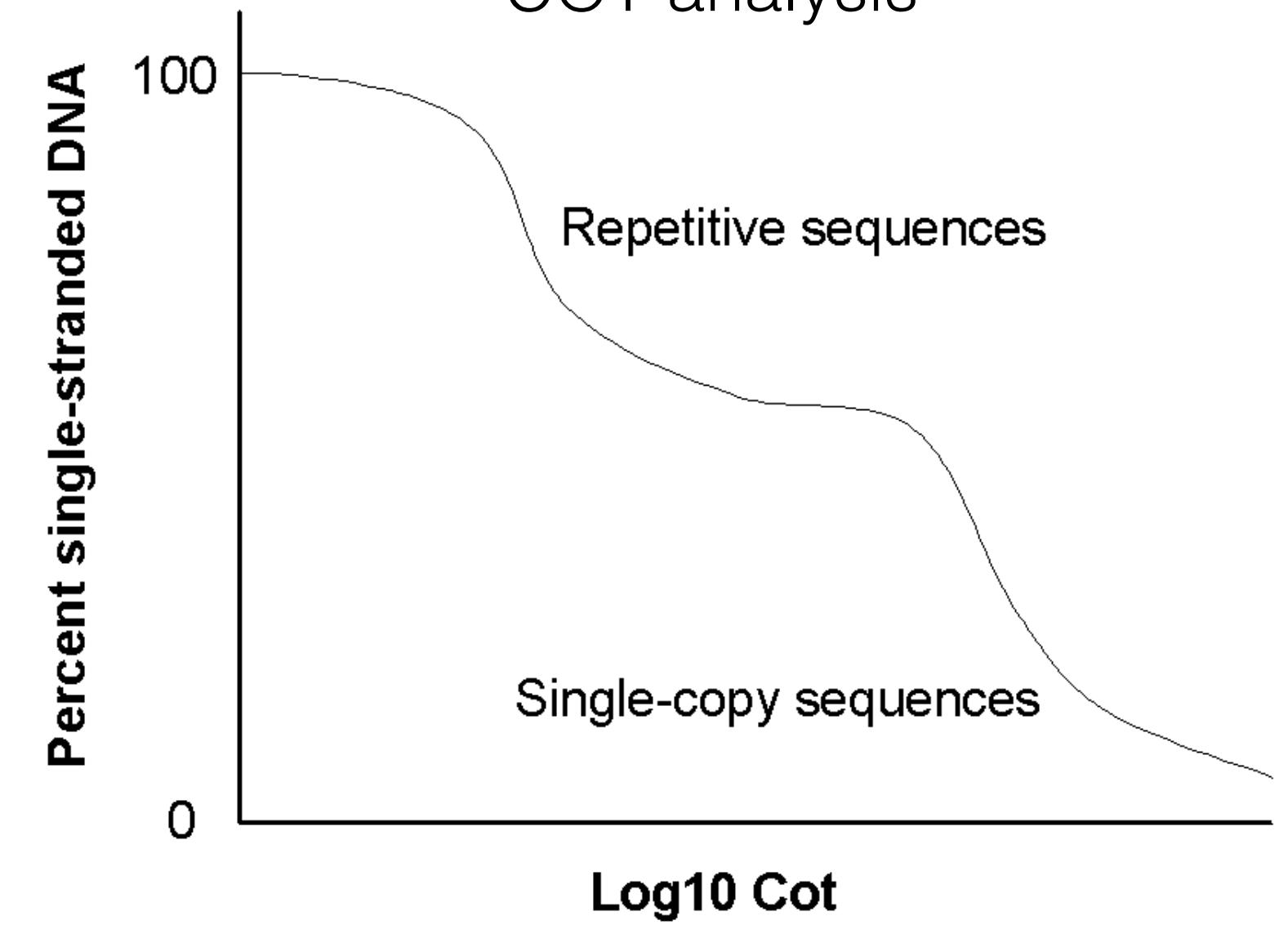
Karyotype



Polytene chromosomes

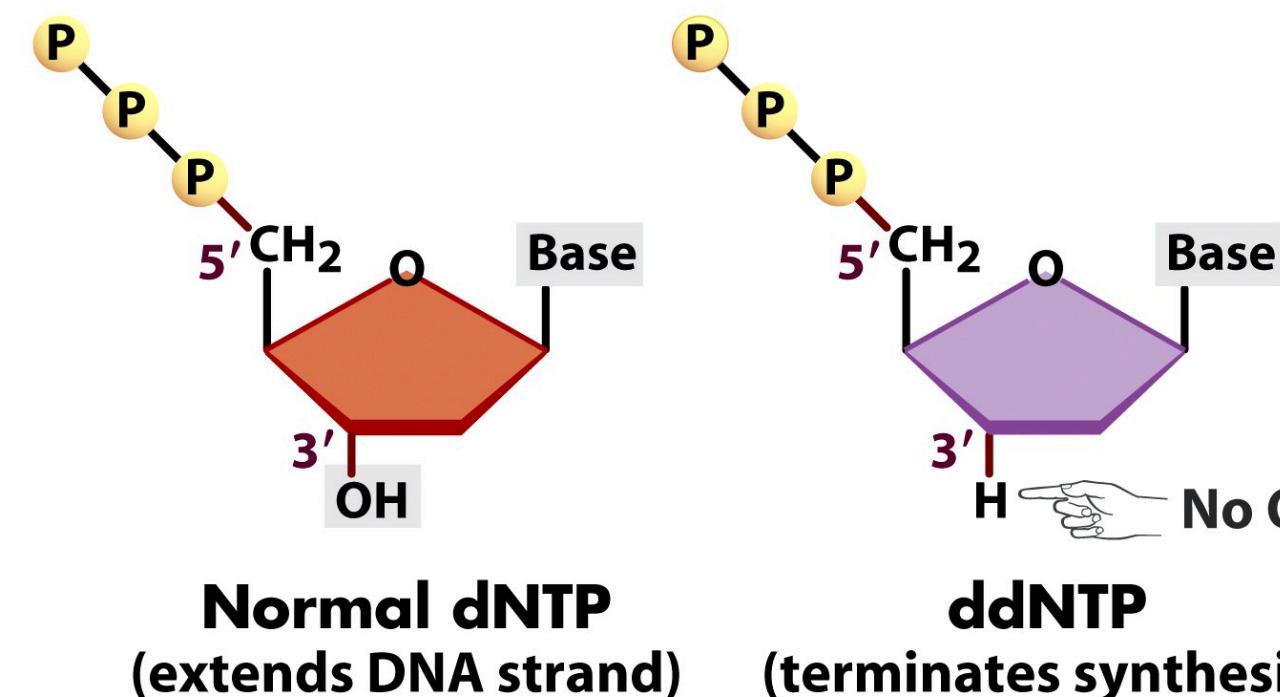


COT analysis



# Sanger Sequencing (1977): sequencing 1 target at a time

**ddNTPs terminate DNA synthesis.**

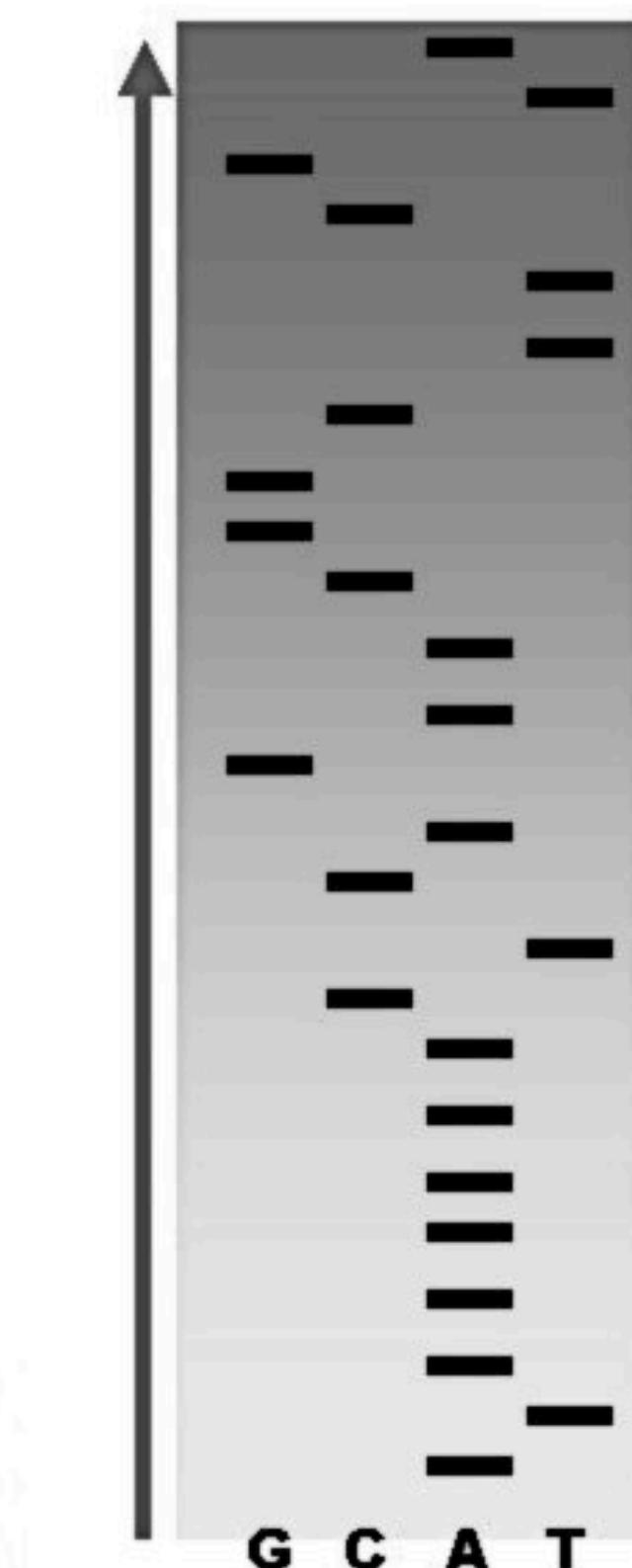
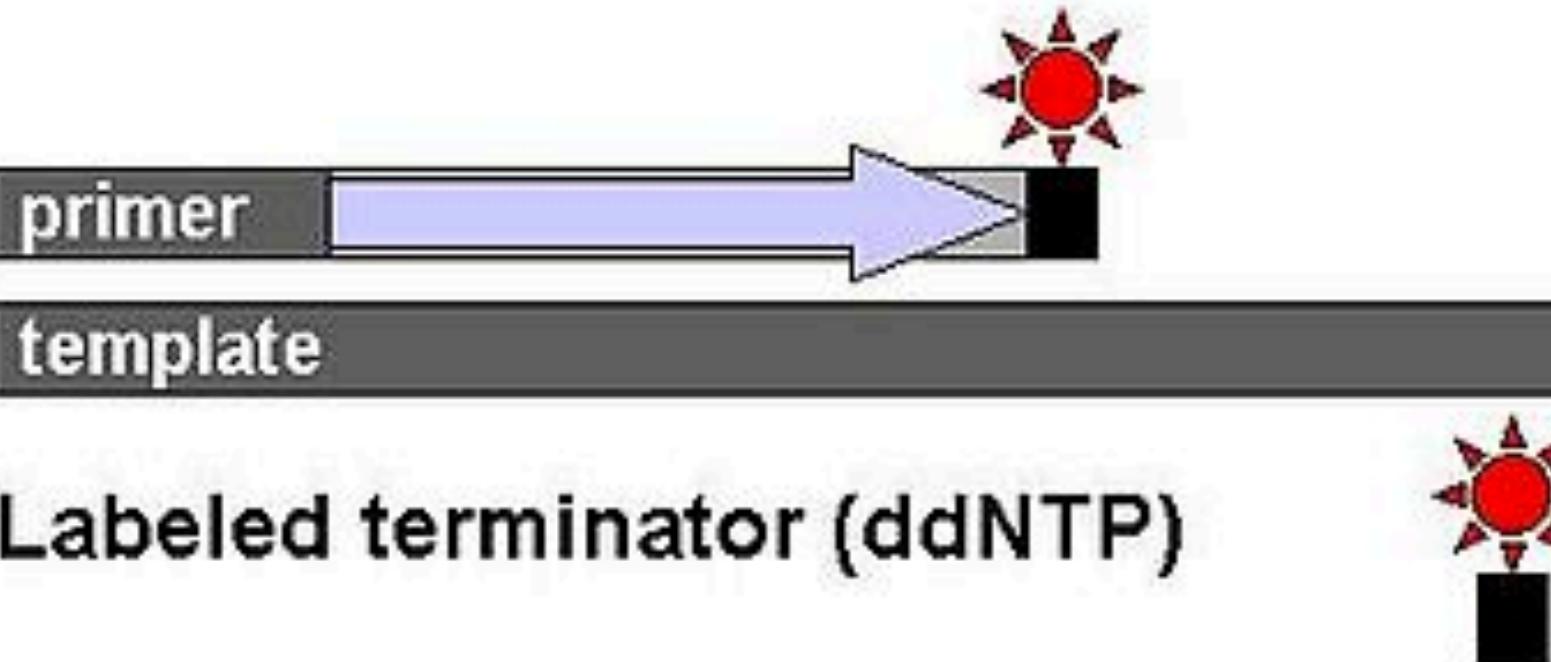


**Normal dNTP**  
(extends DNA strand)

**ddNTP**  
(terminates synthesis)

Figure 19-6a Biological Science, 2/e

© 2005 Pearson Prentice Hall, Inc.



Slide courtesy Dan Sloan. Image credits: Sanger et al (1977) and Wikipedia

Next generation sequencing (NGS) ~ deep sequencing ~ high throughput sequencing (HTS)

All simultaneously sequence **many molecules in parallel**

### Short read sequencing (Illumina)

- Millions of reads
- Relatively short: ~50-300 nt (Illumina)
- Relative low error rates
- Cheaper per base pair of data generated



MiSeq

\$100,000-\$1,000,000

### Long read sequencing

- Fewer, longer reads
- >1 kb (PacBio), up to 100s of kb (Oxford Nanopore)
- Relative high error rates

Oxford Nanopore MinION

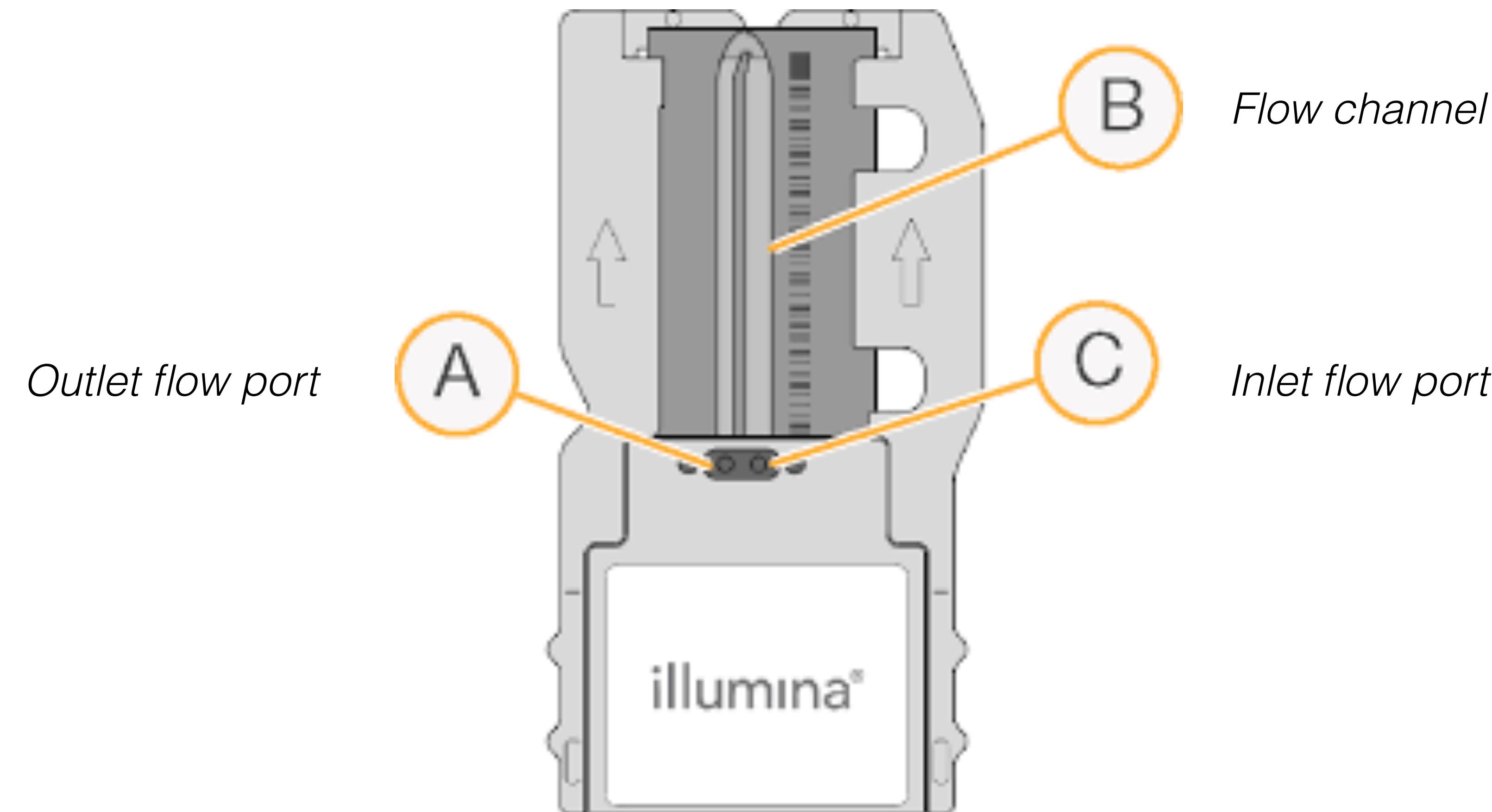


\$1000

PacBio RS-II

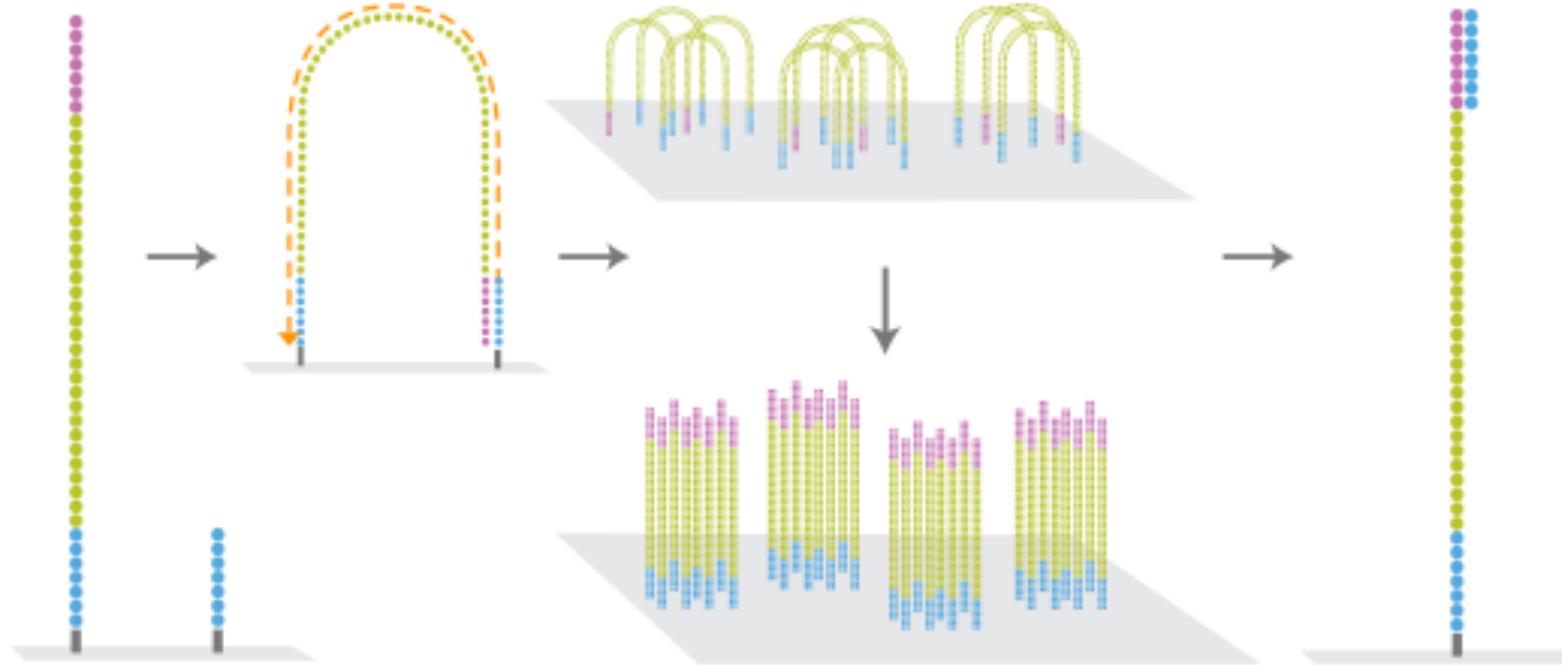


Illumina sequencing happens on flow cells



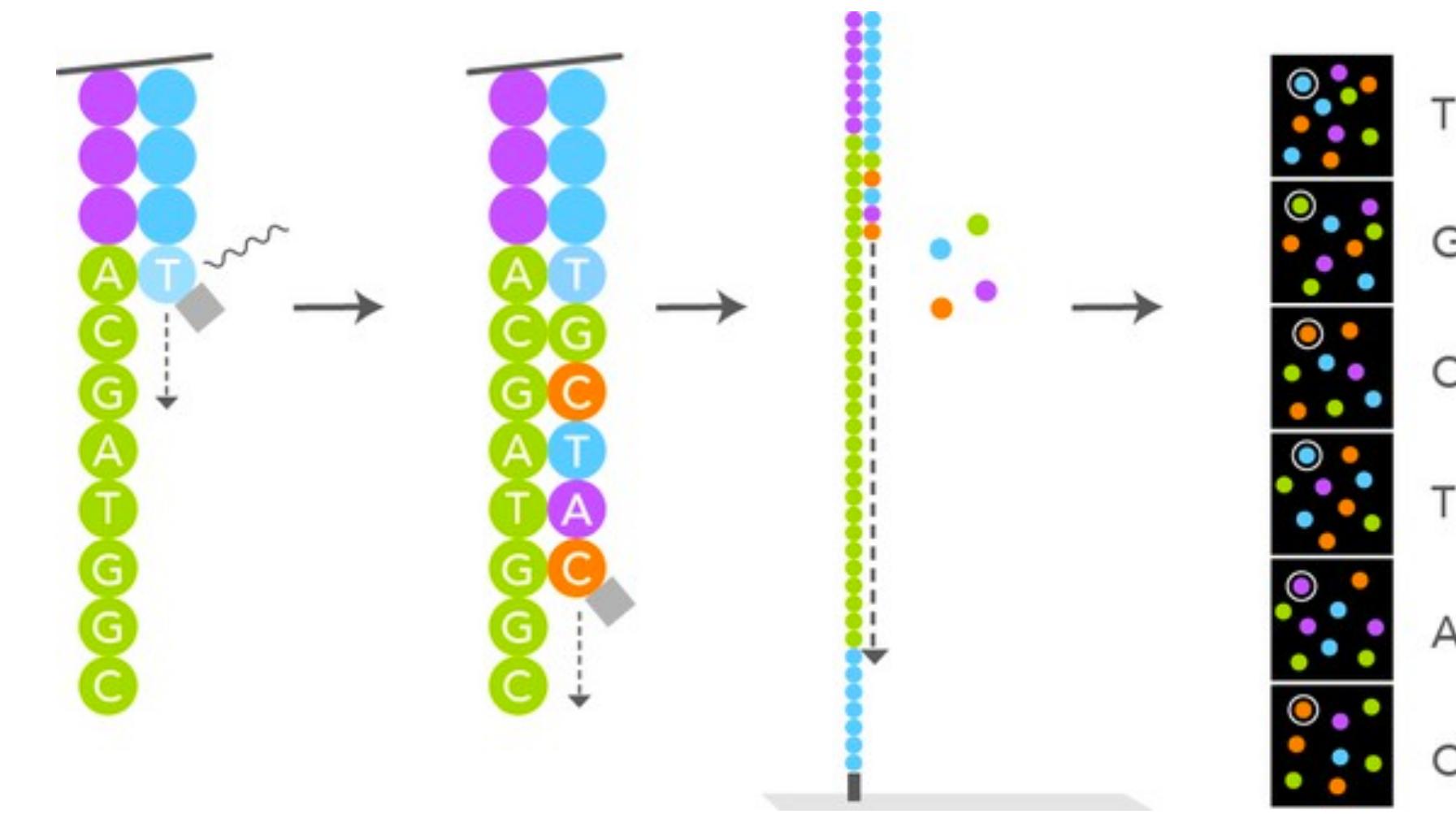
*Image credit:*  
Illumina

# Illumina instruments use sequencing by synthesis (SBS)



Millions of clusters per flow cell

Each cluster contains 1000s of clonal copies of a library molecule

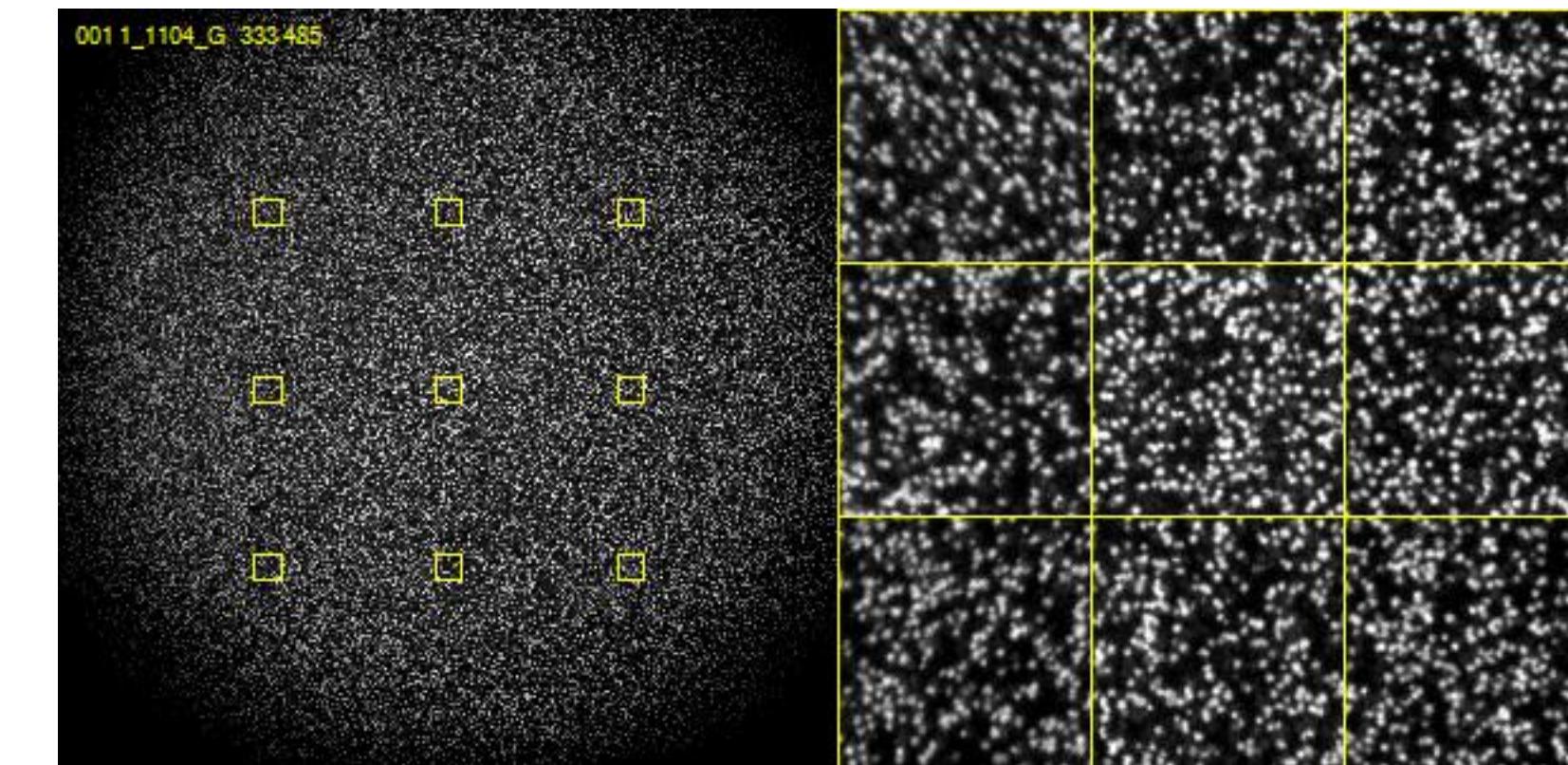


Library molecules are sequenced by primer extension reactions that incorporate chain-terminated, fluorescent nucleotides

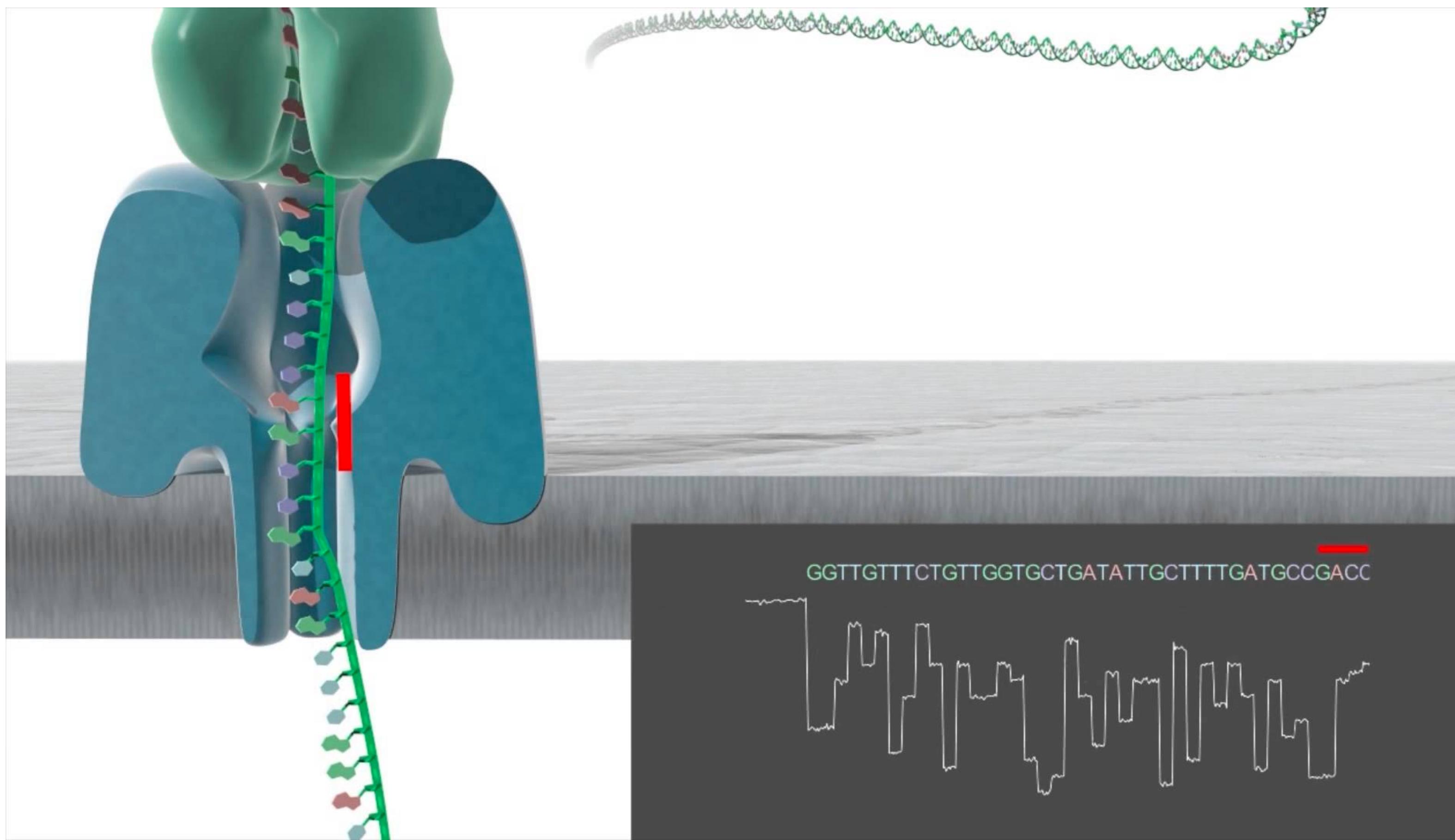
This technology is very similar to Sanger sequencing in principle

real raw Illumina sequencing data

Image credit: Illumina

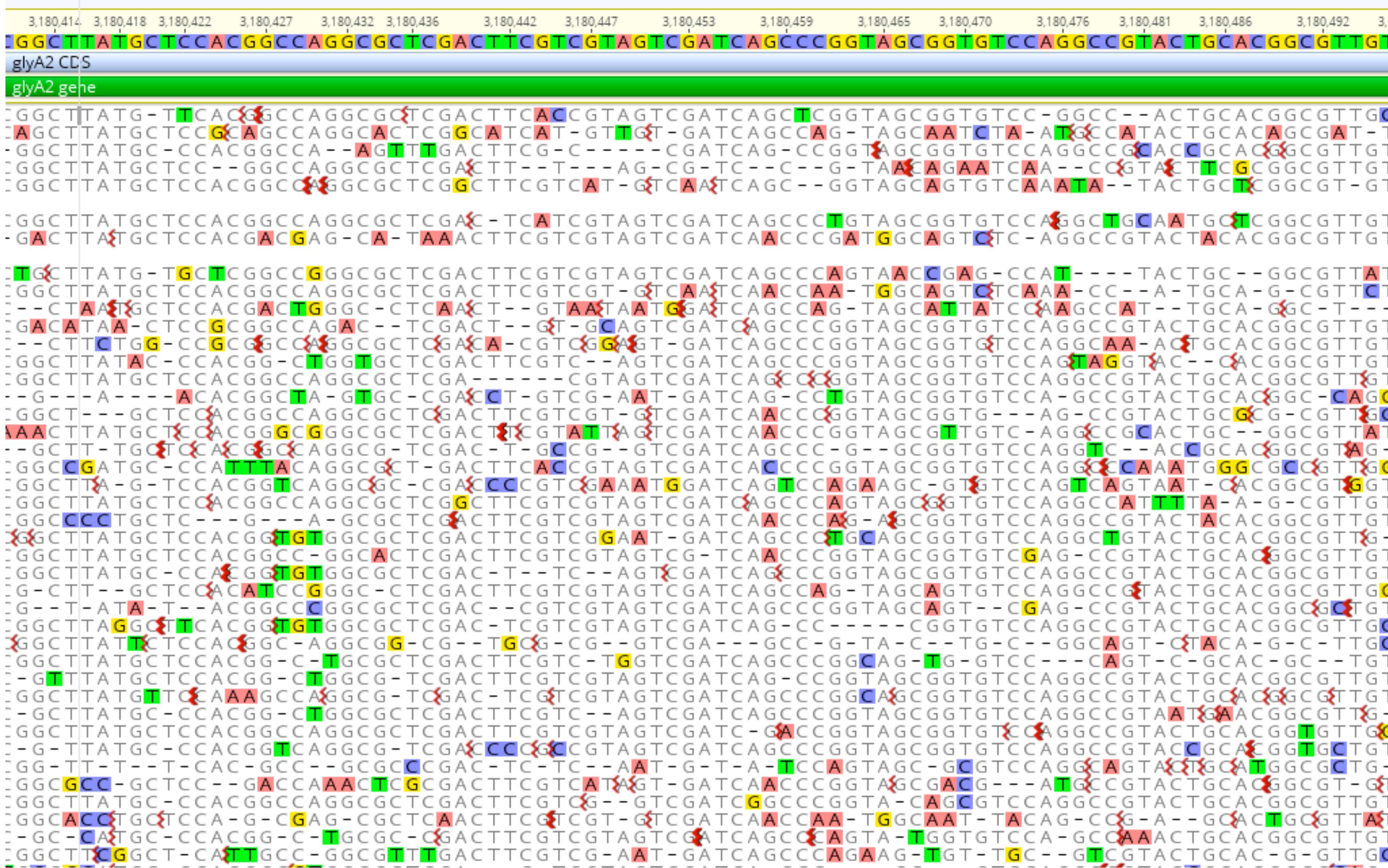


# Long read sequencers sequence single molecules

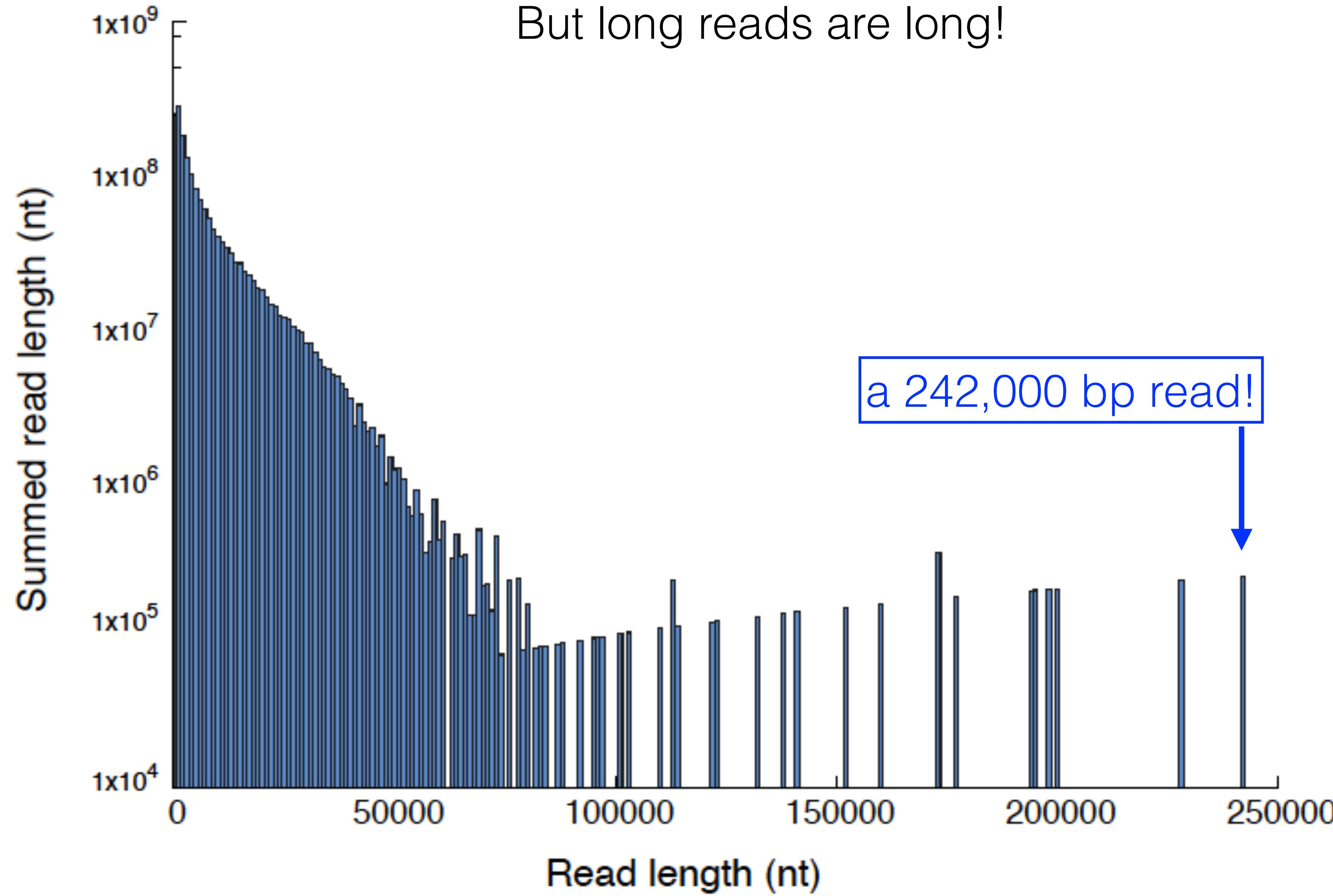


Nanopore sequencing: Much longer reads, but with much higher error rates

# Long reads have relatively high error rates (up to 5-10%)



Improvements to nanopore chemistry and software is pushing down error rate



Justin Lee

# Some nanopore sequencers are portable



Oxford Nanopore @nanopore

Our new field kits can be stored at ambient temperature for up to a month, the workflow is simple with no compromise on throughput. Sequencing anywhere made easier: [bit.ly/2v1ilCG](https://bit.ly/2v1ilCG) #ECCMID2019



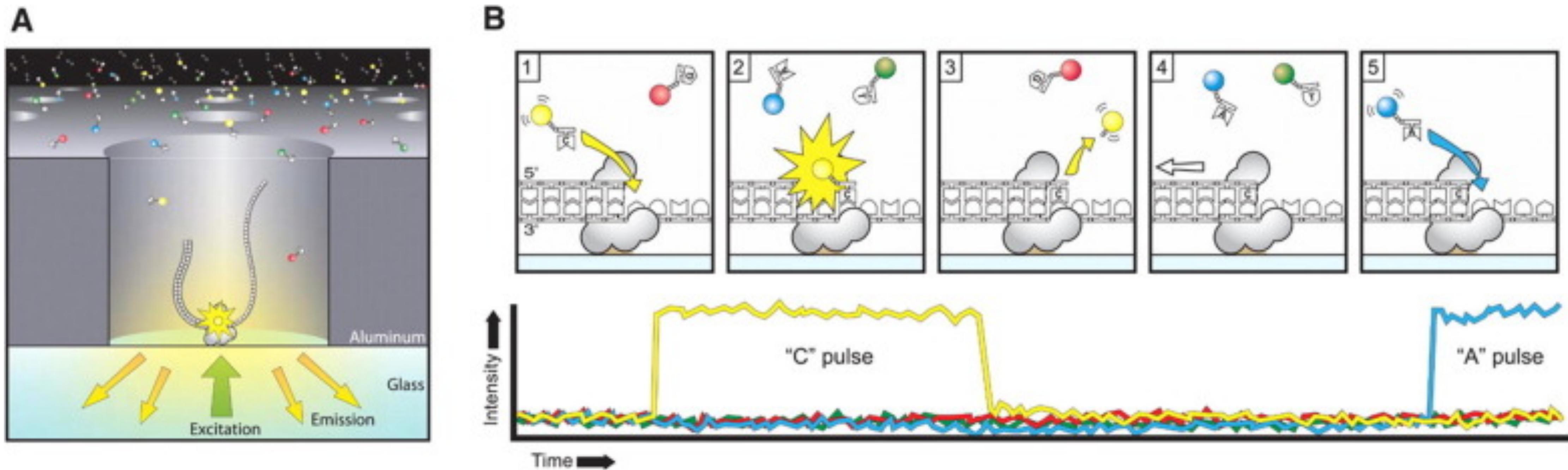
2:08 AM · Apr 12, 2019 · HubSpot



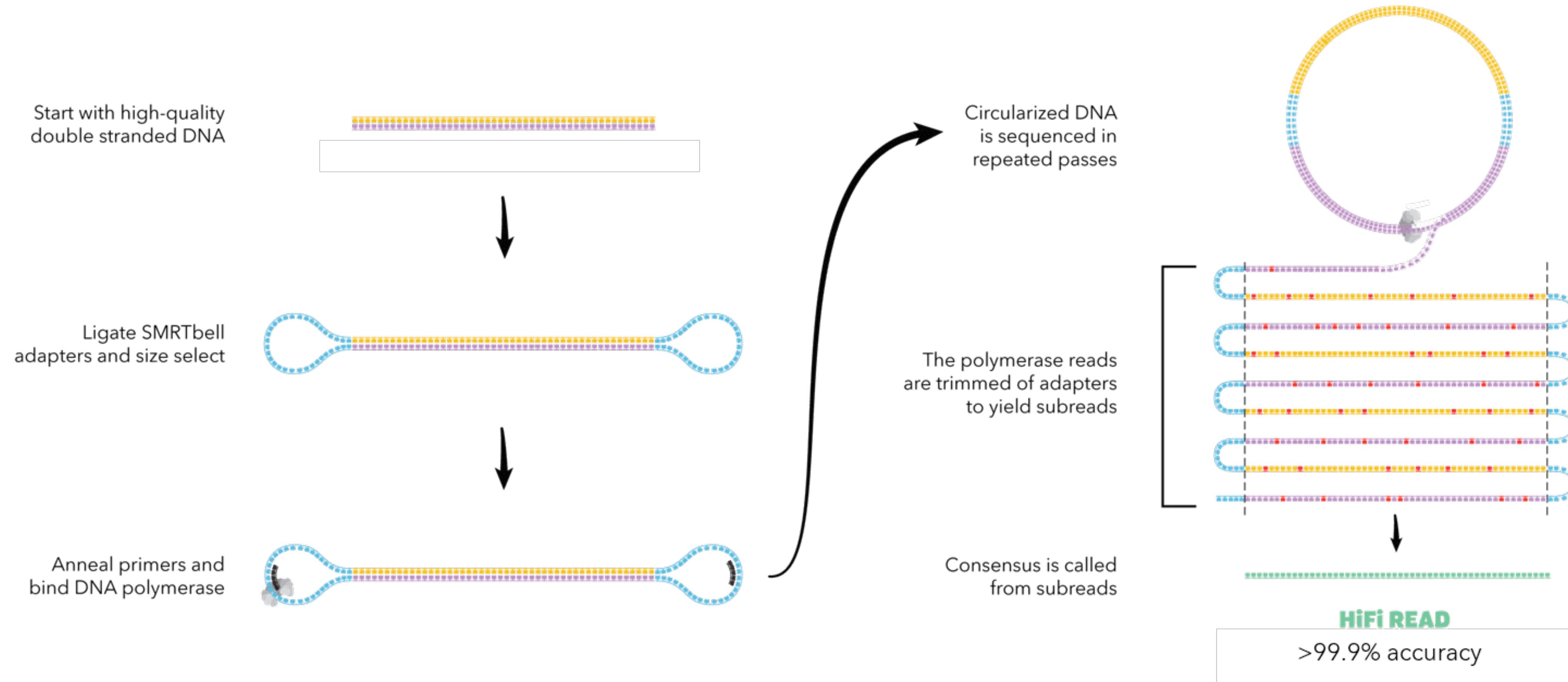
Illumina reads have much lower error rates (~0.1% – 1%)



# PacBio single molecule real-time (SMRT) sequencing is the other main long-read technology



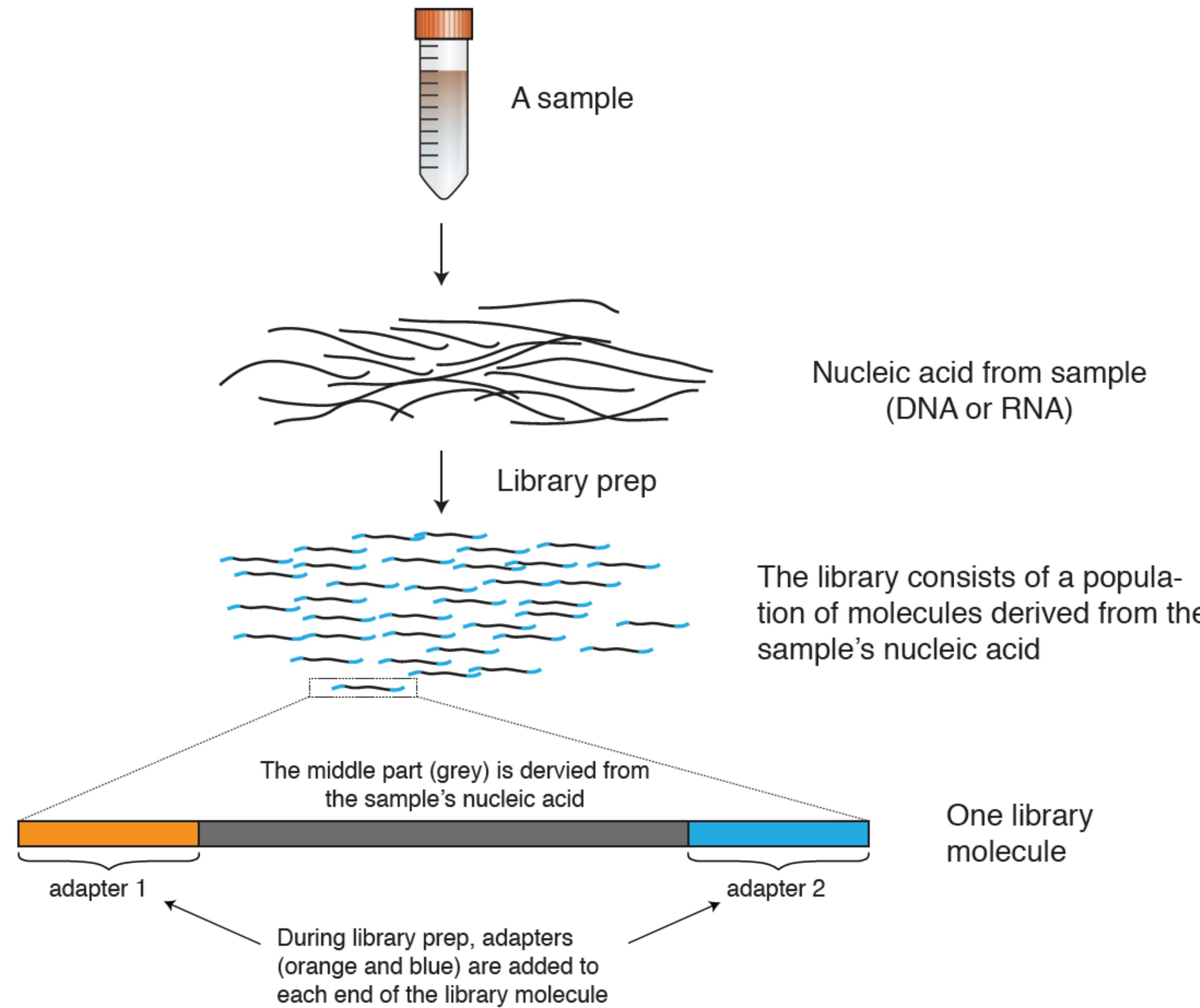
# PacBio HiFi sequencing sequences the same molecule many times to reduce error rate



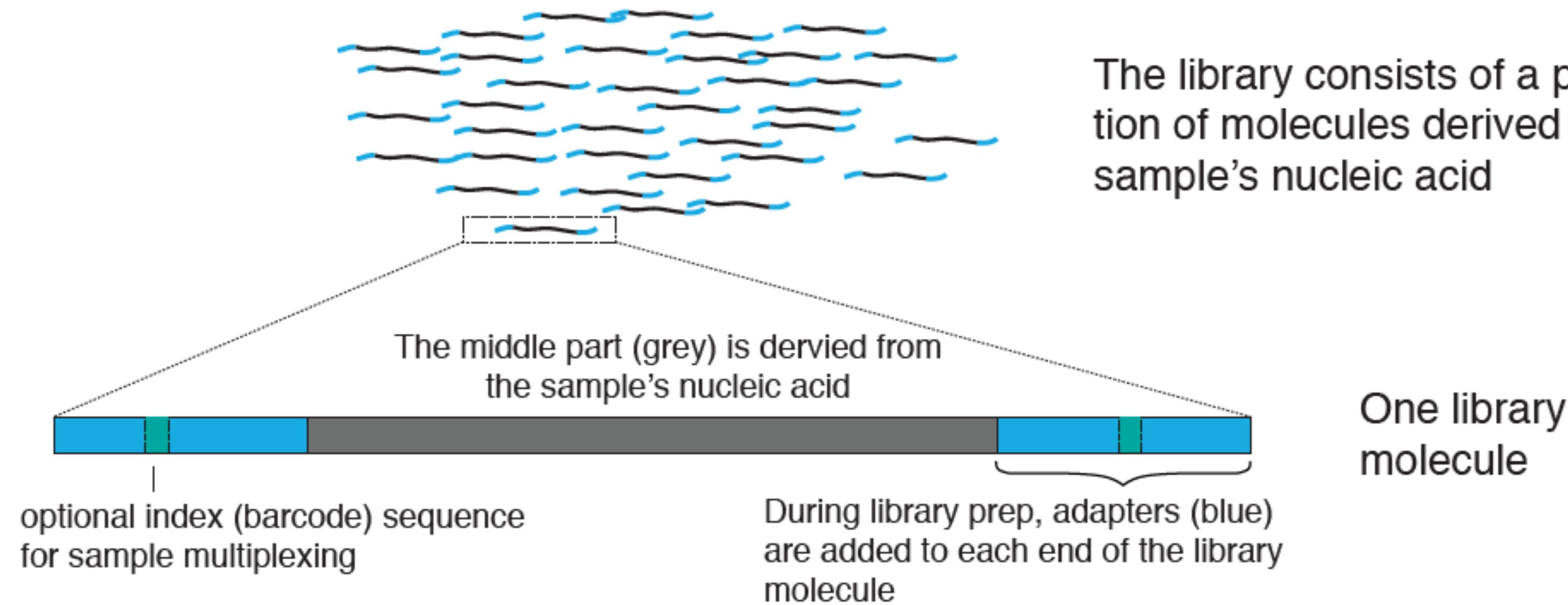
Best of both worlds: long reads with low error rate

Image: PacBio

# Library prep converts nucleic acids into a form suitable to be sequenced

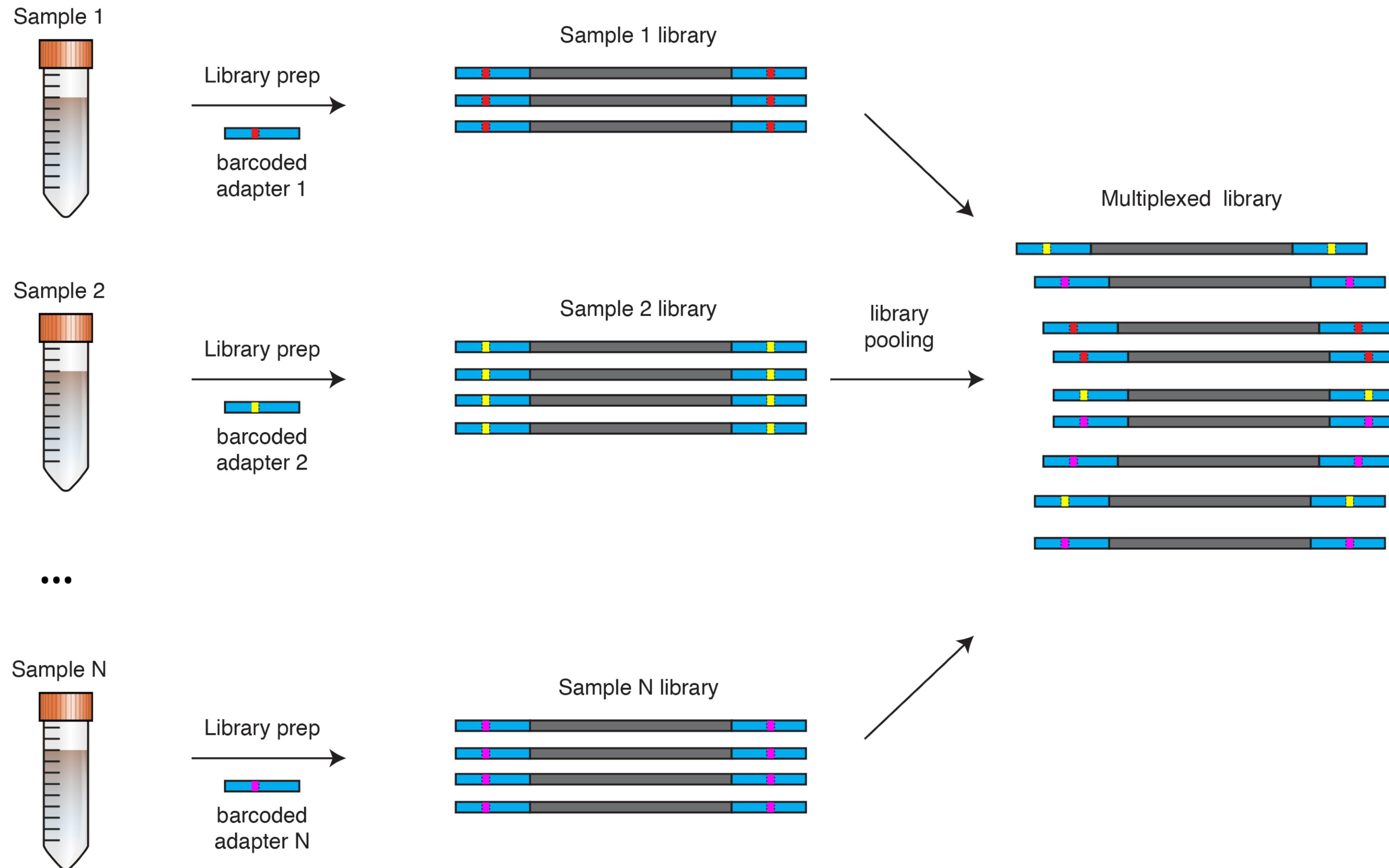


# Library prep converts nucleic acids into a form suitable to be sequenced



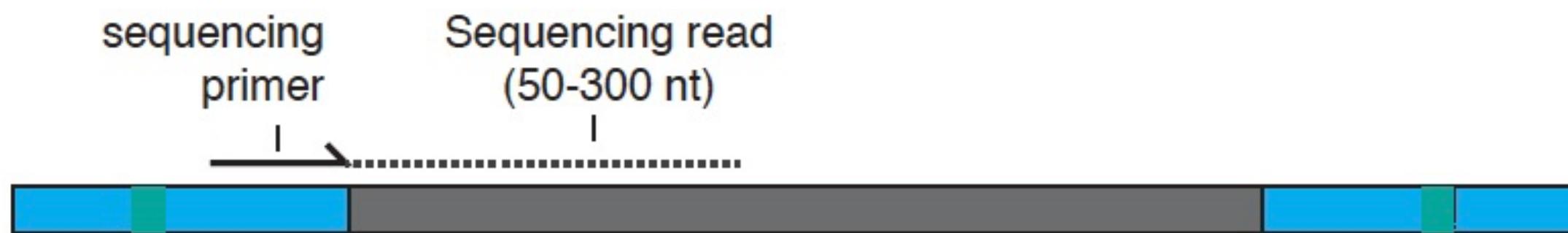
An example Illumina library molecule

# Barcodes (or indexes) allow sample multiplexing



# Illumina sequencing produces 1-4 reads per library molecule

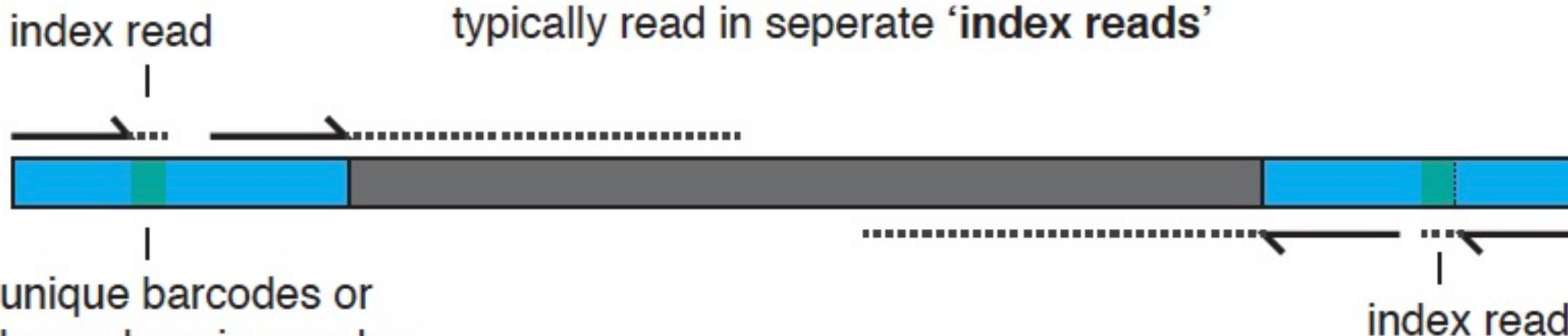
In **single end sequencing**, a library molecule is sequenced from one end



In **paired end sequencing**, a library molecule is sequenced from both end



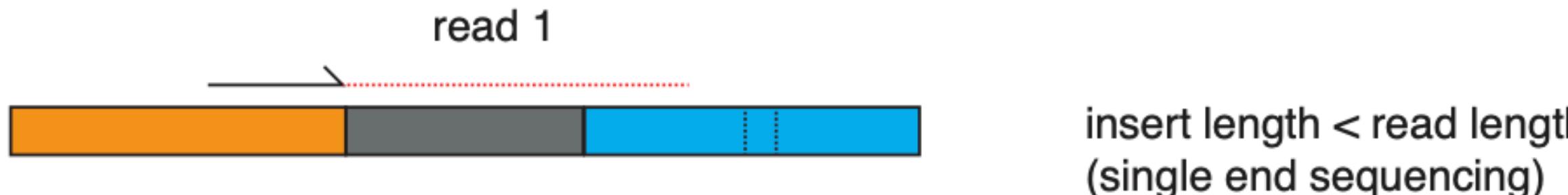
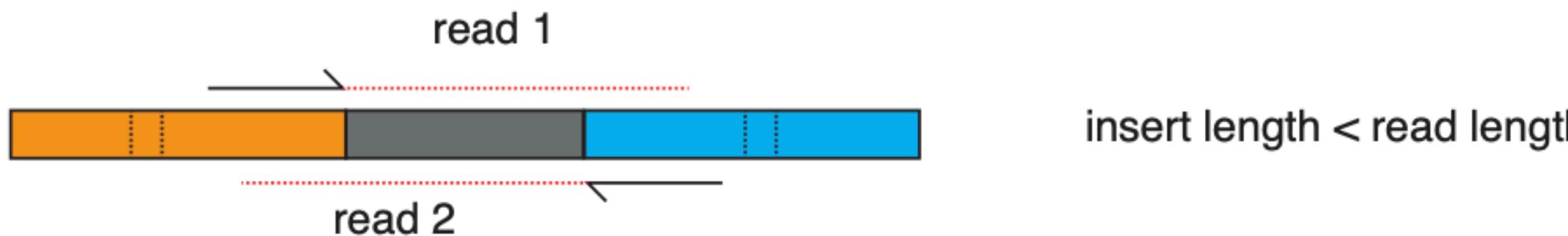
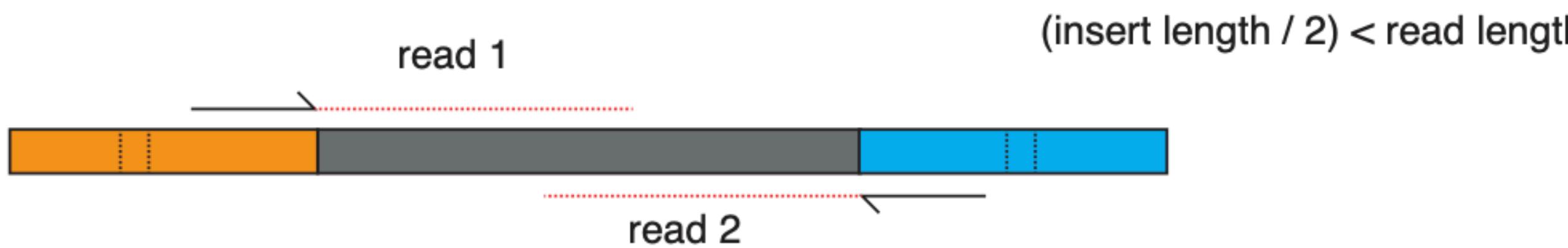
The library molecule's barcodes (indexes) are typically read in separate 'index reads'



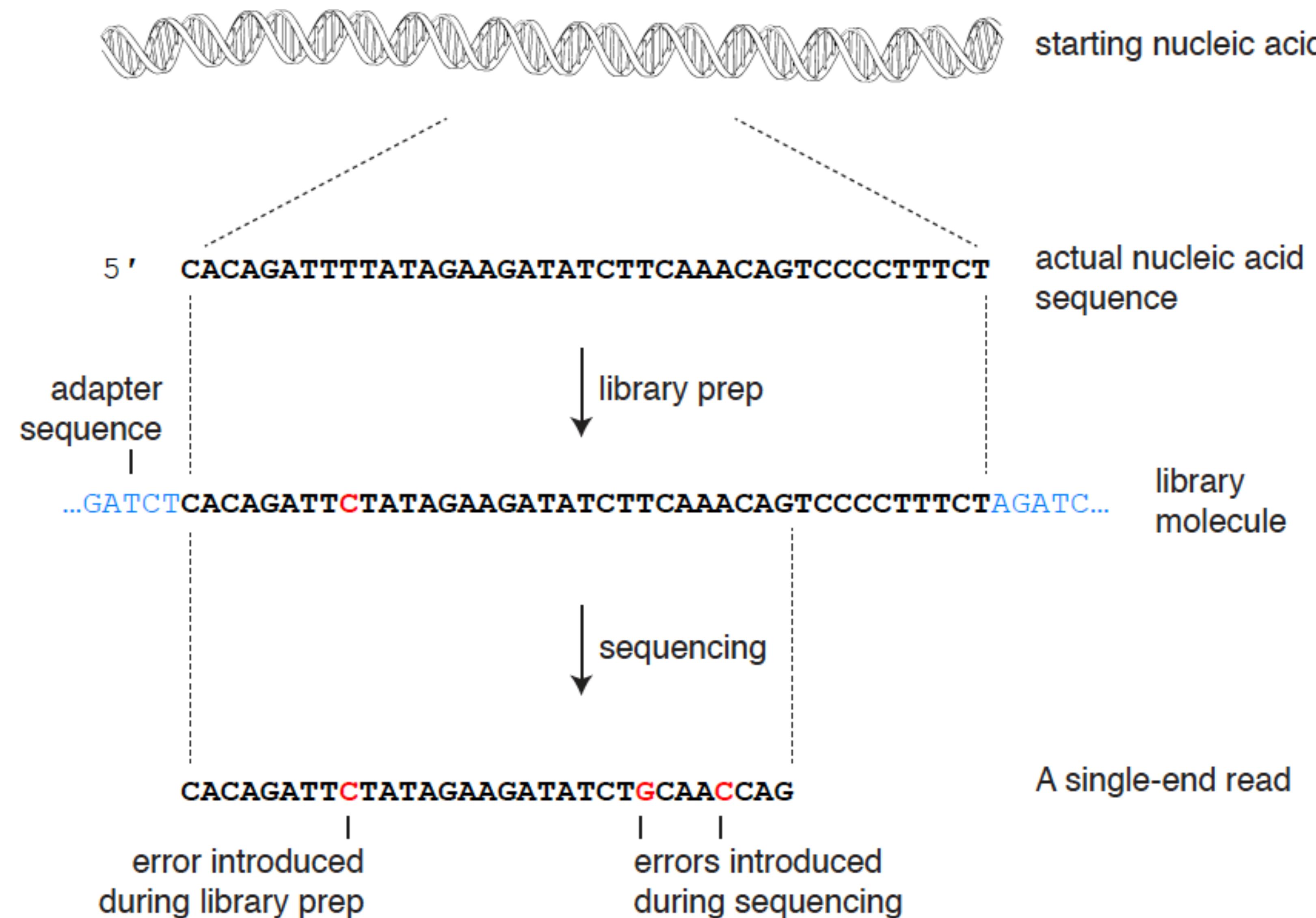
unique barcodes or barcode pairs can be used to differentiate multiplexed samples

index read

Whether or not paired reads overlap and whether or not a read extends into the opposite adapter is a function of insert size and read length



Reads are measurements of the real sequence that often contain errors and are usually shorter than the starting nucleic acid

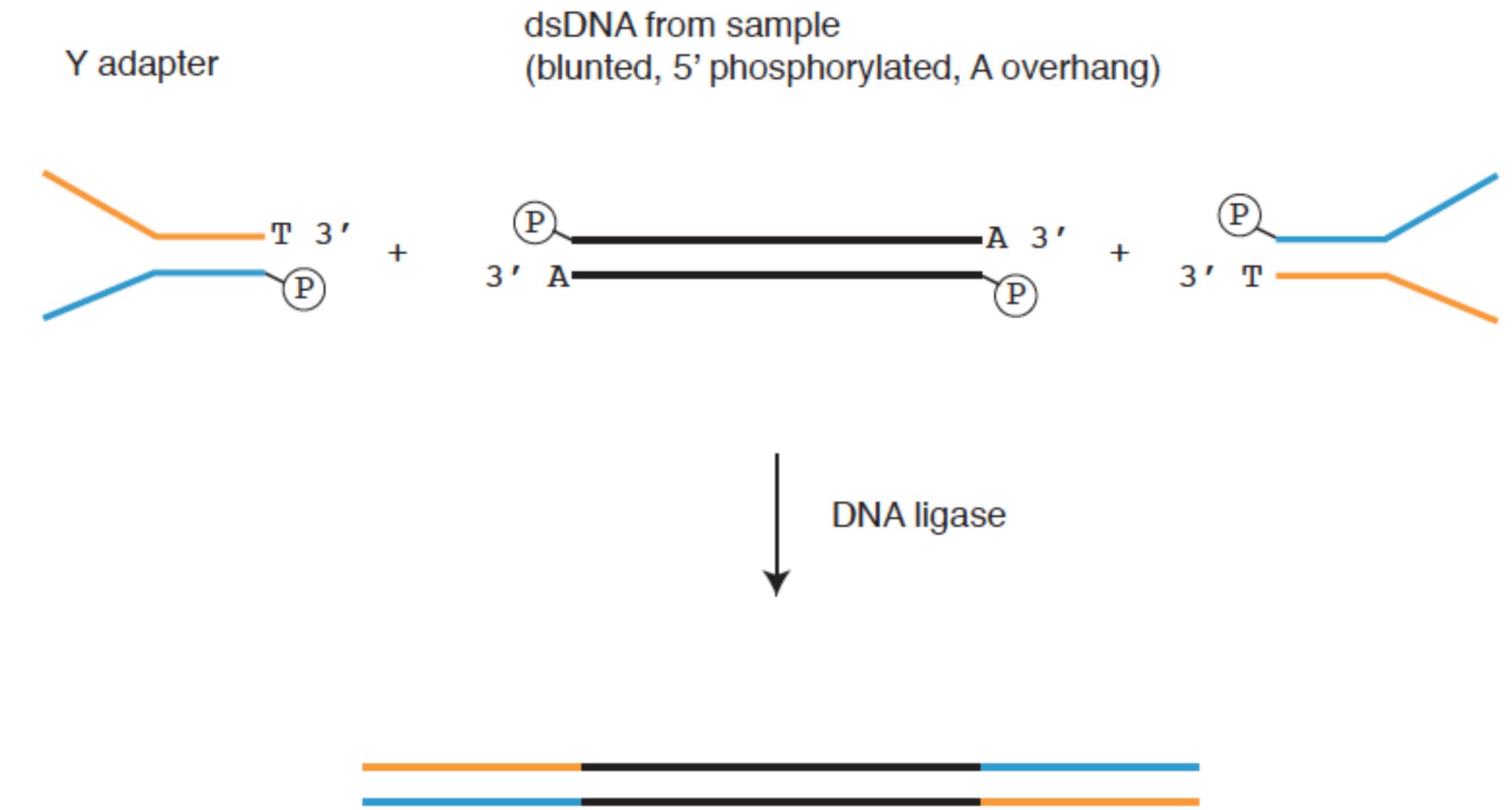


# There are many good ways to make sequencing libraries

## Common Illumina library prep steps (not always included and not always in this order)

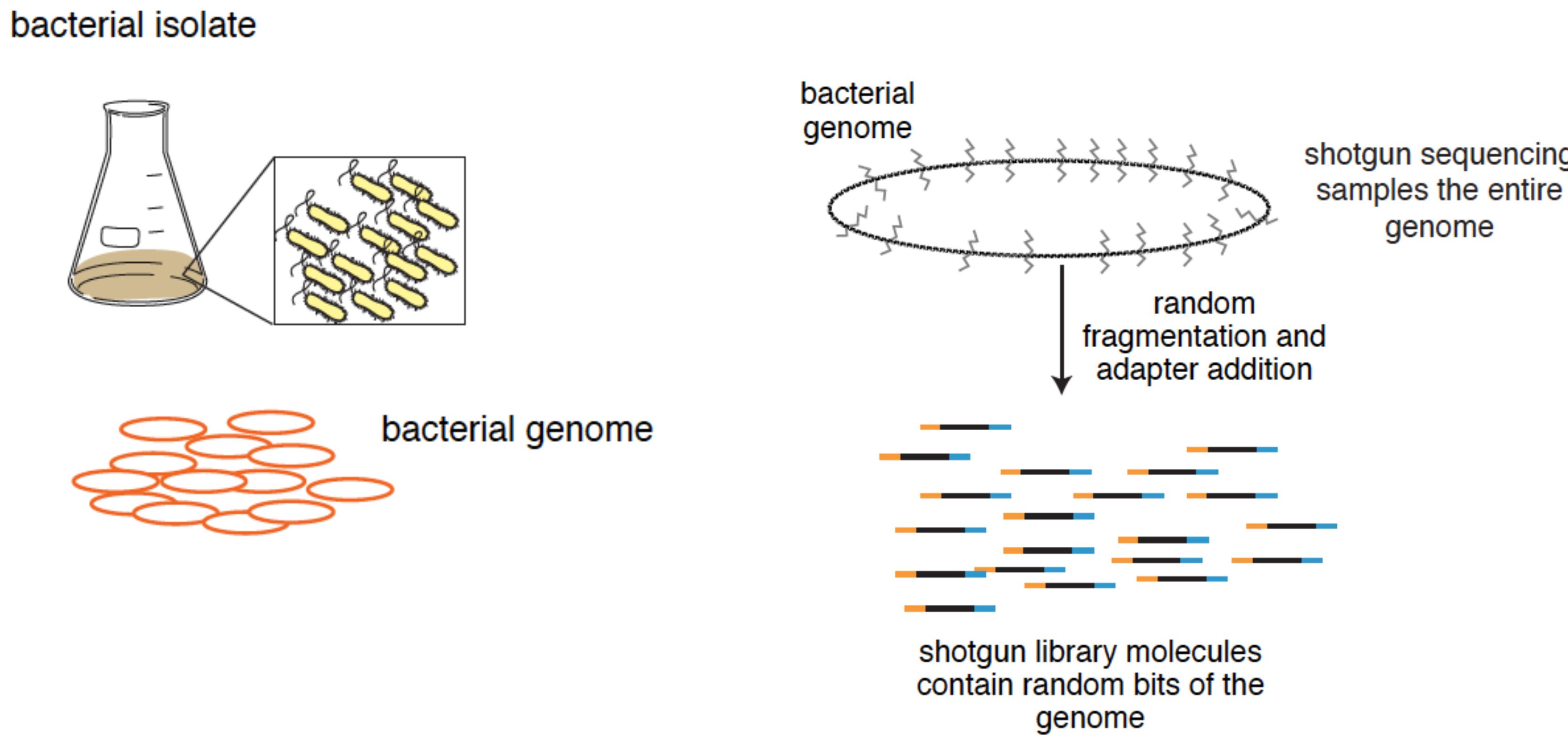
- Nucleic acid isolation
- Enrichment (of nucleic acid subtypes you want) or subtraction (of those you don't want)
- Nucleic acid fragmentation
- Conversion of RNA into dsDNA (for RNA sequencing)
- Addition of adapters to ends of library molecules, possibly with barcodes for multiplexing
- Library amplification
- Pooling of multiplexed samples
- Library QC / quantification

Adapters can be added to sample-derived dsDNA by ligation



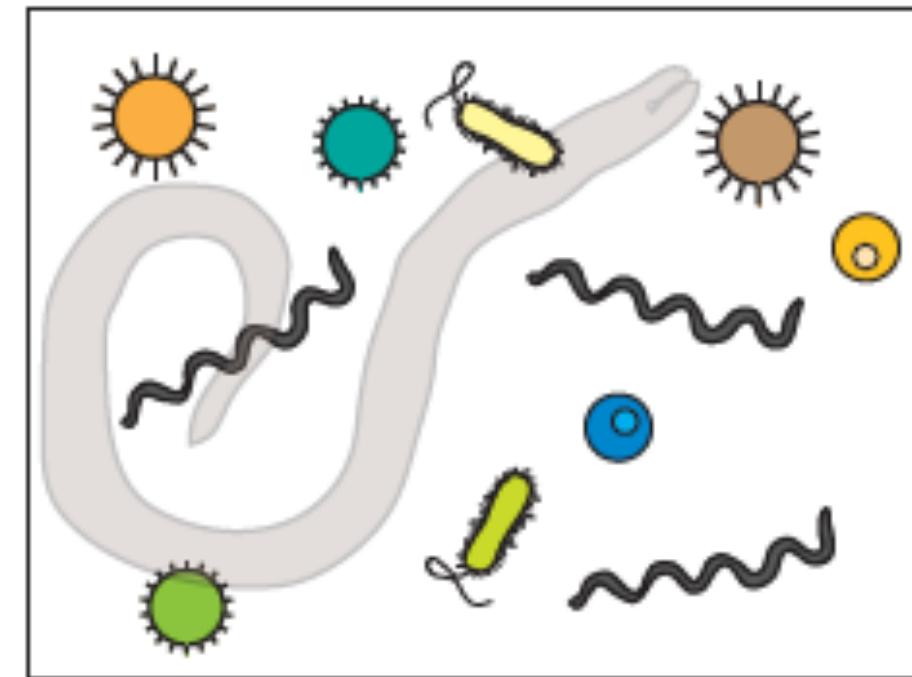
# How you make a library determines what type of sequencing you're doing

For instance, if you make a 'shotgun library' from a single organism, you're doing whole genome sequencing (WGS)



# Metagenomic sequencing involves sequencing of genomes from more than one organism

soil community



Could make a 16S or a shotgun library from these genomes

Sequencing of RNAs from a complex sample like this is metatranscriptomics

soil 'metagenome'



“Nothing in Biology Makes Sense Except in the Light of Evolution”  
- Theodosius Dobzhansky

Biological sequences really only make sense  
when you compare them to each other

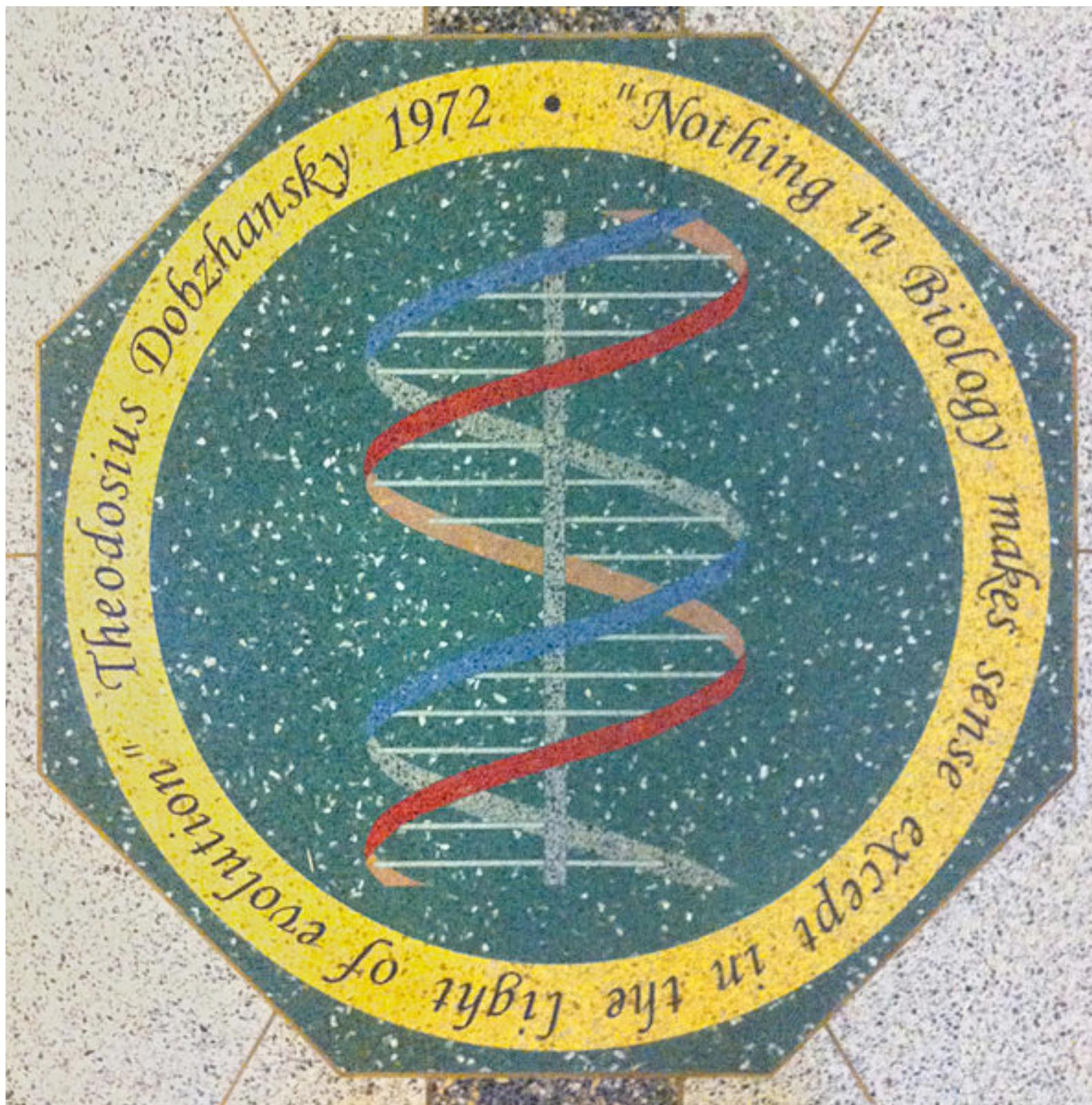
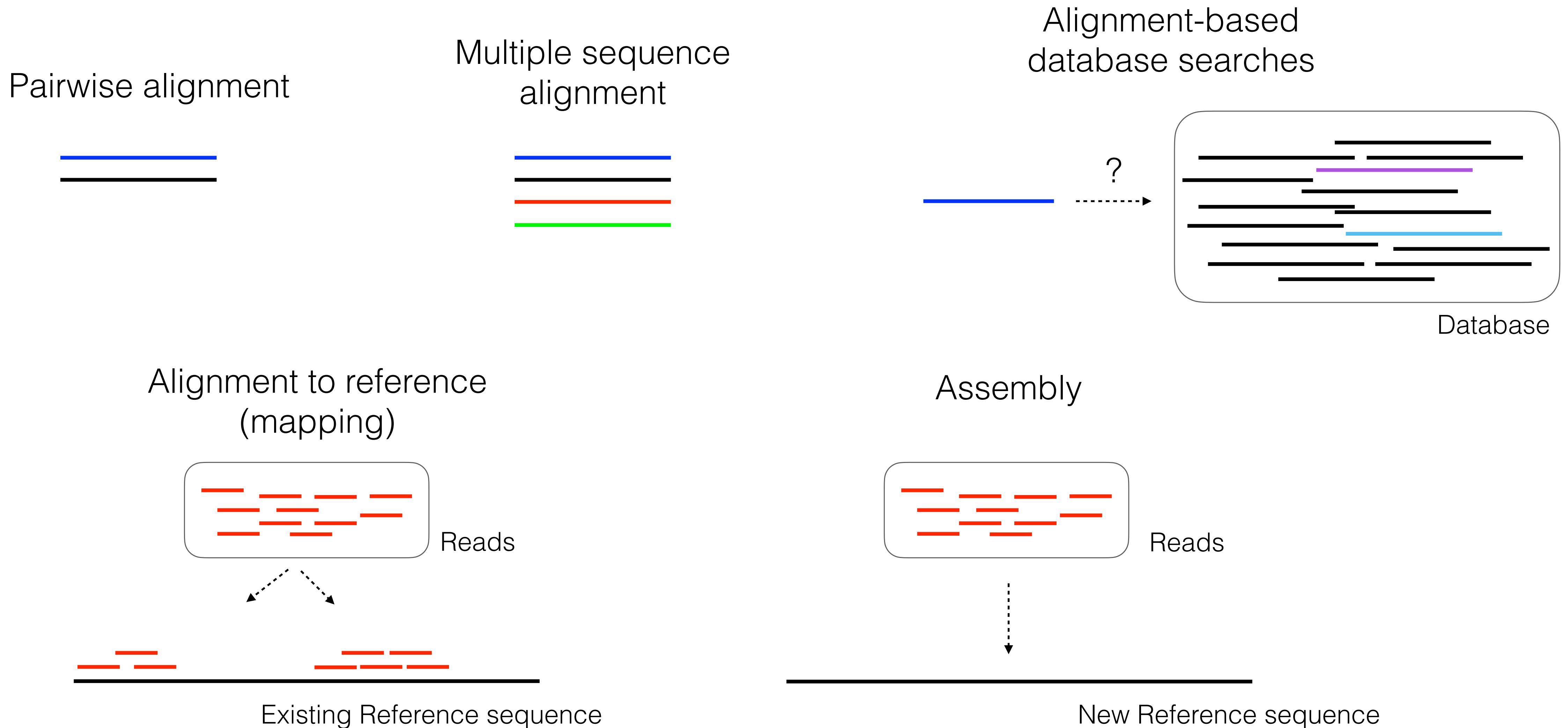


Image credit: Steve McCluskey CC BY-SA 3.0 [Link](#)



Sequence alignment is at the heart of a lot of  
bioinformatics and sequence analysis

# Major categories of sequence alignment



# Major categories of sequence alignment

Alignment type	Purpose	Commonly-used software
Pairwise alignment	Identify the similarities or differences between two sequences	<u><a href="#">Needle</a></u> (global alignment) <u><a href="#">Water</a></u> (local alignment)
Multiple sequence alignment	Identify the similarities or differences between >2 sequences. Input to tree building.	<u><a href="#">MAFFT</a></u>
Alignment-based search	Find the most closely related sequence in a database of sequences	<u><a href="#">BLAST</a></u>
Mapping (alignment to reference)	Determine the most likely location in a reference sequence from which a shorter sequence (a read) derives	<u><a href="#">BWA</a></u> <u><a href="#">Bowtie2</a></u>
Assembly	Create a new reference sequence using overlapping reads	<u><a href="#">SPAdes</a></u>