

An overview of genomics and sequencing terminology and practices

Mark Stenglein, GDW



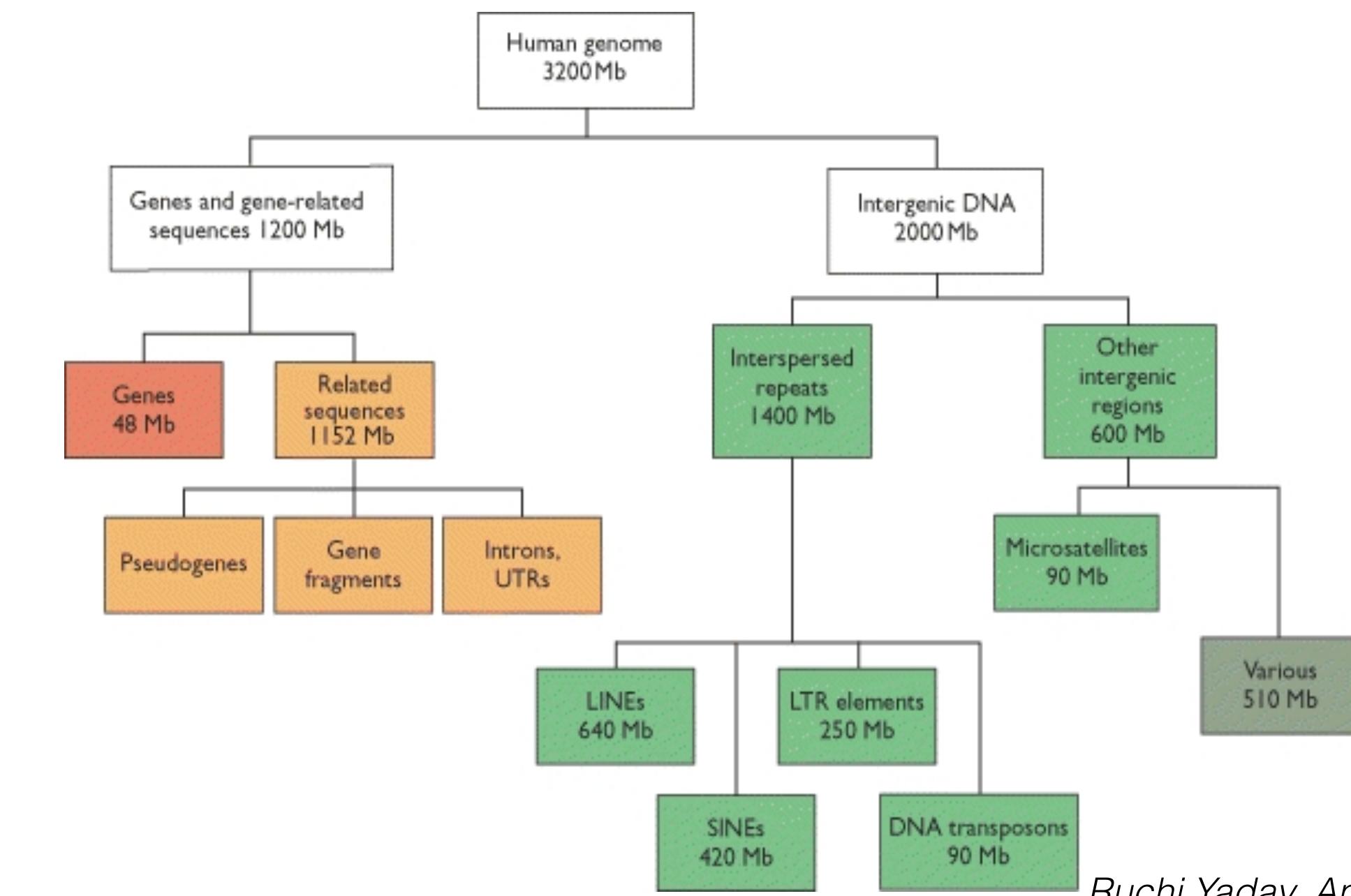
The jargon and terminology associated with genomics and ‘next gen’ sequencing can be confusing and intimidating



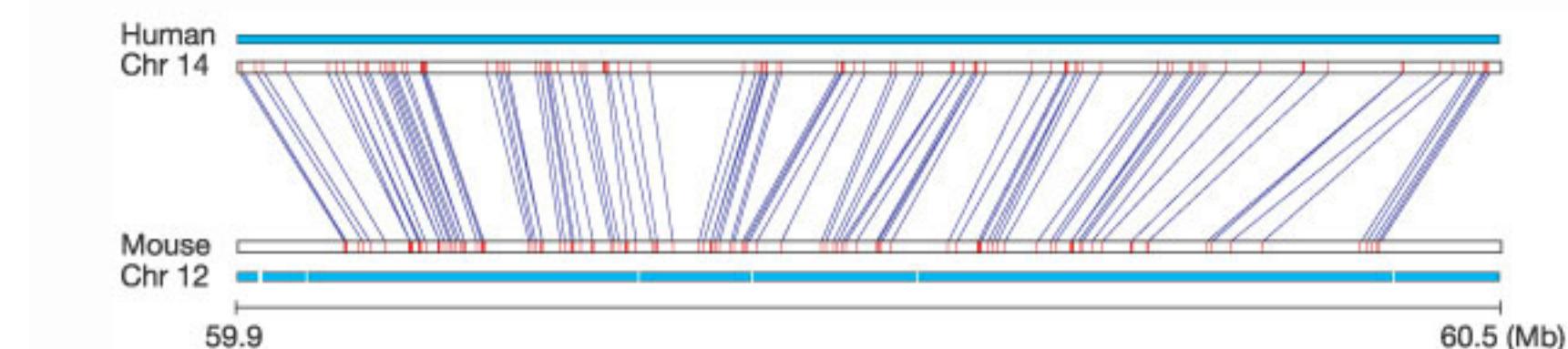
The goal of this lecture is to explain and demystify some common jargon
and explain how sequencing works

Genomics is the study of any of a number of attributes of genome or genomes

- Genome:
 - size
 - sequence
 - structure / variation
 - evolution
- Gene:
 - structure
 - expression
- Comparative genomics
- Epigenomics
- Metagenomics
- Transcriptomics
- Other -omics: Proteomics/Metabolomics

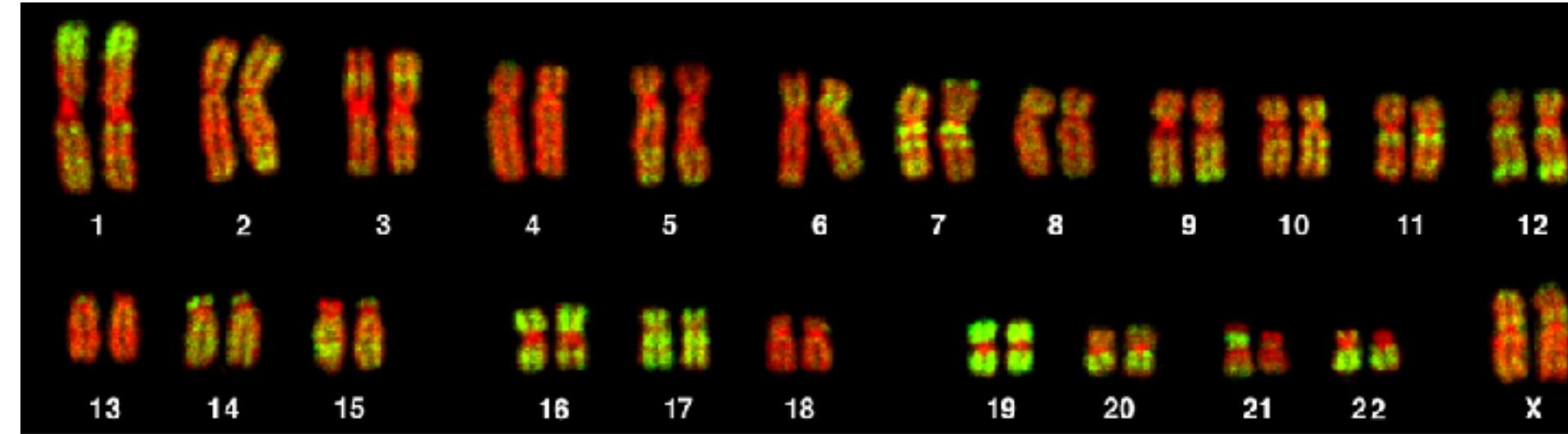


Ruchi Yadav, Amity University



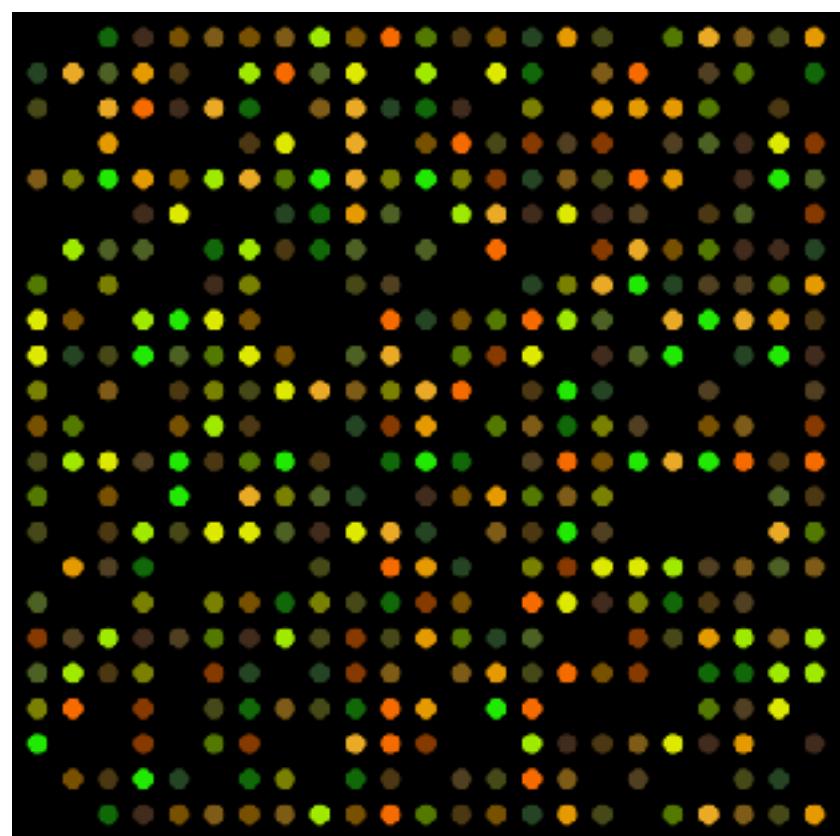
Nature (2002) Mouse Genome

Genomics isn't the same thing as sequencing, but they're increasingly related

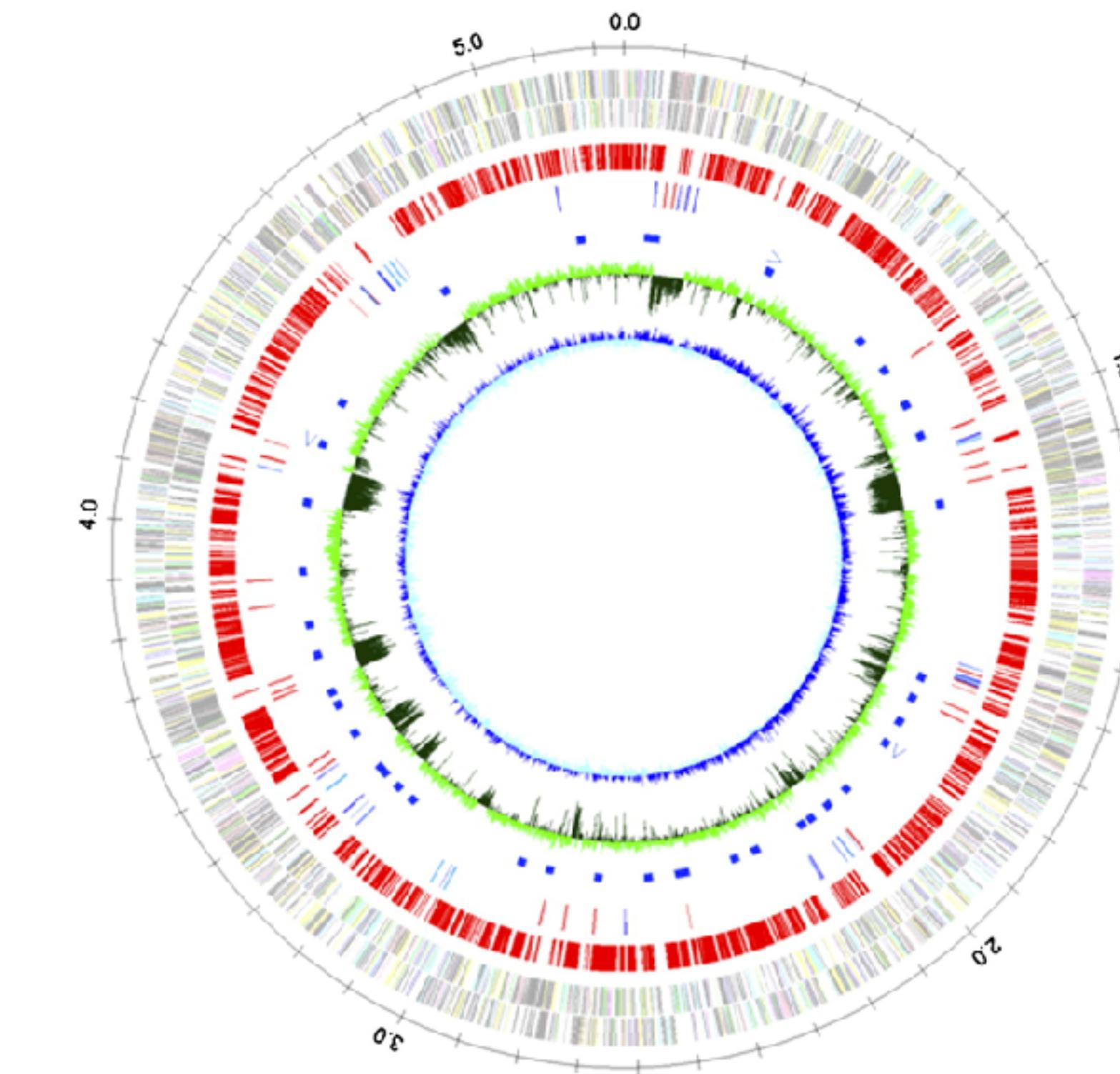


Bolzer et al (2005) PLoS Biol

Microarray

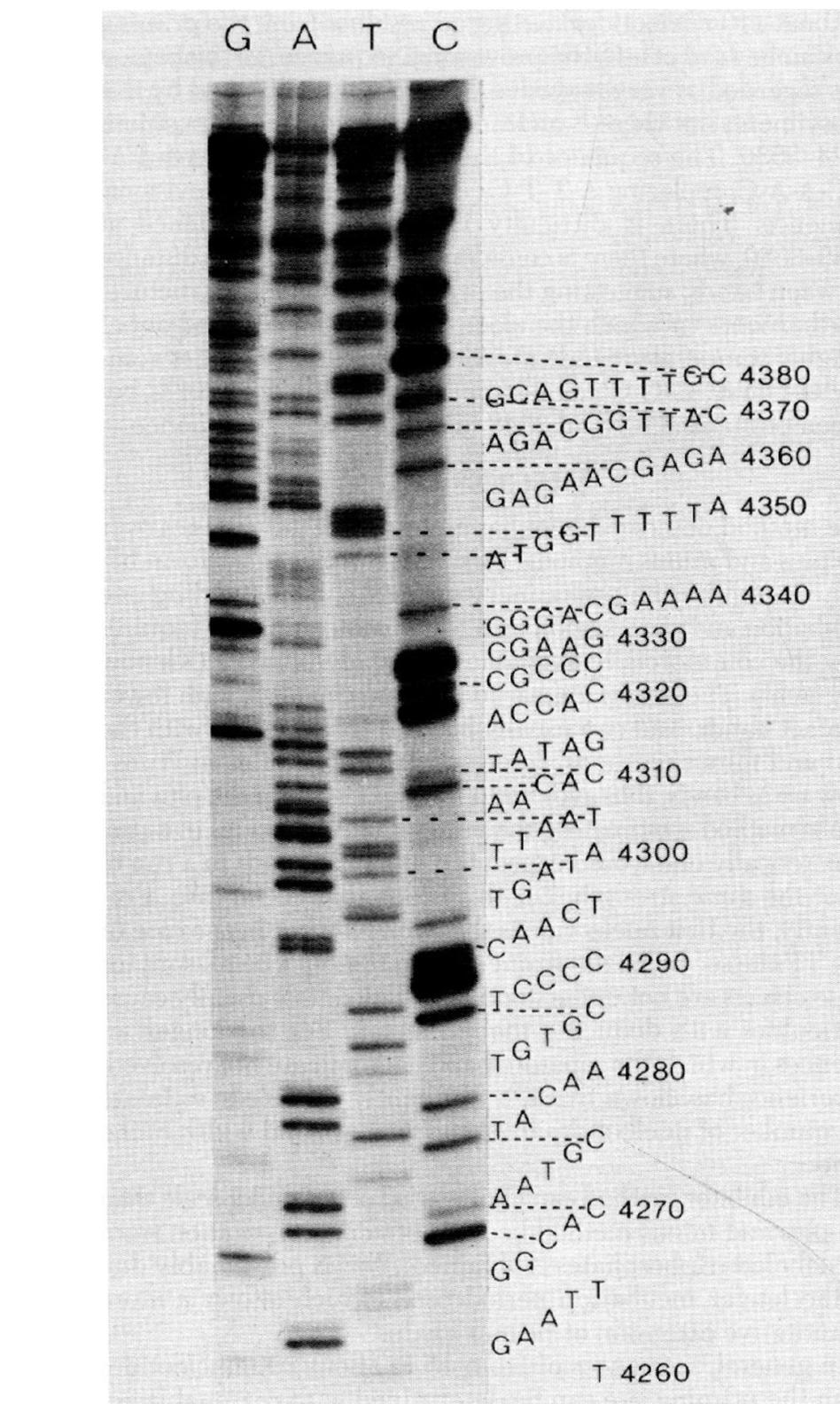
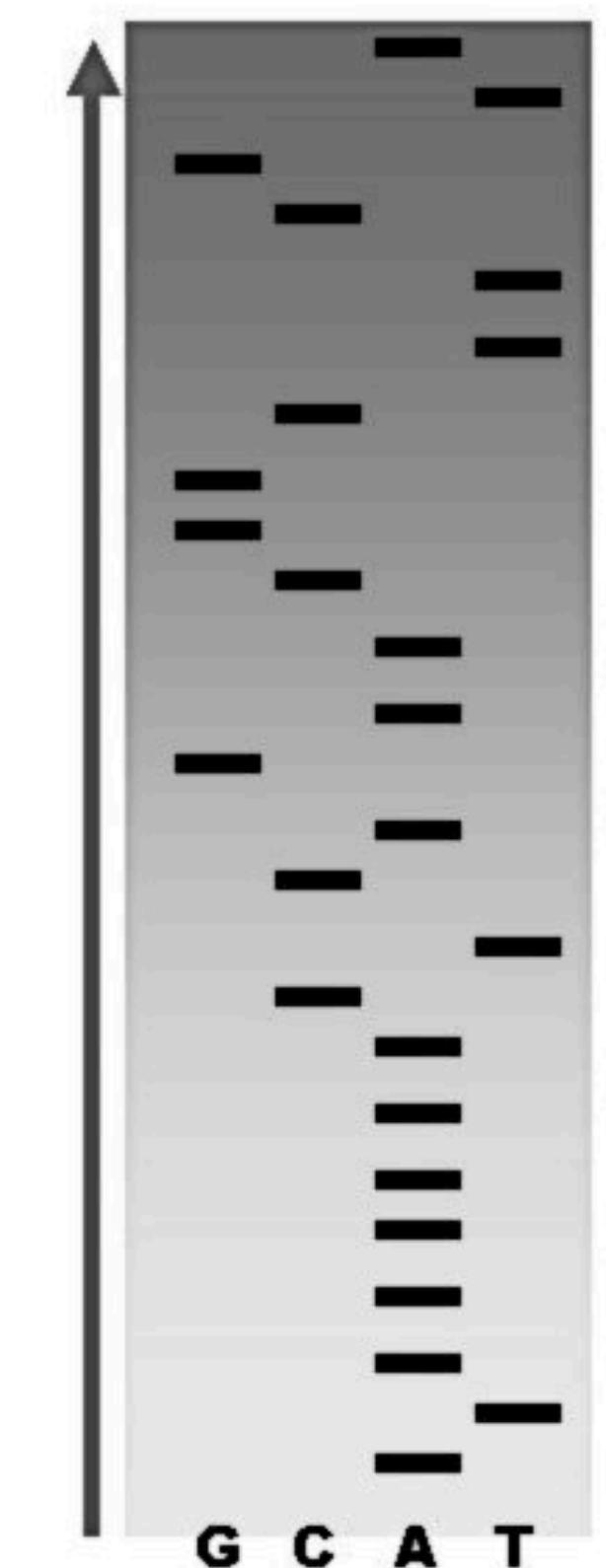
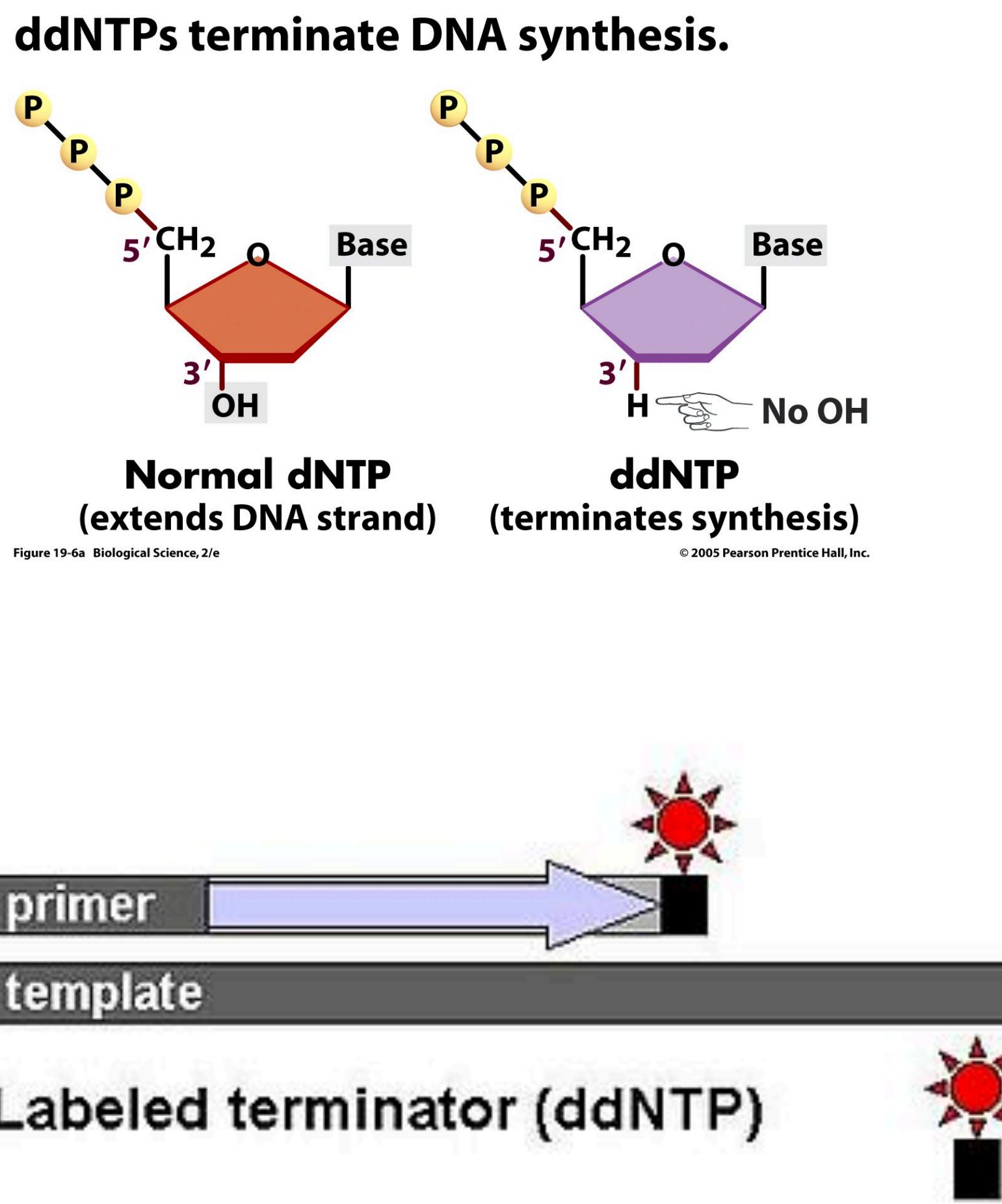


Wikimedia commons



Nakazawa et al (2009) Genome Research

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



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

- Millions of reads
- Relatively short: ~50-300 nt (Illumina)
- Relative low error rates
- Illumina has virtually all of the market share



MiSeq

Long read sequencing

- Fewer, longer reads
- >10 kbp (PacBio), up to 100s of kbp (Oxford Nanopore)
- Relative high error rates

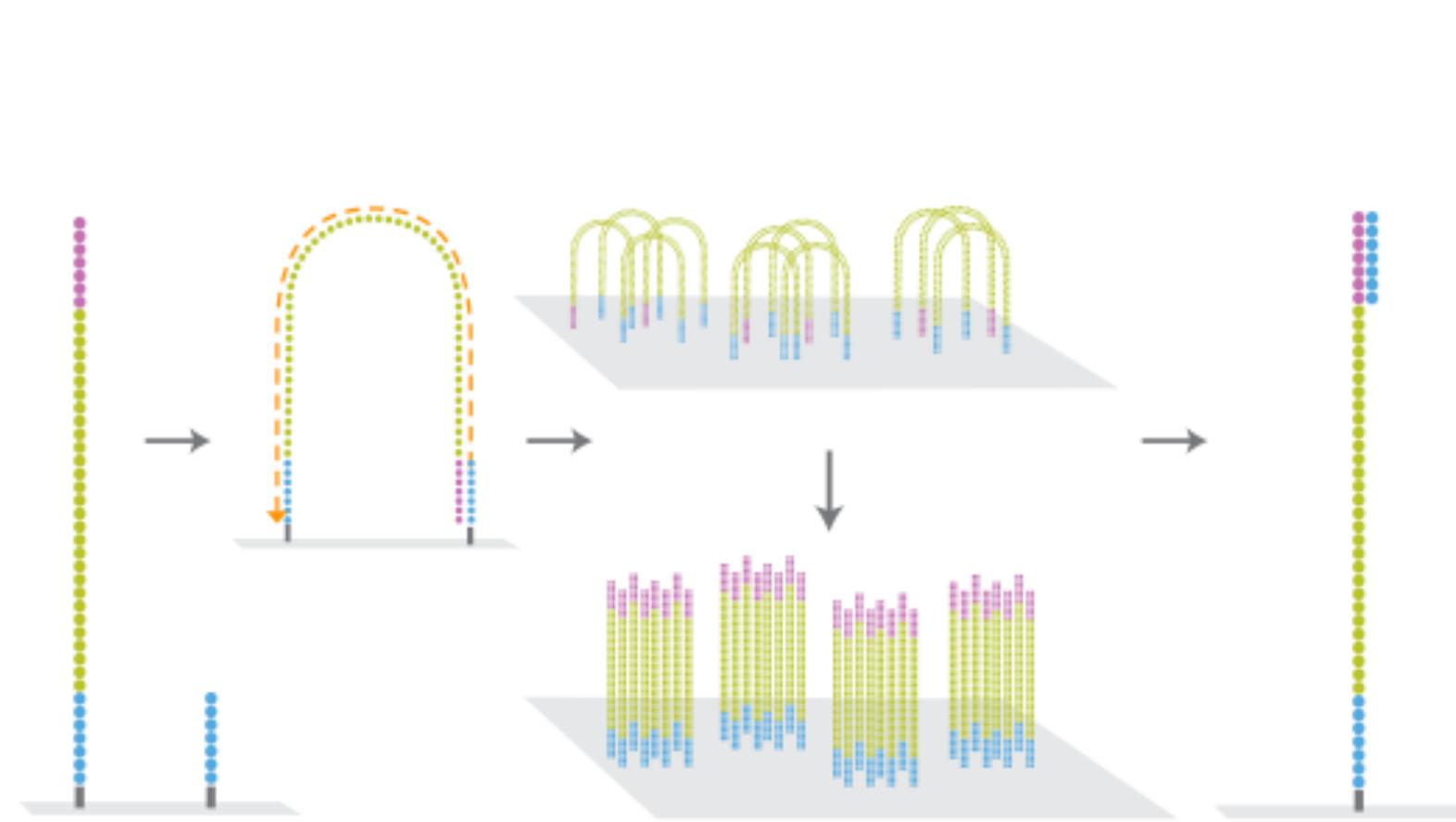
Oxford Nanopore MinION



PacBio RS-II

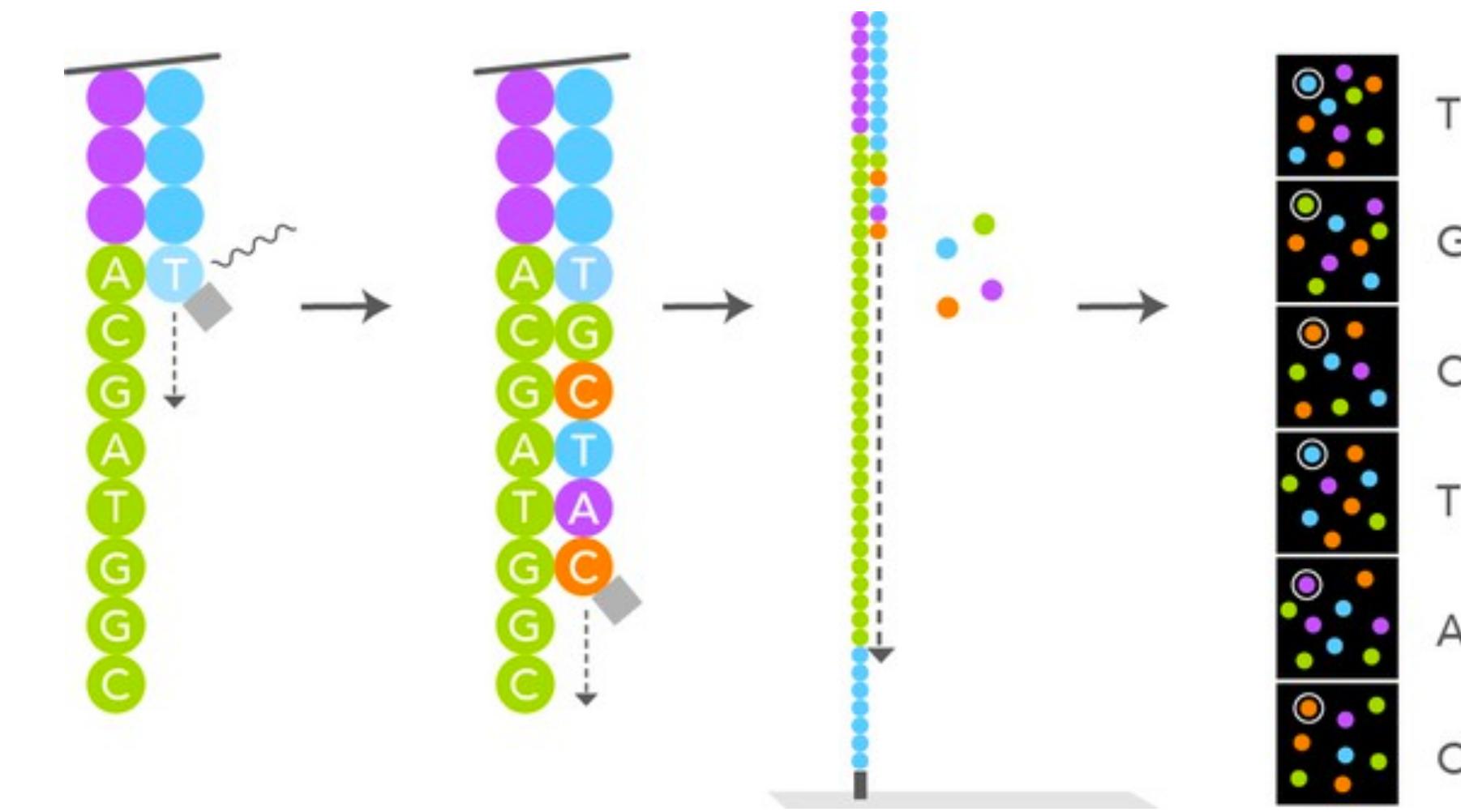


Illumina instruments use sequencing by synthesis (SBS)



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

real raw Illumina sequencing data

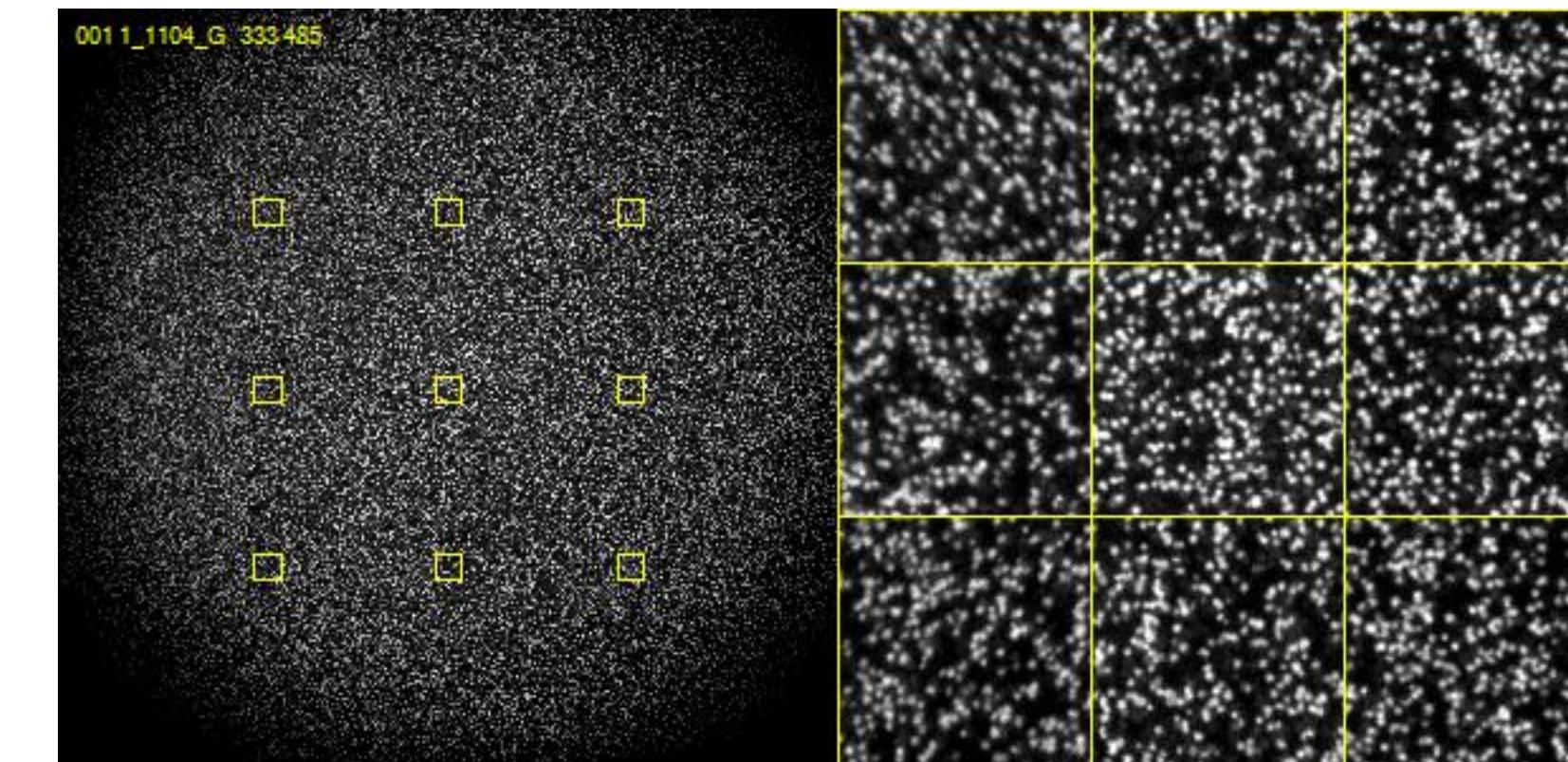
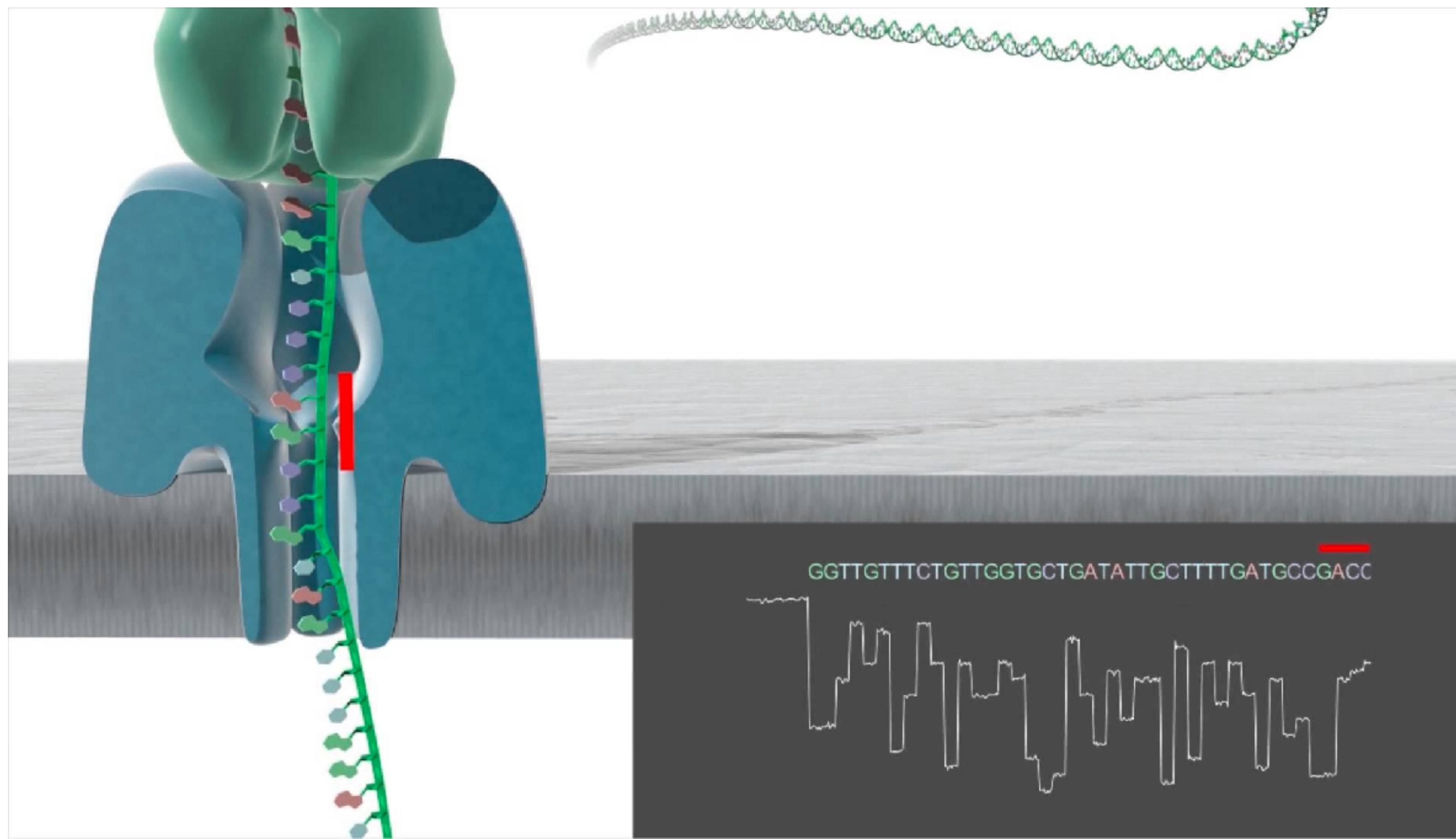


Image credit: Illumina

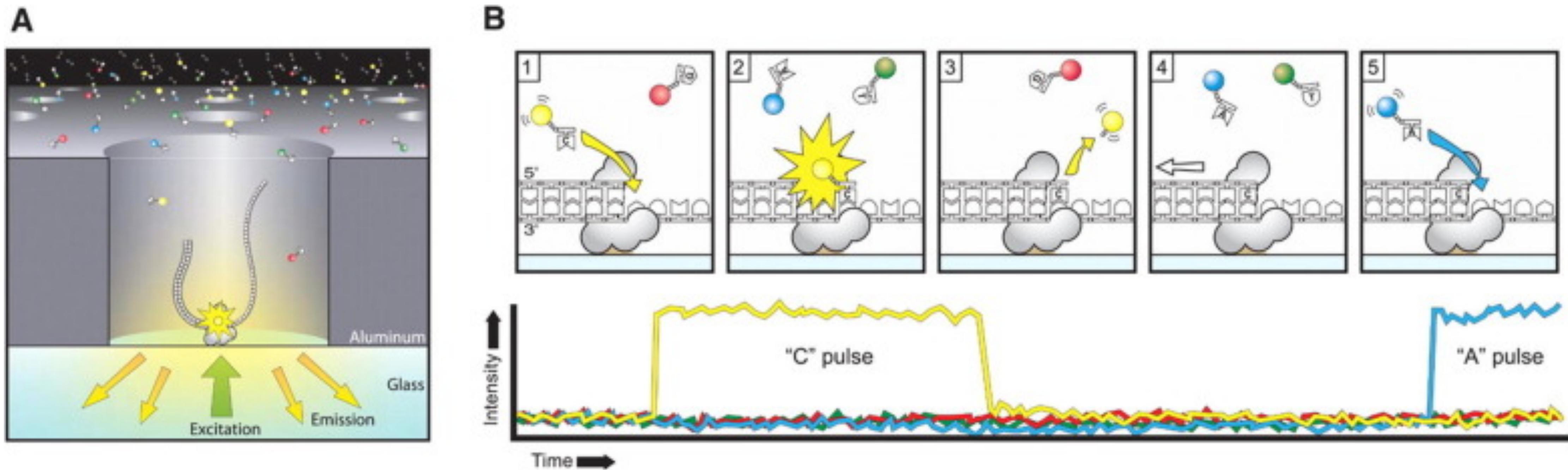
Long read sequencers sequence single molecules



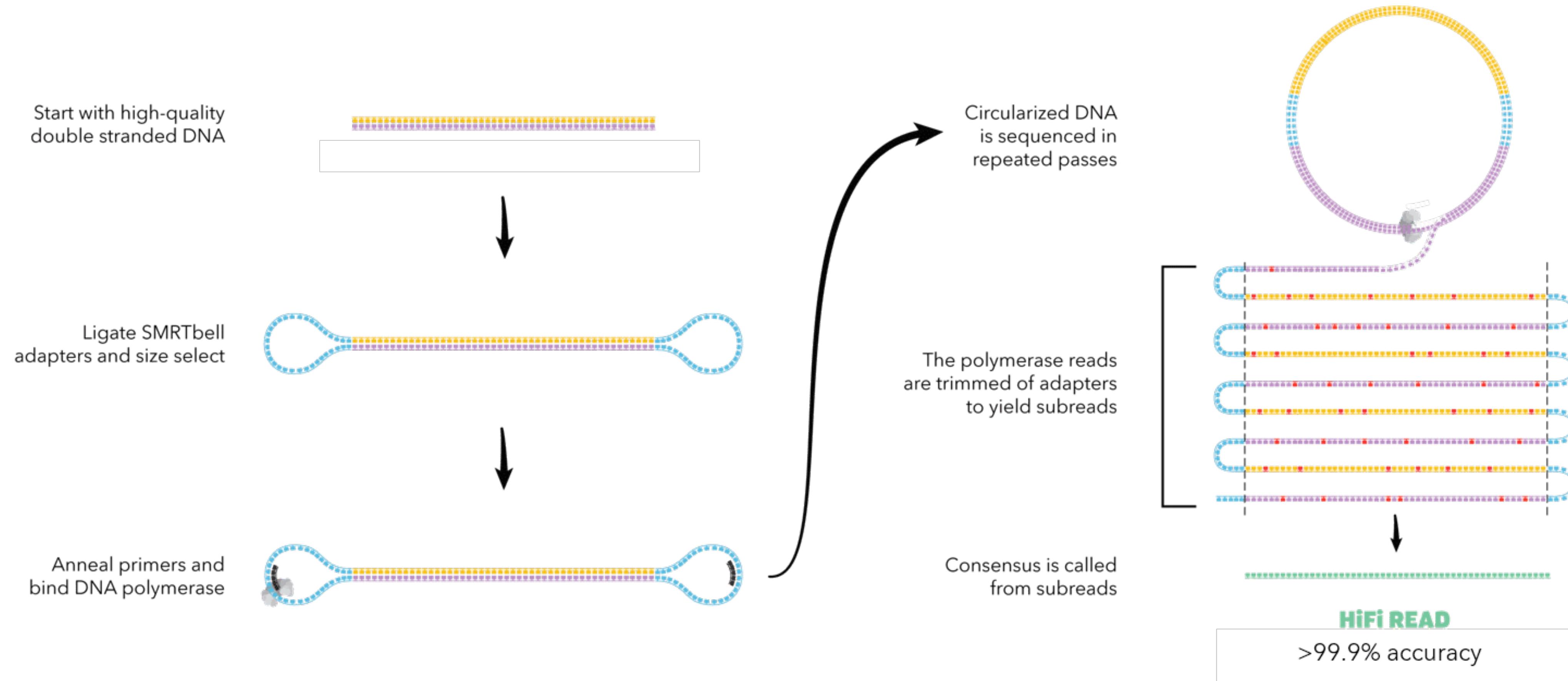
Nanopore sequencing: Much longer reads, but with much higher error rates

Image: Oxford Nanopore

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

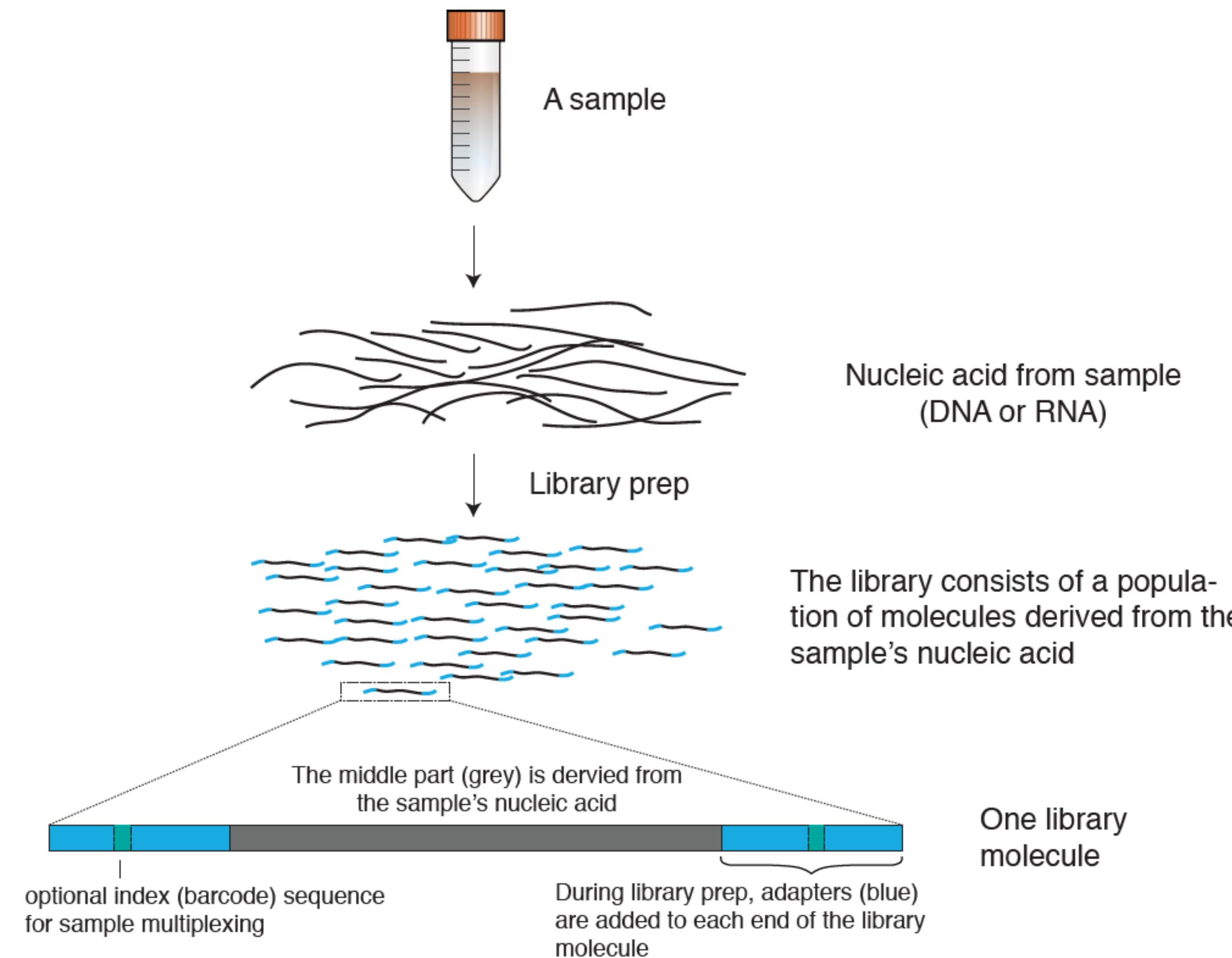


Best of both worlds: long reads with low error rate

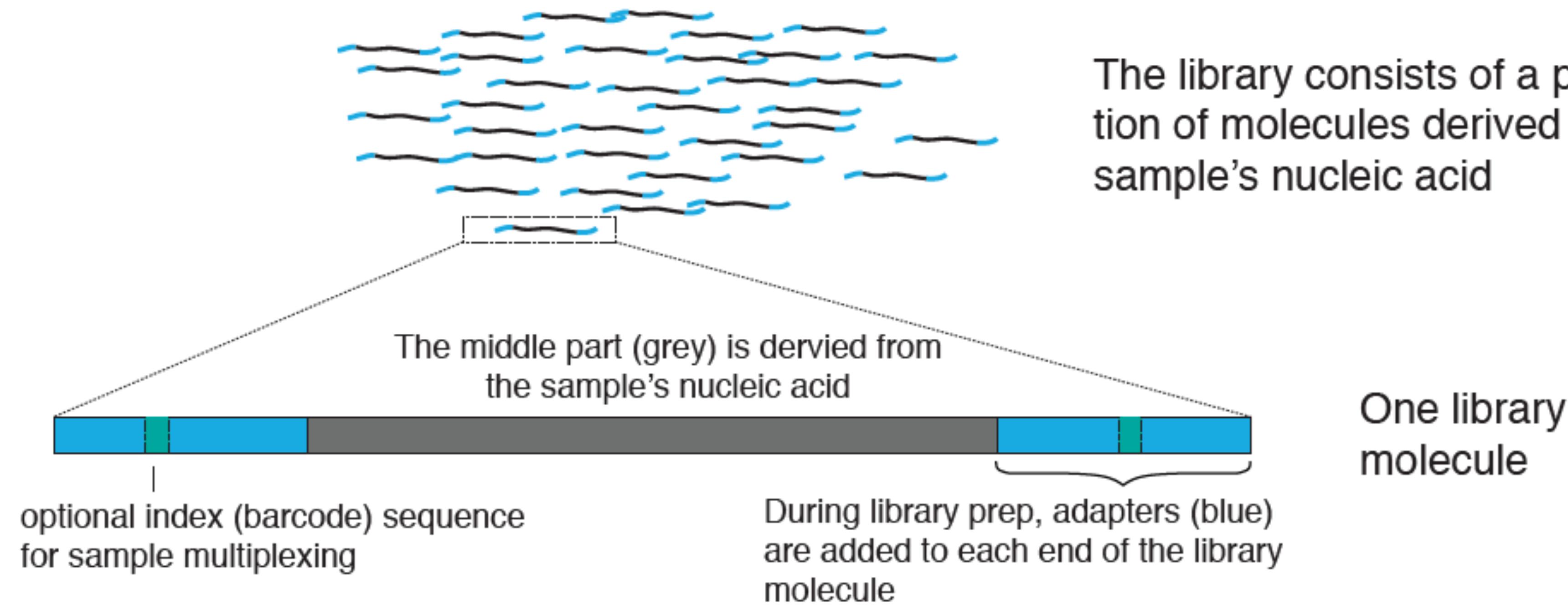
Image: PacBio

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

Details differ by sequencing platform

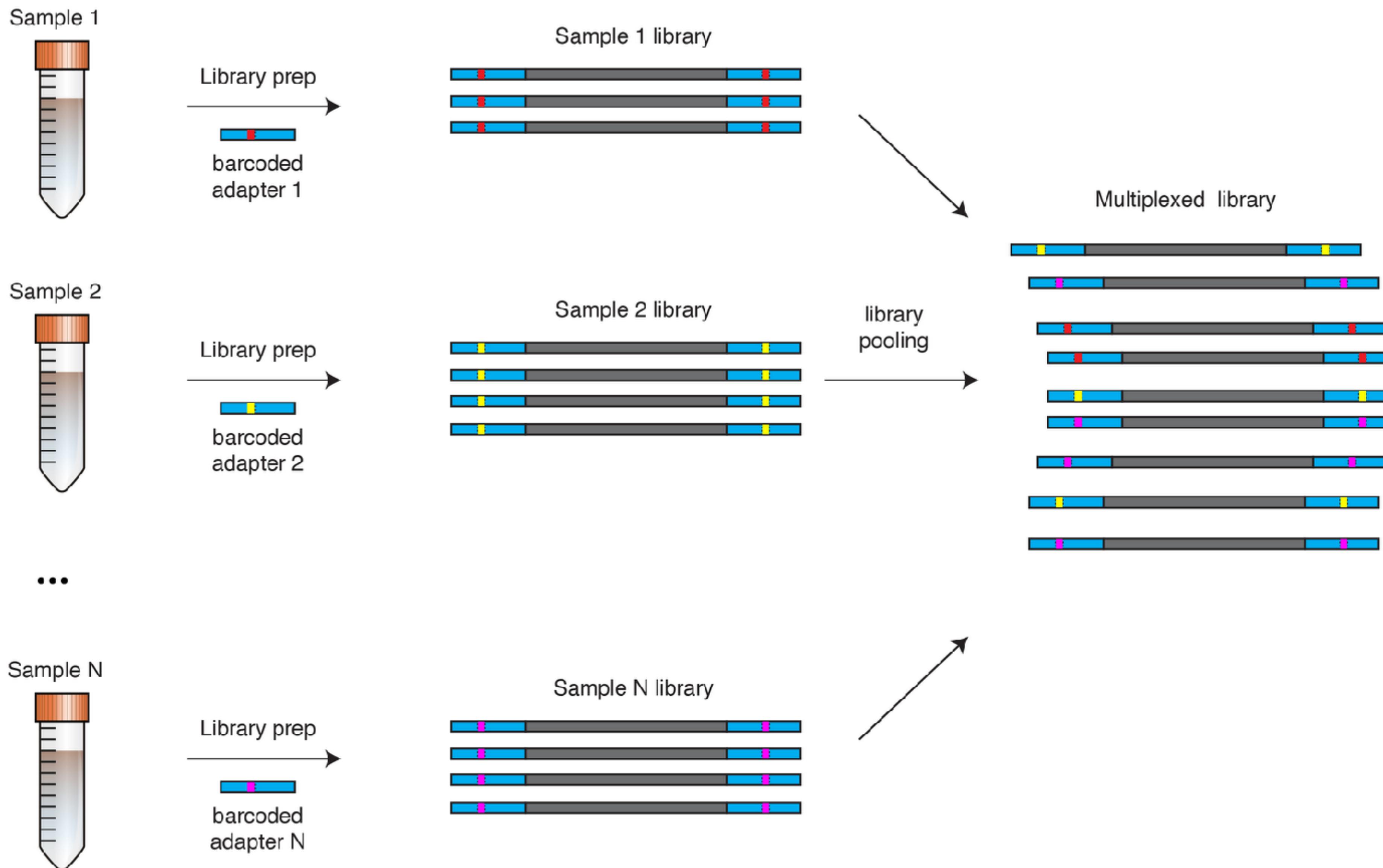


Library prep converts nucleic acids into a form suitable to be sequenced



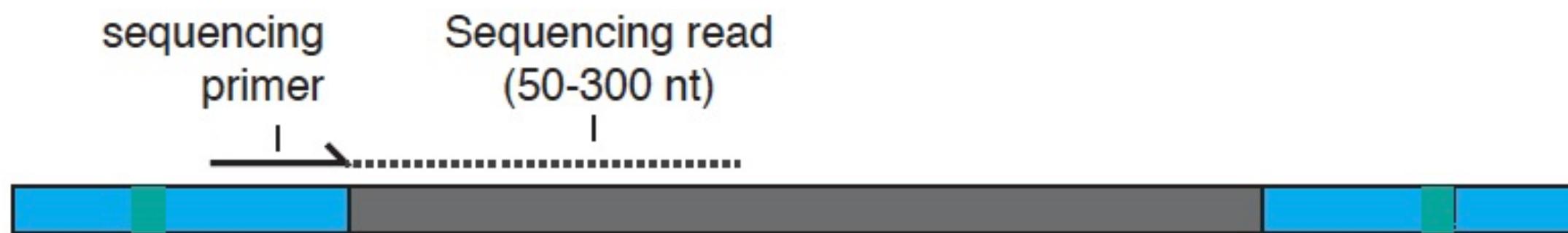
An example Illumina library molecule

Barcodes (or indexes) allow sample multiplexing



Illumina sequencing produces 1-4 reads per library molecule

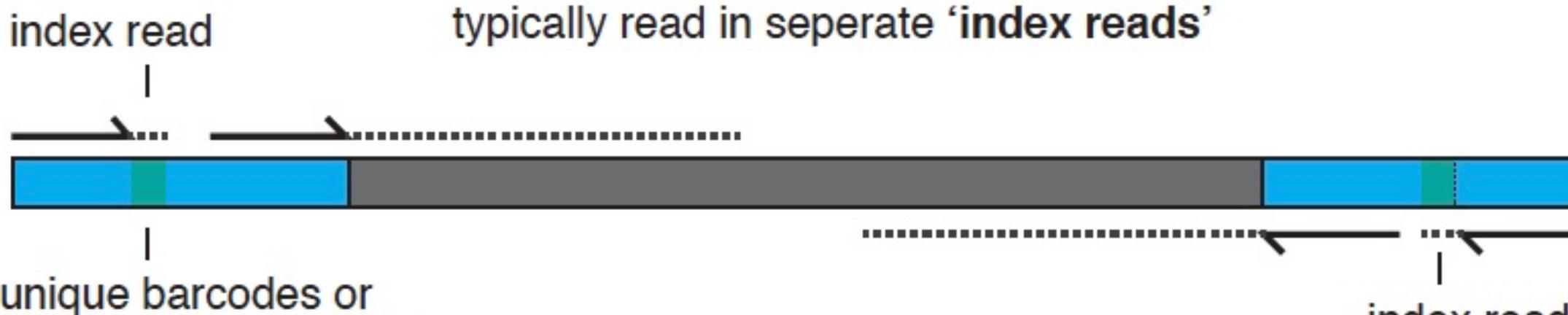
In **single end sequencing**, a library molecule is sequenced from one end



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

Reads are measurements of the real sequence that often contain errors

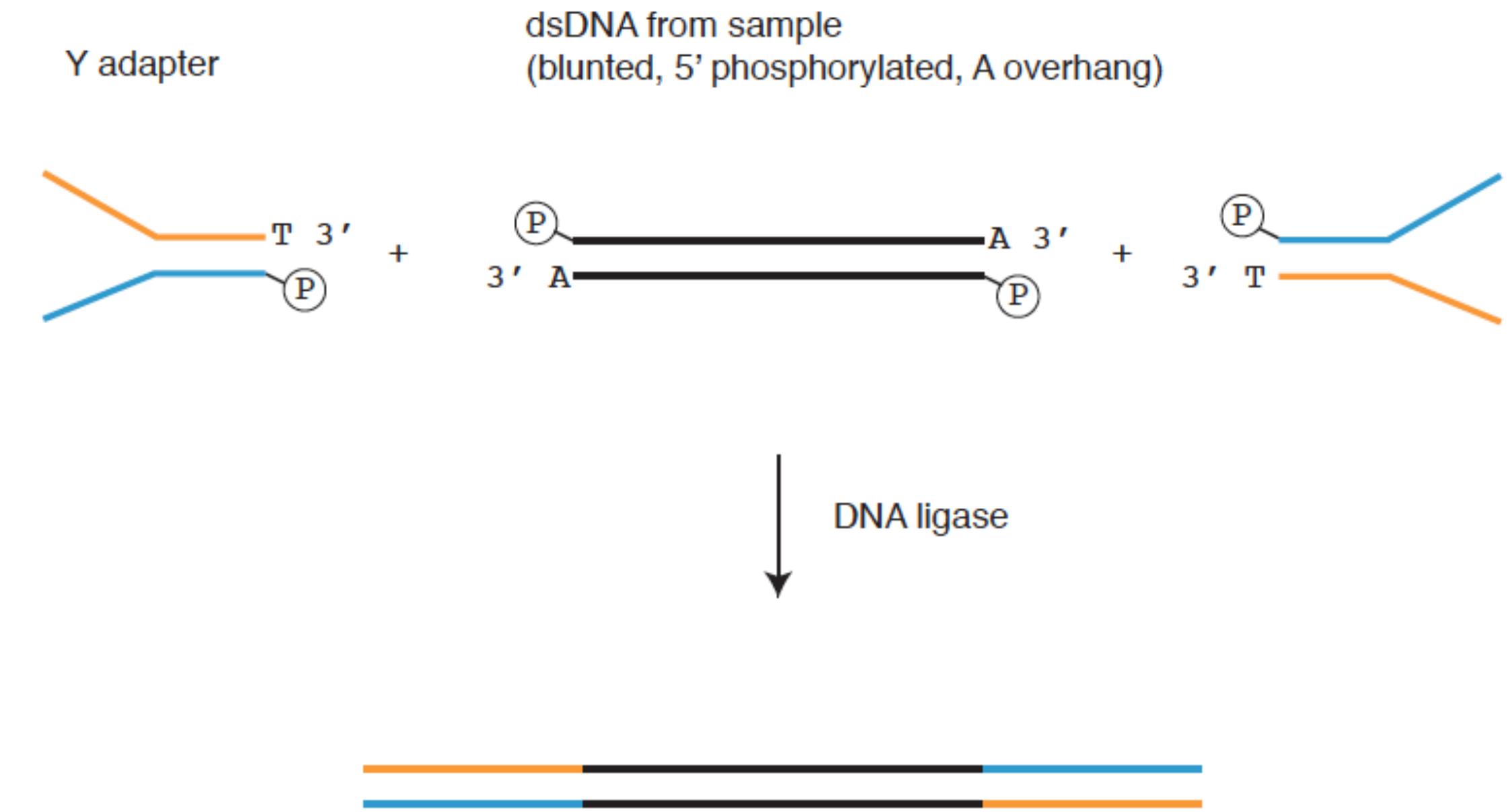


There are many good ways to make sequencing libraries

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

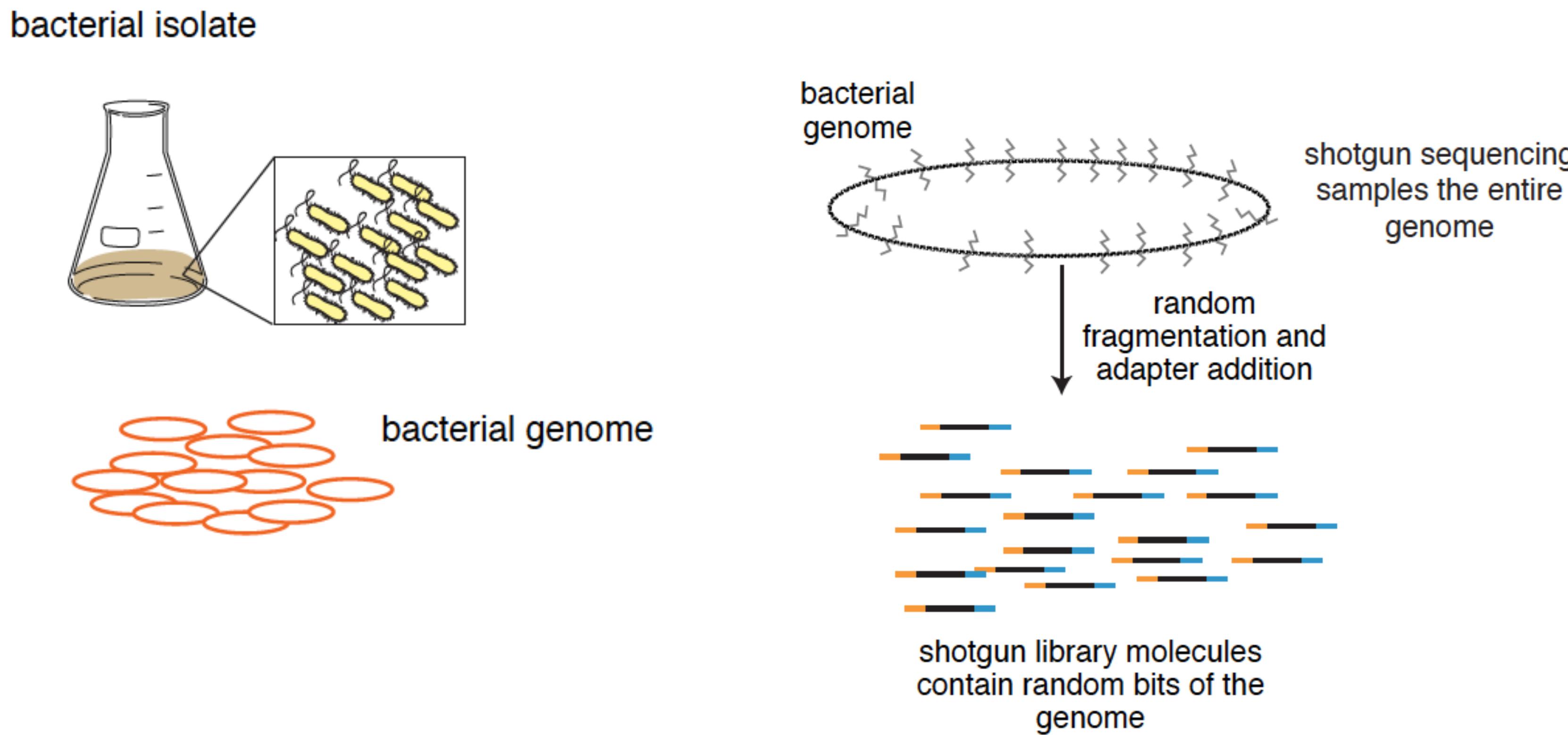
- Nucleic acid isolation
- Enrichment (of nucleic acid subtypes you want) or subtraction (of those you don't want)
- Nucleic acid fragmentation
- Conversion of RNA into dsDNA (for RNA sequencing)
- Addition of adapters to ends of library molecules, possibly with barcodes for multiplexing
- Library amplification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

Adapters can be added to sample-derived dsDNA by ligation



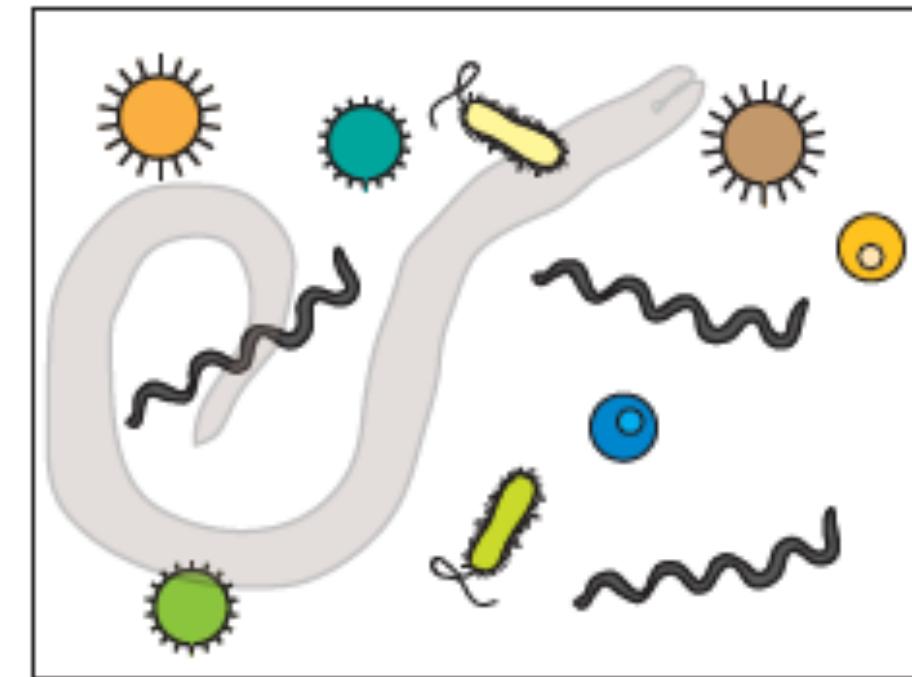
How you make a library determines what type of sequencing you're doing

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

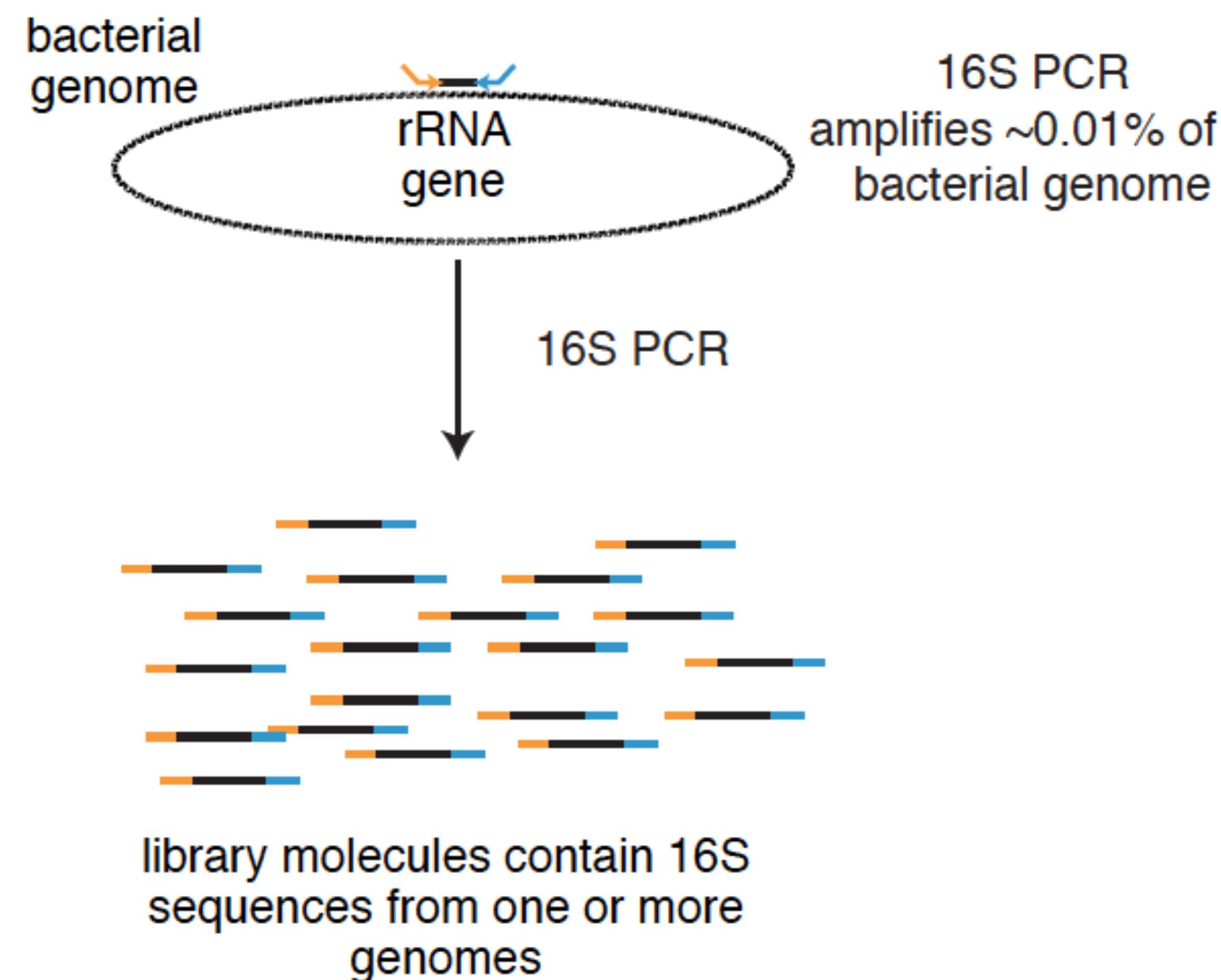
Sequencing of RNAs from a complex sample like this is metatranscriptomics

soil 'metagenome'

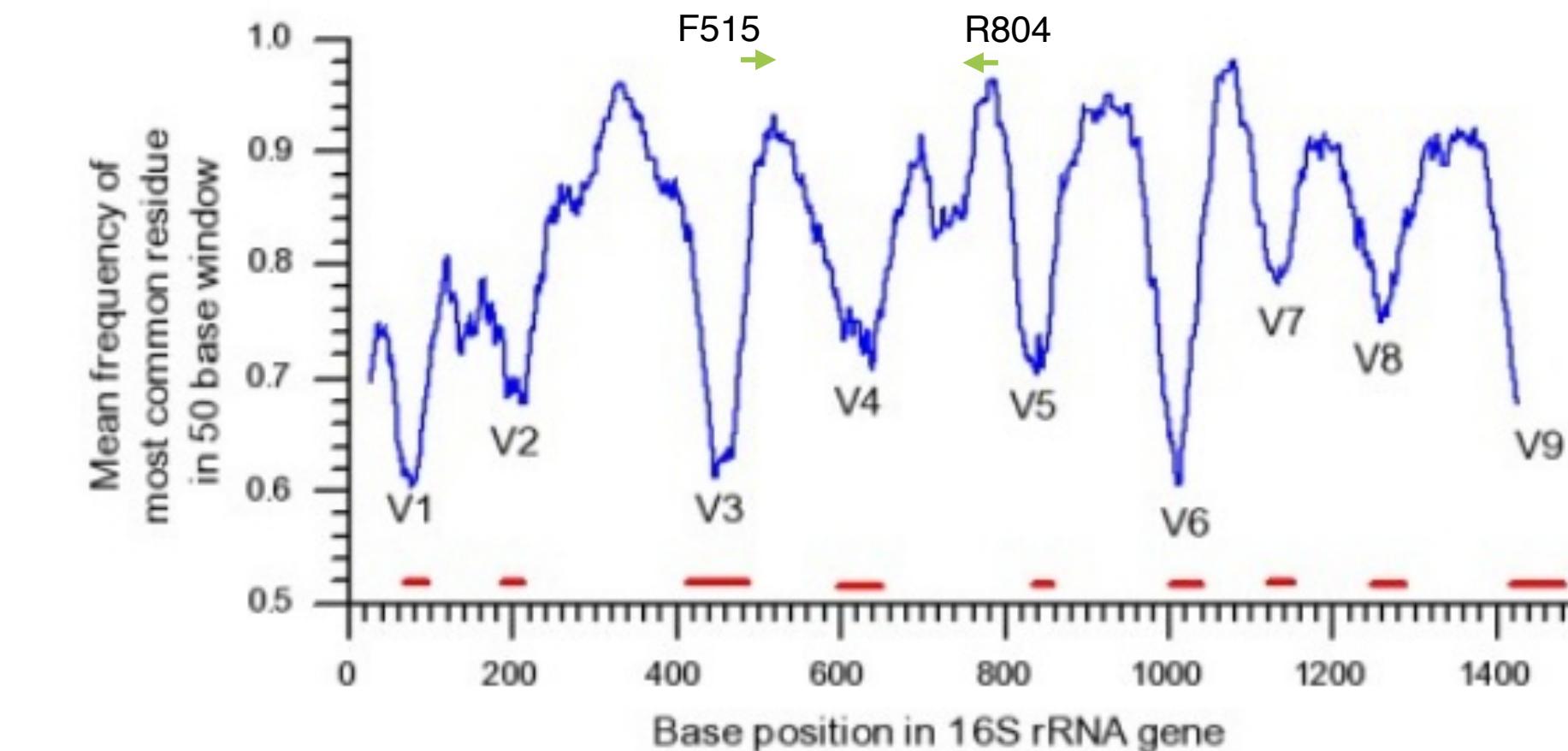


Microbiome sequencing often means **16S rRNA** sequencing

16S sequencing is one type of
'amplicon sequencing'

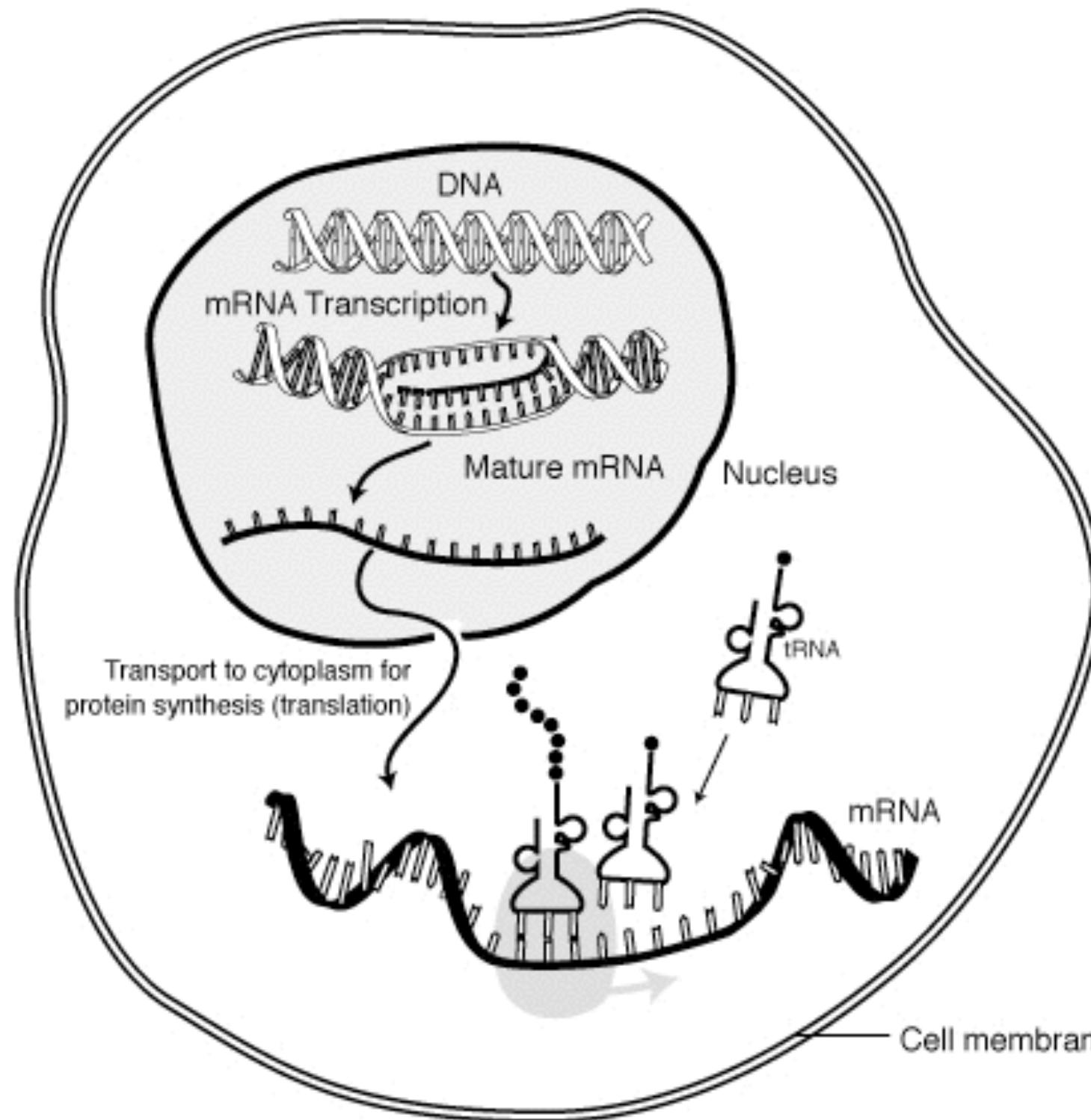


16S rRNA genes have highly conserved regions flanking variable regions



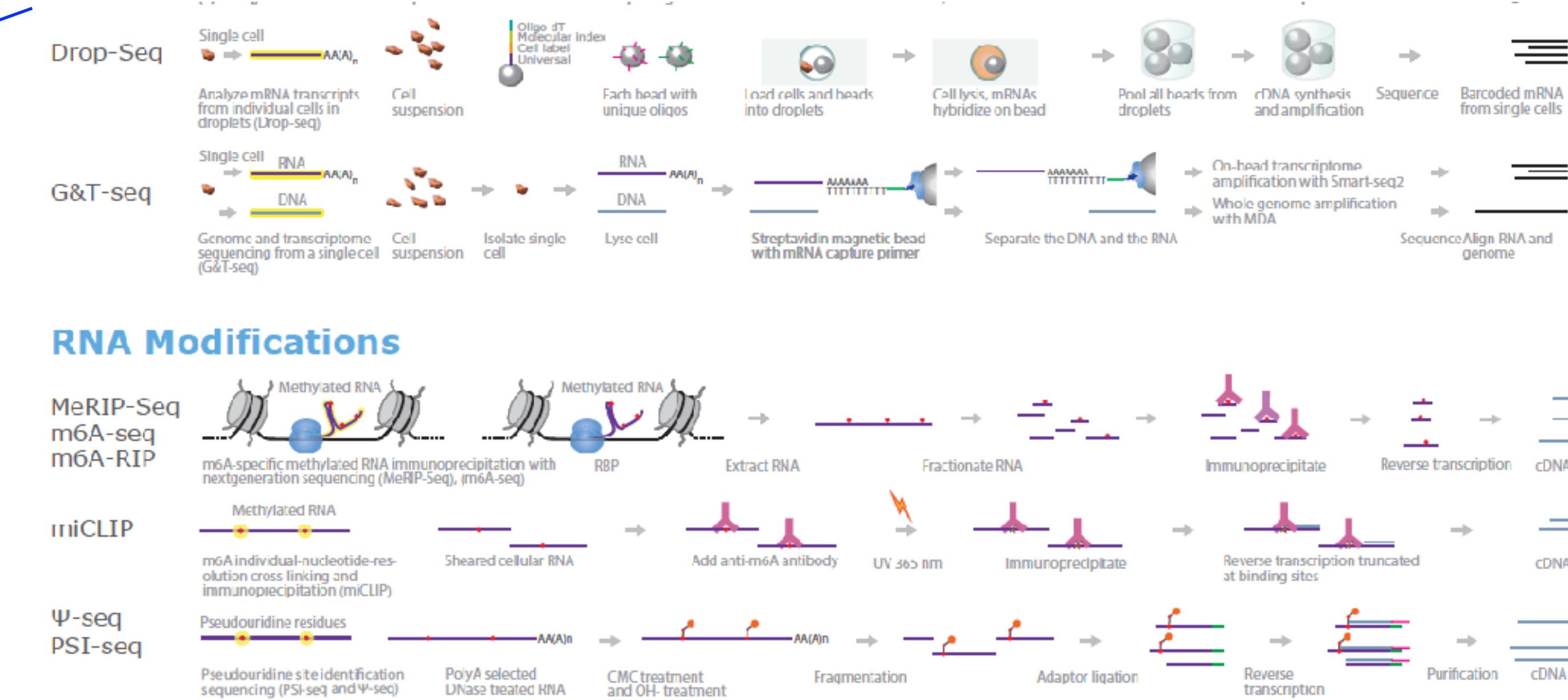
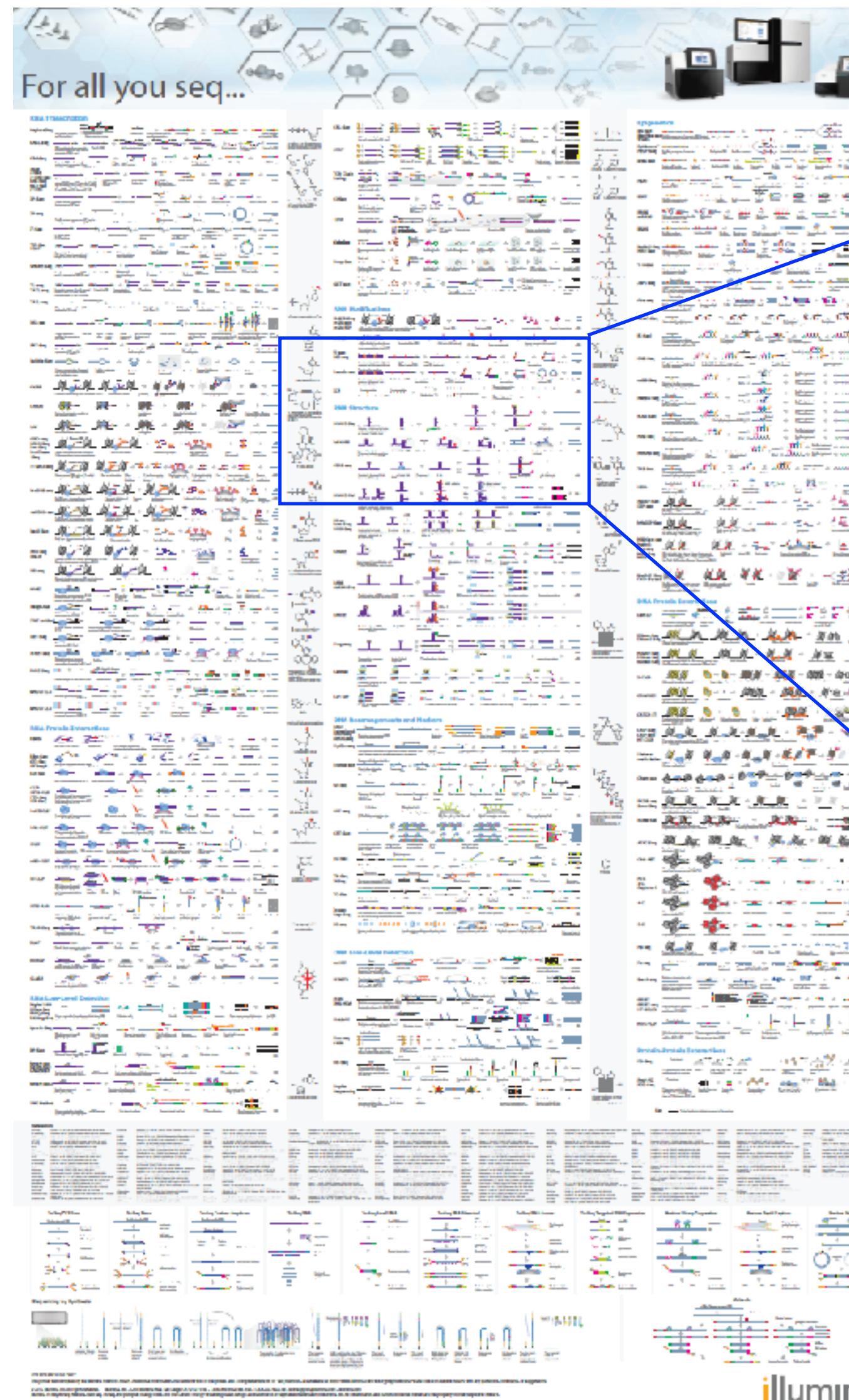
How you make a library determines what type of sequencing you're doing

If you make a library from mRNA, that is RNA-Seq (transcriptome sequencing)



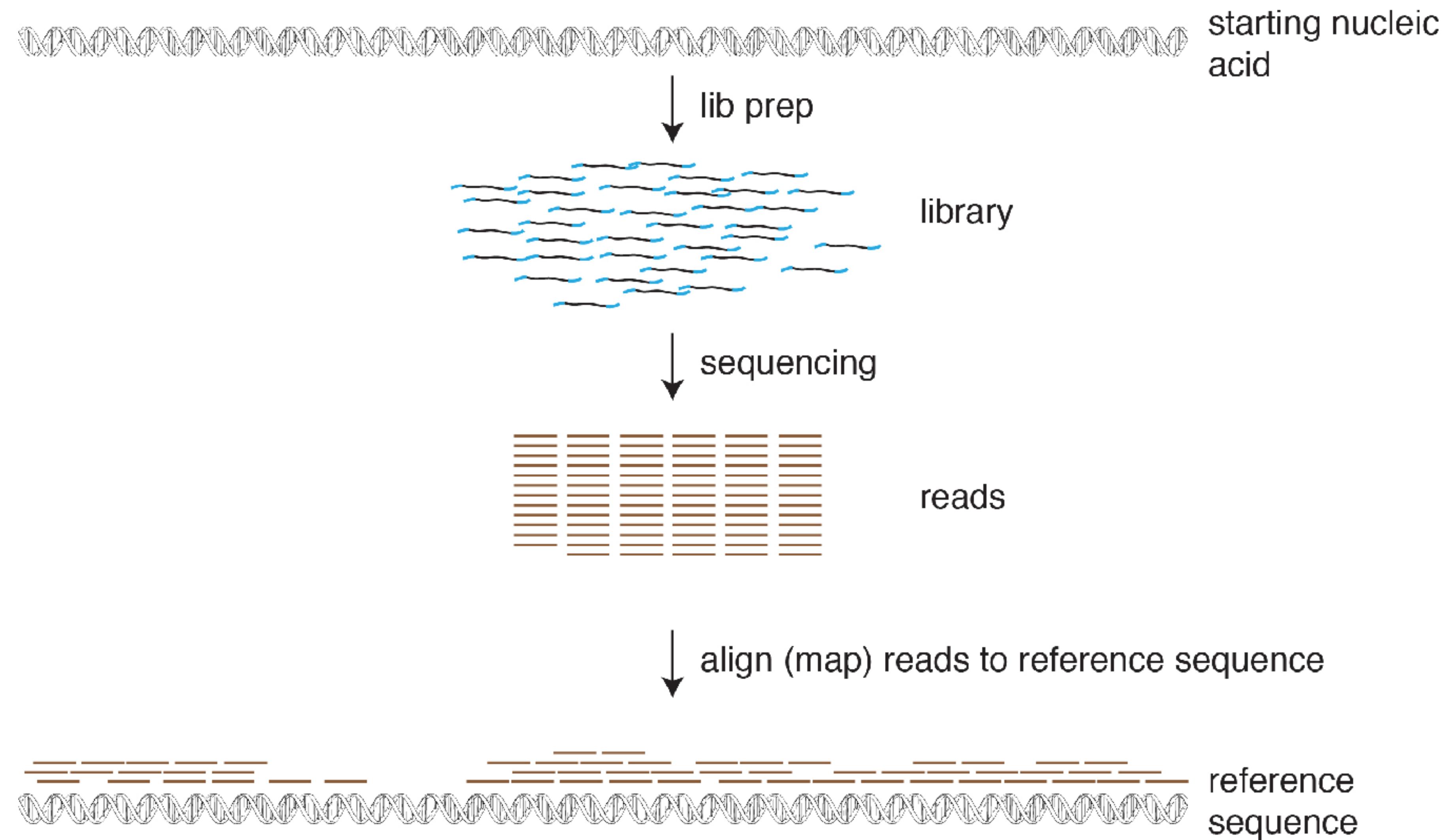
The abundance of reads from a particular mRNA is proportional to that mRNA's abundance in the cell

There are **5 gajillion** ways to make libraries and to do sequencing (all have names that end in -seq)

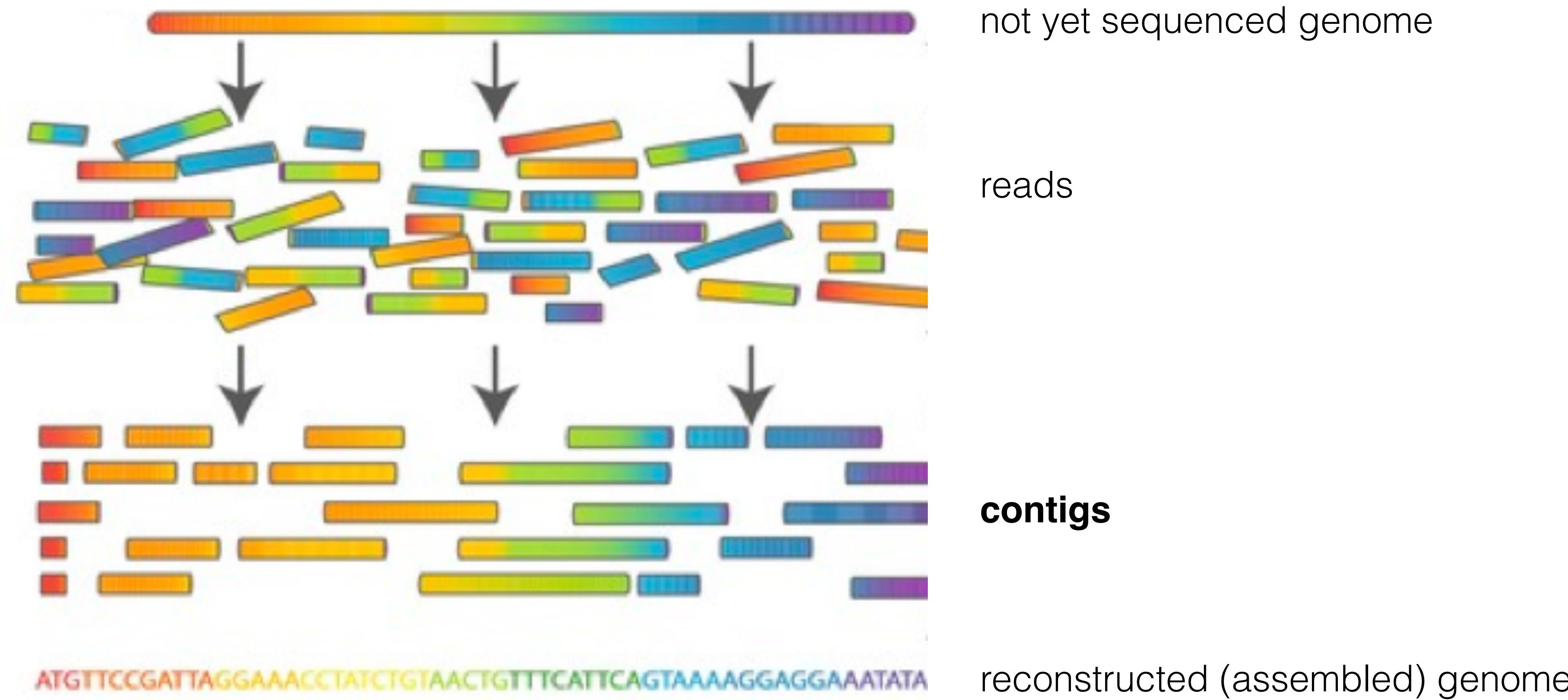


They're all variations on a few themes. Don't let it overwhelm you. Most sequencing is of a few simple types, and it's better to focus on the Biology and experimental design.

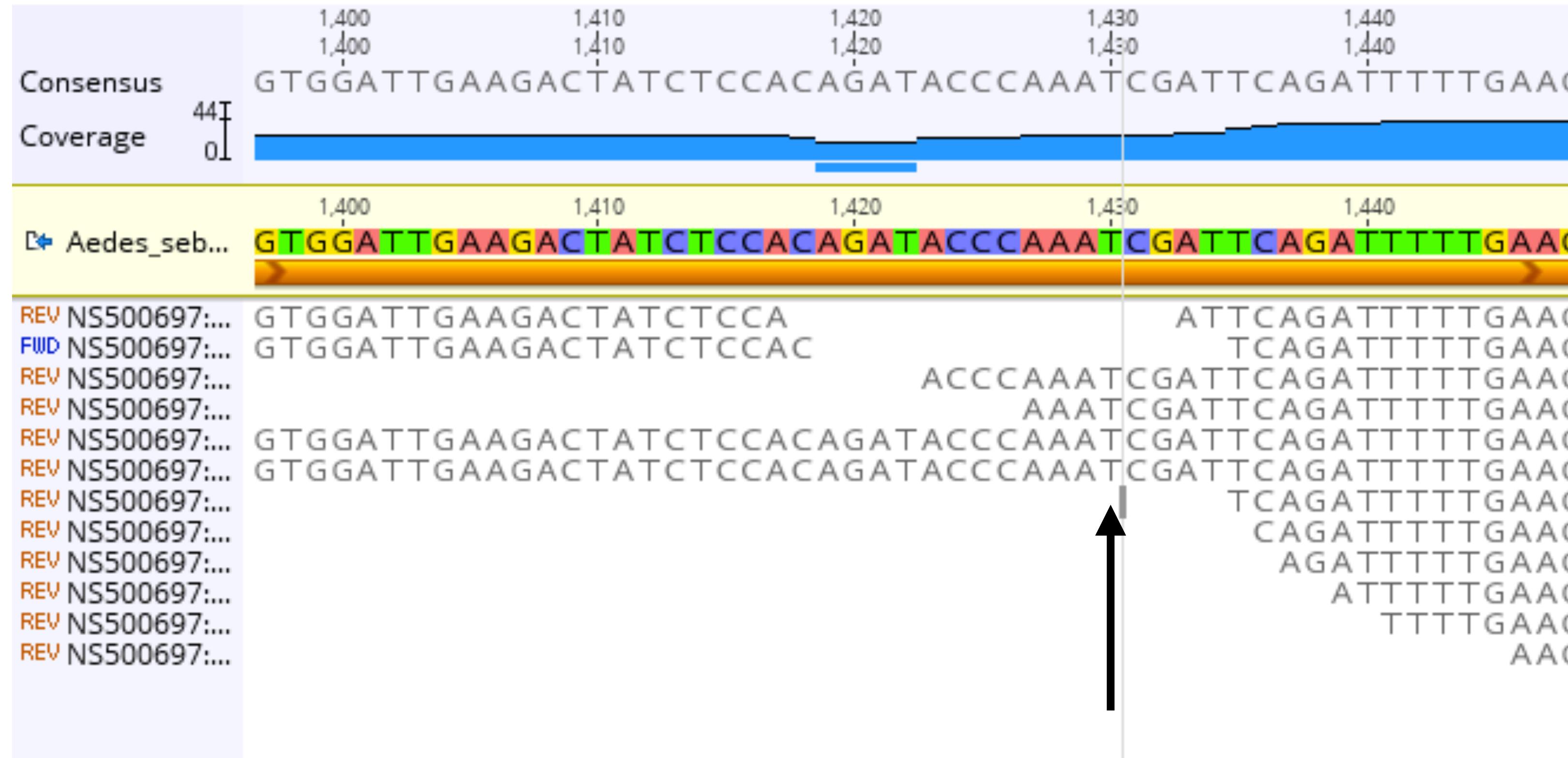
Mapping is the process by which sequencing reads are aligned to the region of a genome from which they derive.



(De novo) **genome assembly** is the process of trying to reconstruct a genome sequence from reads



Coverage is the number of individual aligned reads that support a particular nucleotide in a reference sequence

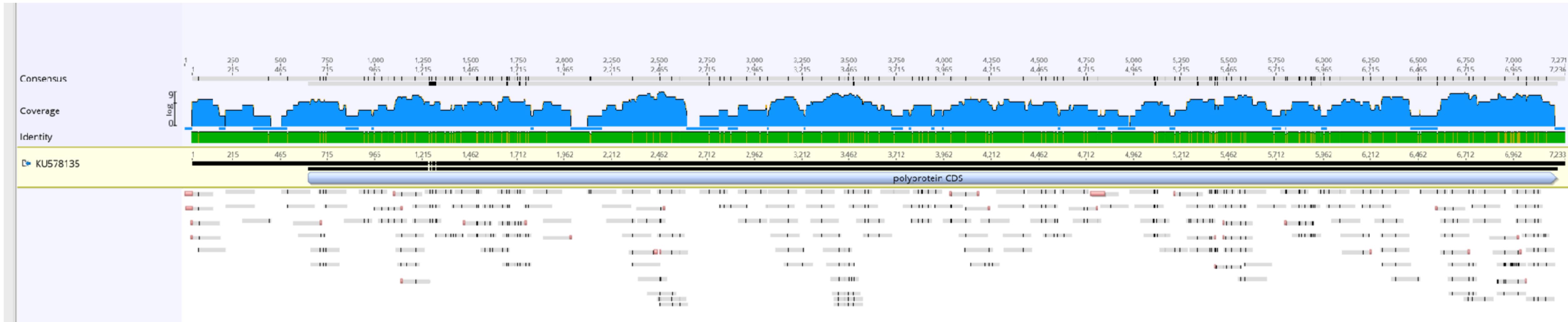


coverage is often referred to as
'depth' or 'depth of coverage'

This T has 4x coverage

Coverage is also used to describe the fraction of a genome with >0x read coverage

reads from human oral swab RNA aligned to a coxsackie virus genome



96% genome coverage (96% of bases have >0x coverage)
3.4x average coverage depth (range 0-9x)



(Mayo clinic)

Questions?

Is there genomics or sequencing jargon about which you're not certain?