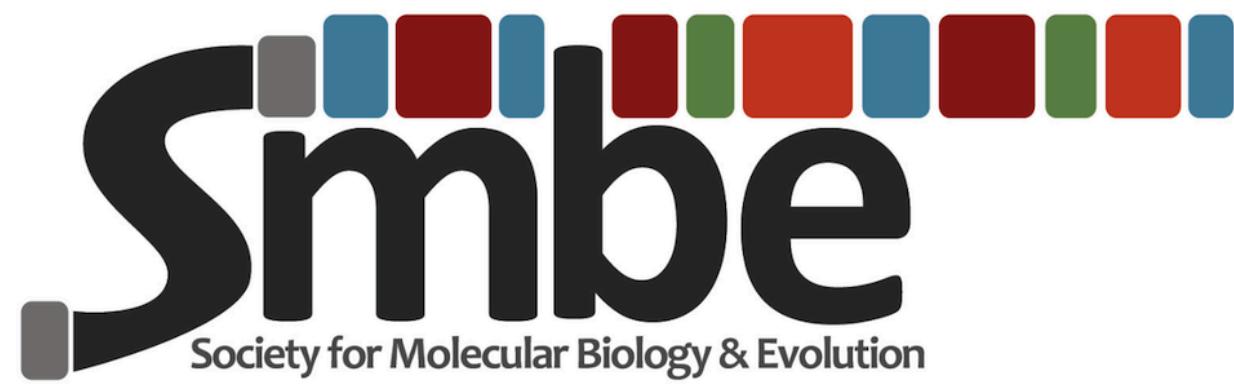


An overview of genomics and sequencing terminology and practices

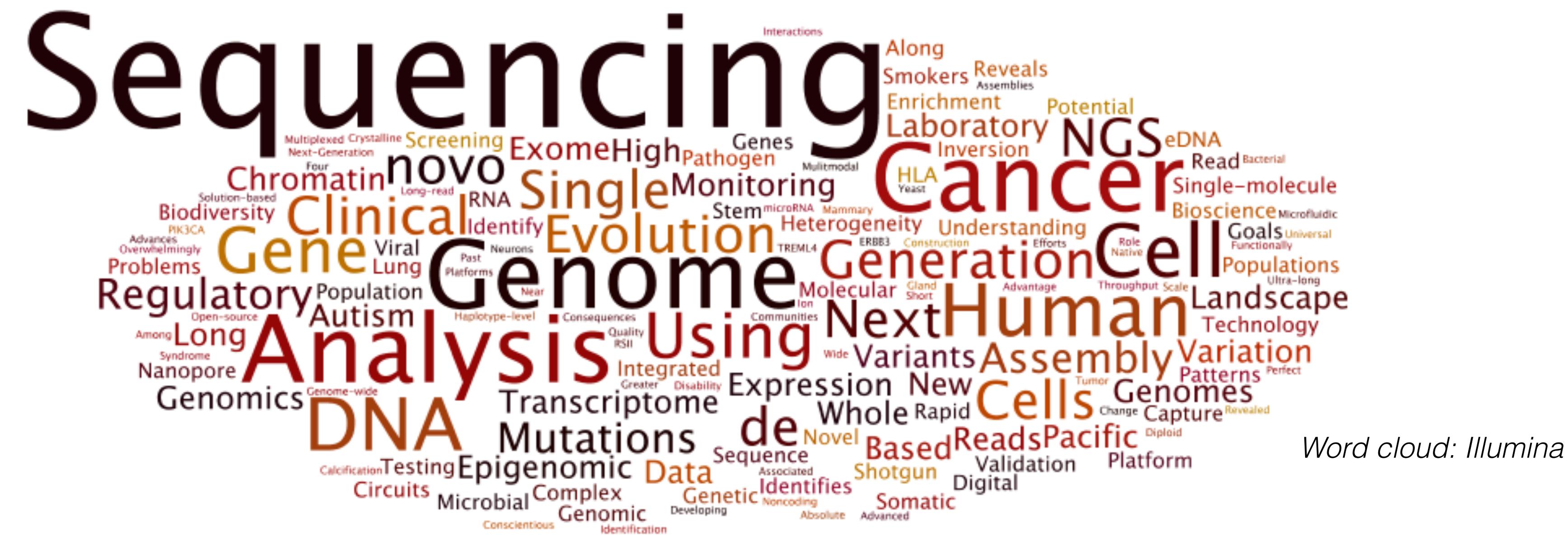


SMBE Regional Workshop
on Computational Biology
in Todos Santos, Mexico

Mark Stenglein

 Colorado State University Todos Santos Center
April 8-12, 2019

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

There is a glossary available online that explains many of these terms

- [Transcriptome](#)
- [Variant](#)
- [WGS](#)

16S

The **16S** ribosomal RNA gene is present in all bacterial and archaeal genomes. This gene is sufficiently conserved that primers that anneal to conserved regions of the gene will amplify essentially any prokaryotic 16S rRNA gene. These PCR products (amplicons) can be sequenced to provide a survey of microbial diversity in a sample.

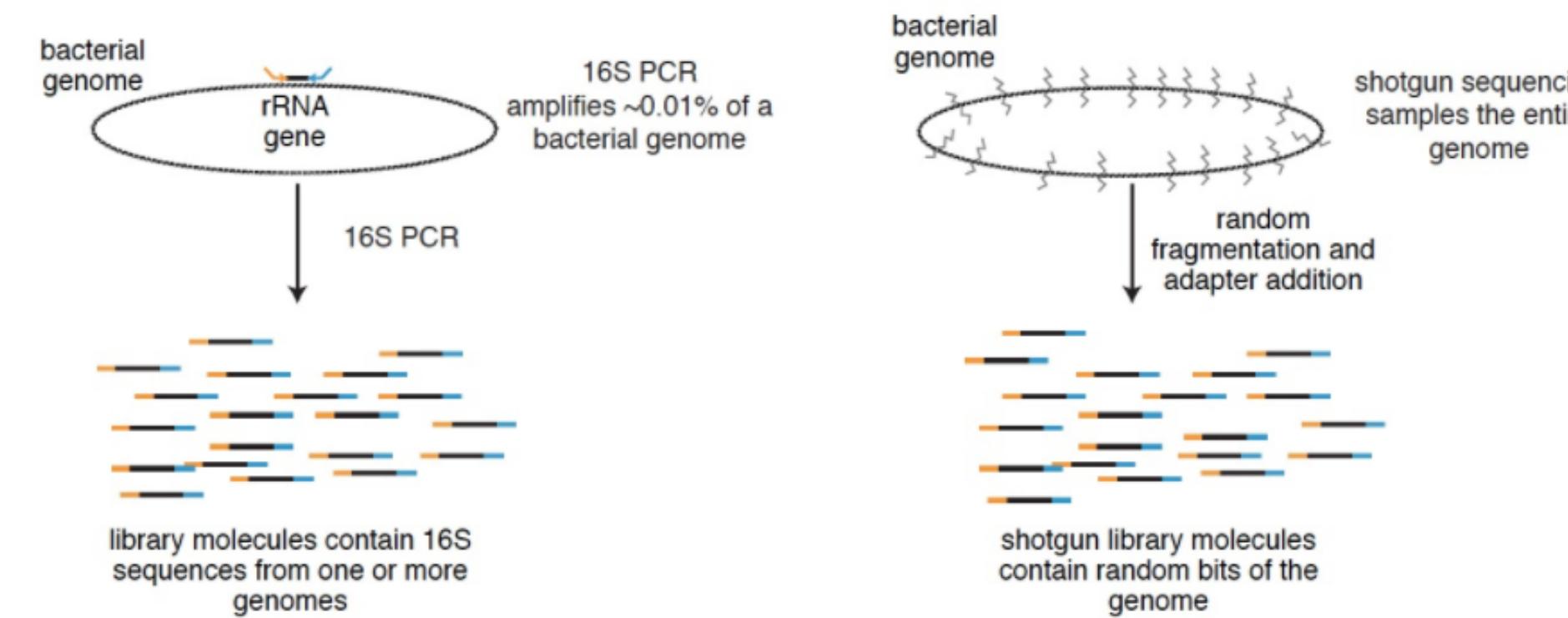


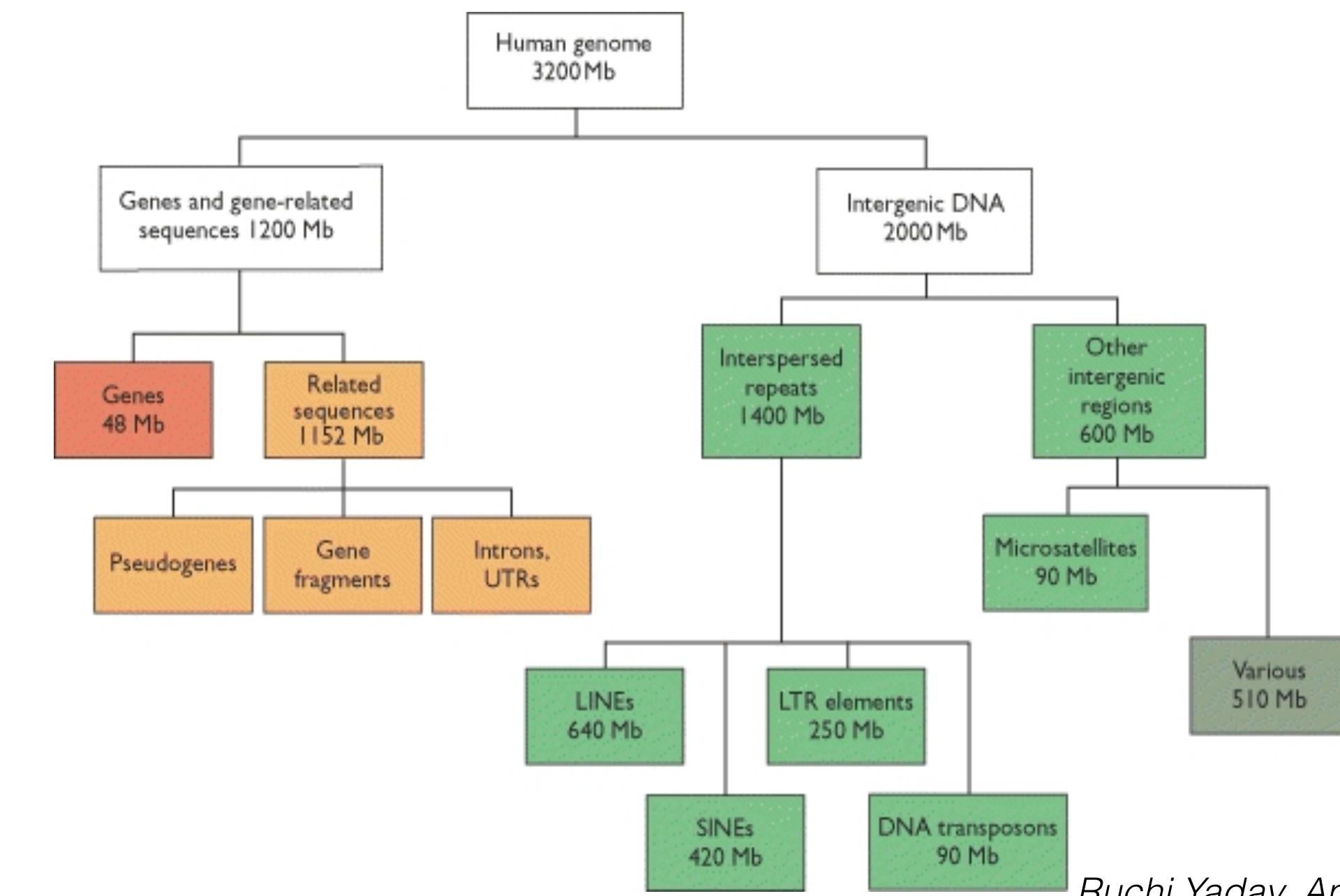
Figure: 16S vs. shotgun sequencing.

Adapter

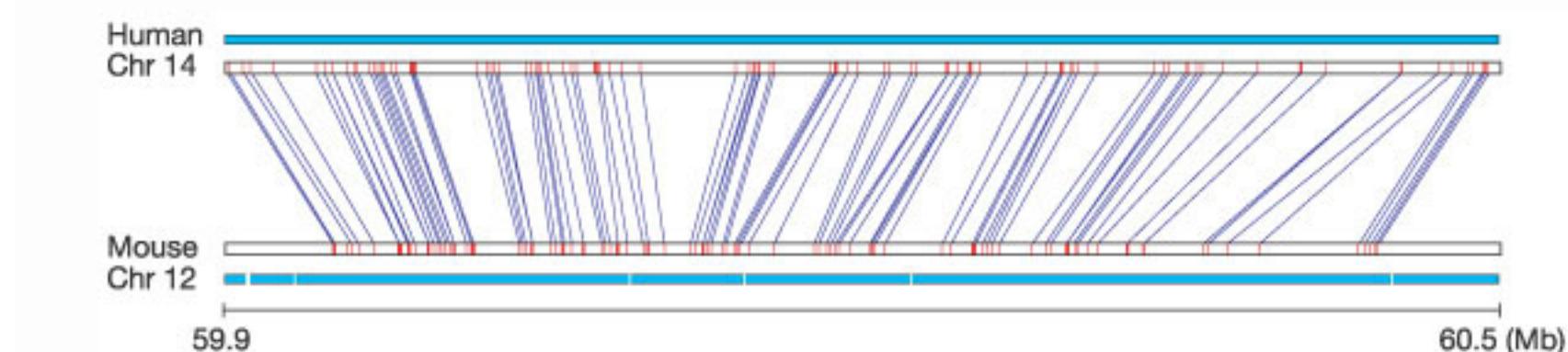
Most NGS instruments require that dsDNA of known sequence be added to the 2 ends of [library](#) molecules that will be sequenced on the instrument. Adapters can be added in a variety of ways to starting nucleic acid molecules during [library](#)

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

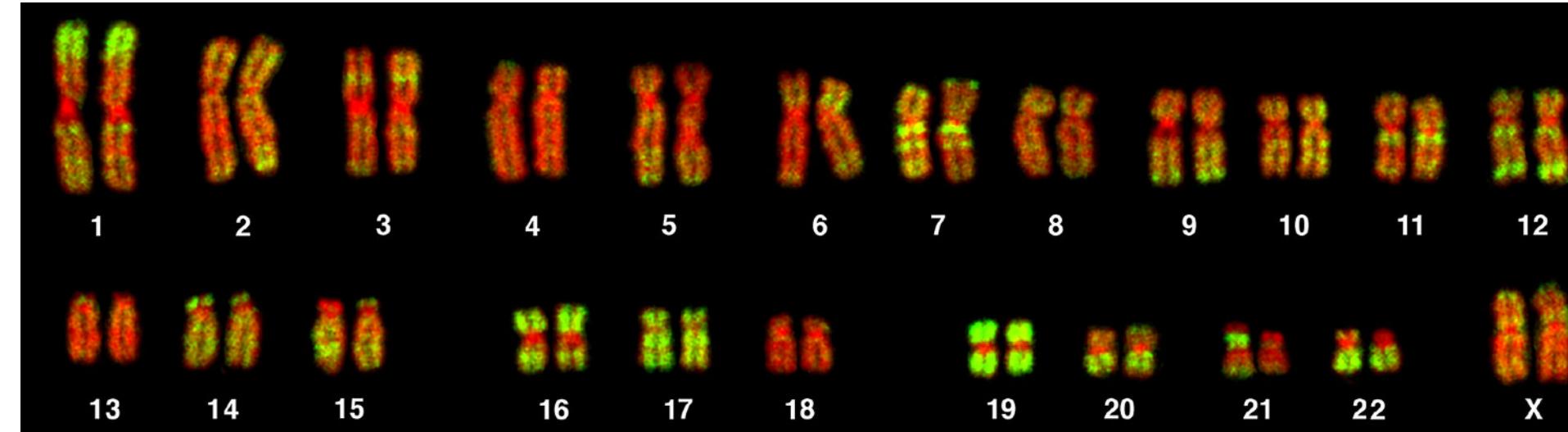


Ruchi Yadav, Amity University

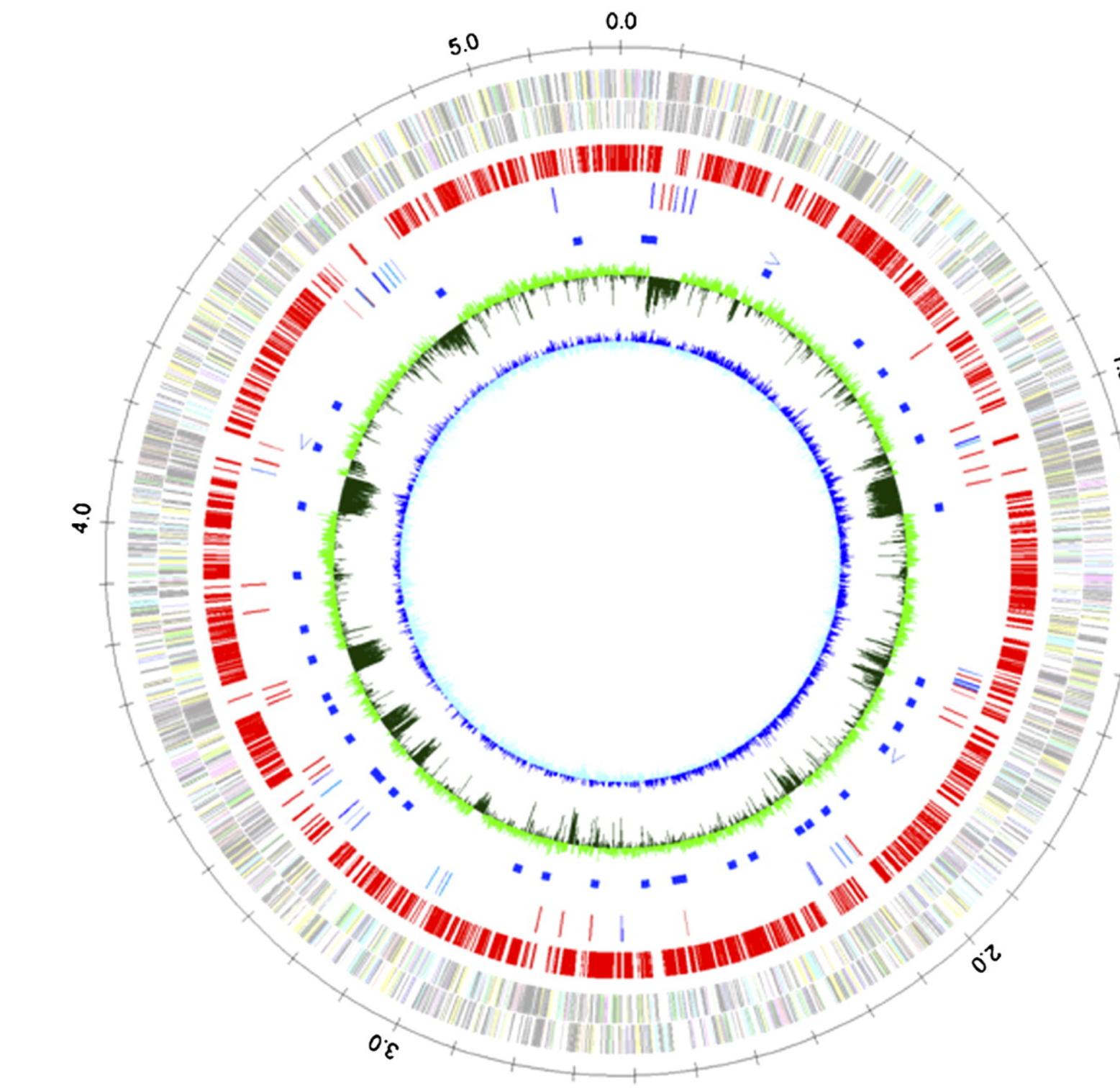


Nature (2002) Mouse Genome

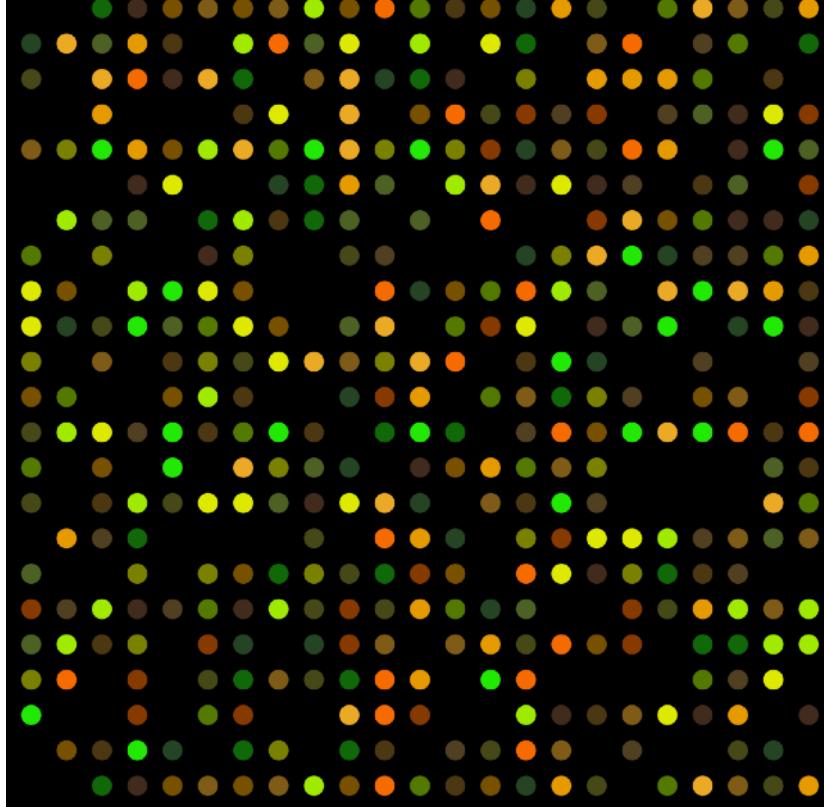
Genomics isn't the same thing as sequencing, but they're increasingly related



Bolzer et al (2005) PLoS Biol

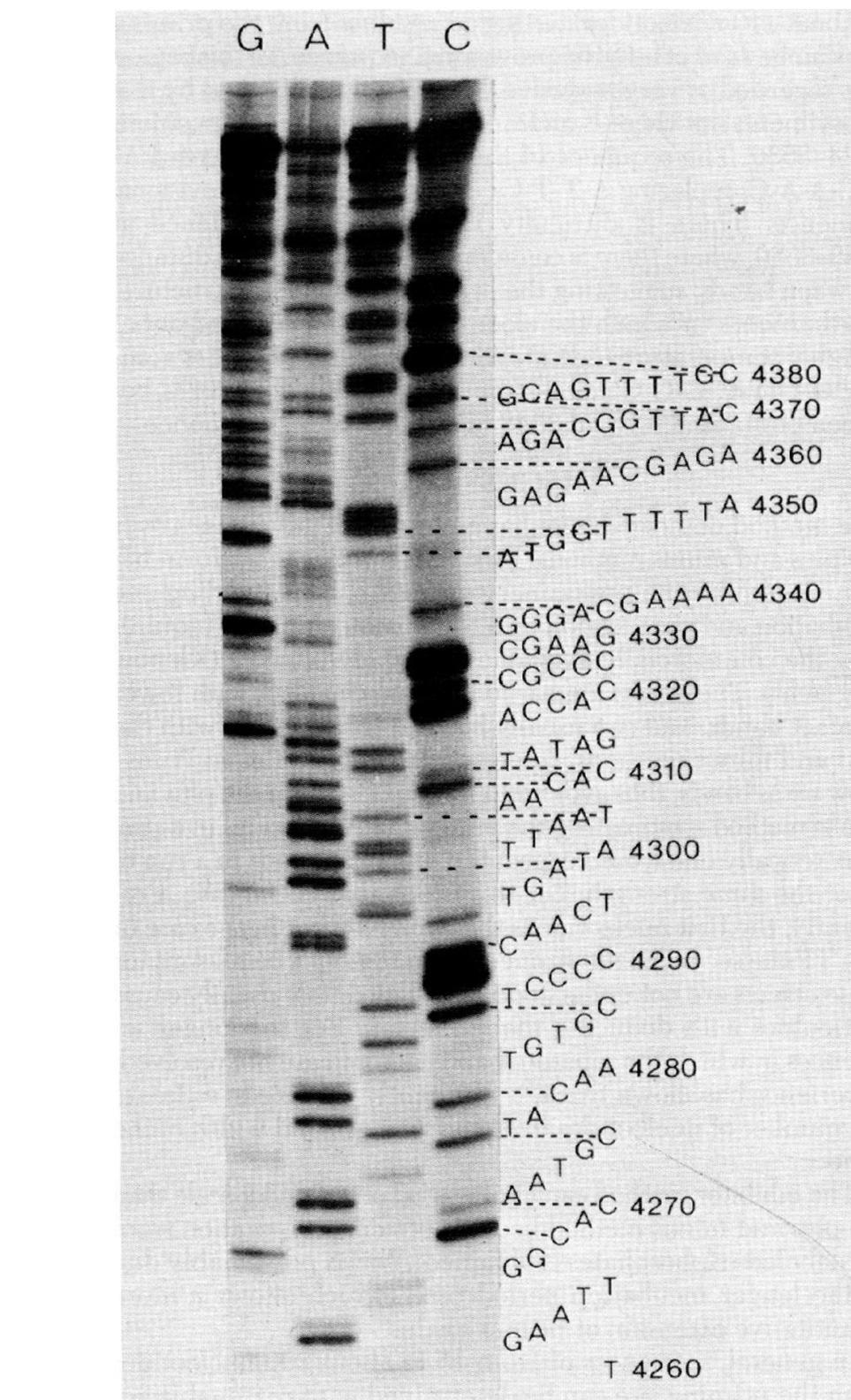
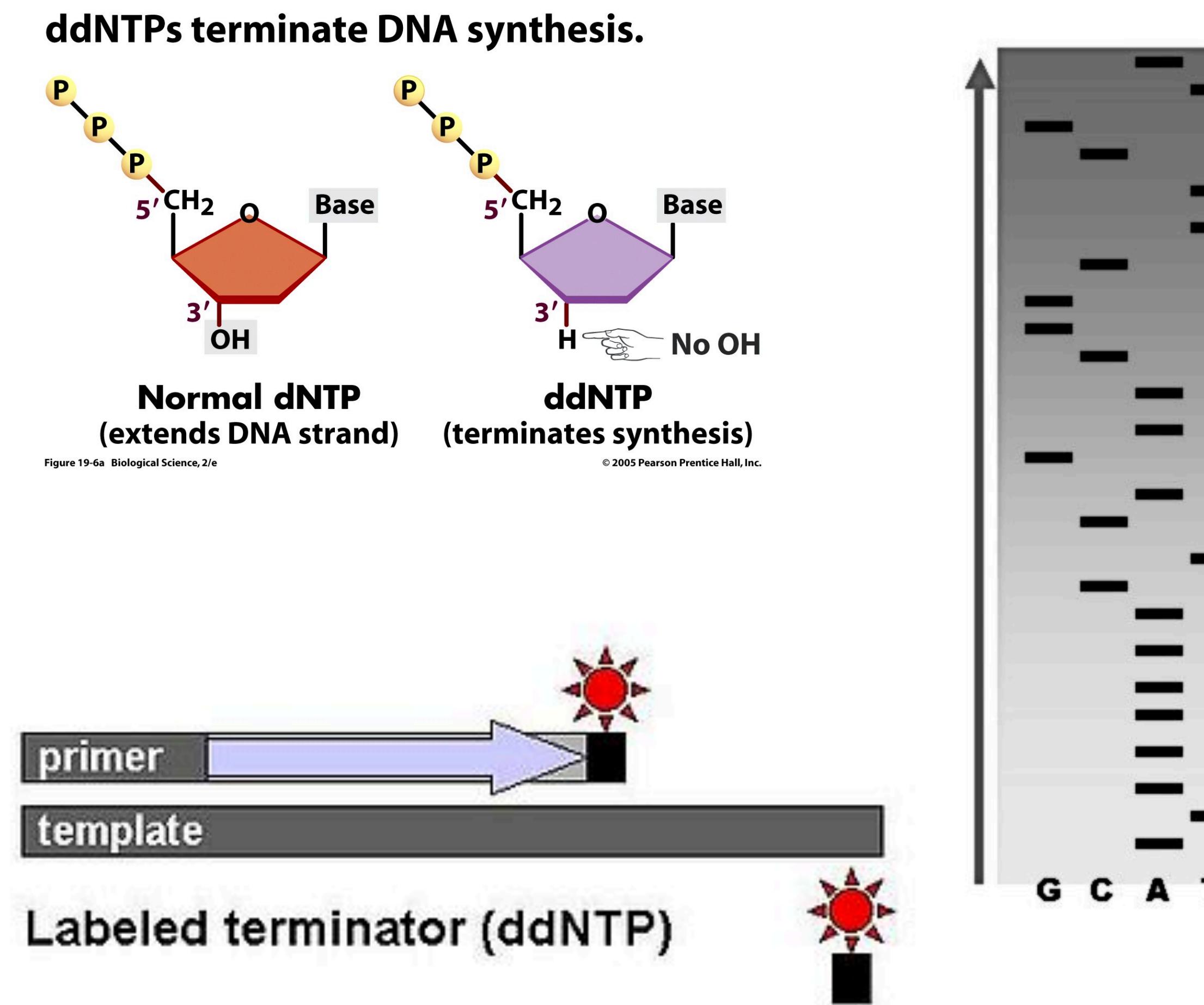


Nakazawa et al (2009) Genome Research



Wikimedia commons

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



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

Improvements to Sanger sequencing and molecular methods allowed the sequencing of increasingly large genomes

1965 – First nucleic acid sequenced: Yeast trnA

1976 – First complete genome sequenced (RNA virus: bacteriophage MS2)

1977 – Maxam-Gilbert and Sanger DNA sequencing methods introduced and first complete DNA genome (Phage Φ -X174)

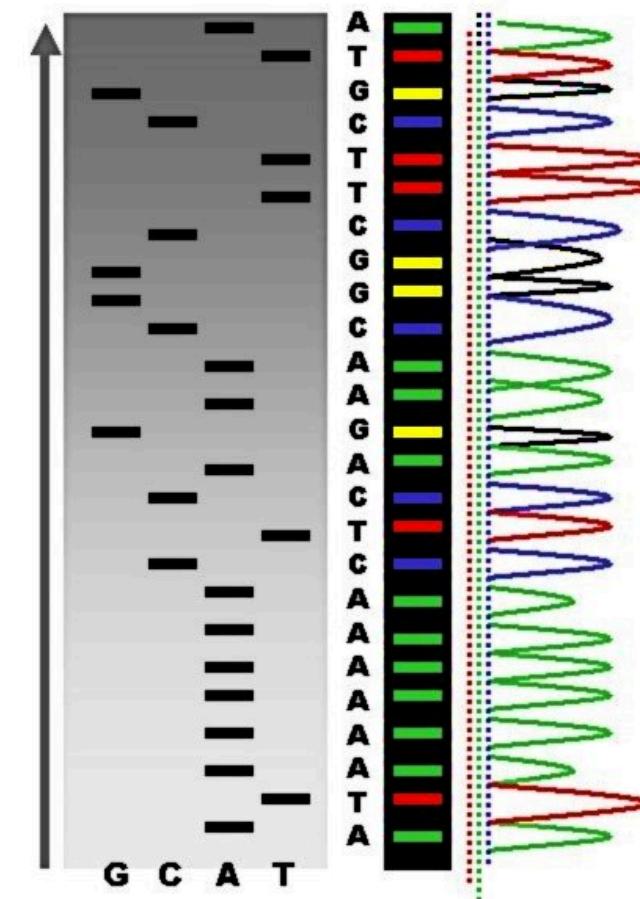
1983 – PCR introduced

1995 – First complete cellular genome (*Haemophilus influenzae*) and eukaryotic genome (yeast) sequenced

2001 – Publication of the first sequenced human genomes

2005 – Introduction of 454 Sequencing and the NGS Revolution

2005 – present: Rapid evolution of NGS technology



Slide courtesy Dan Sloan.

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

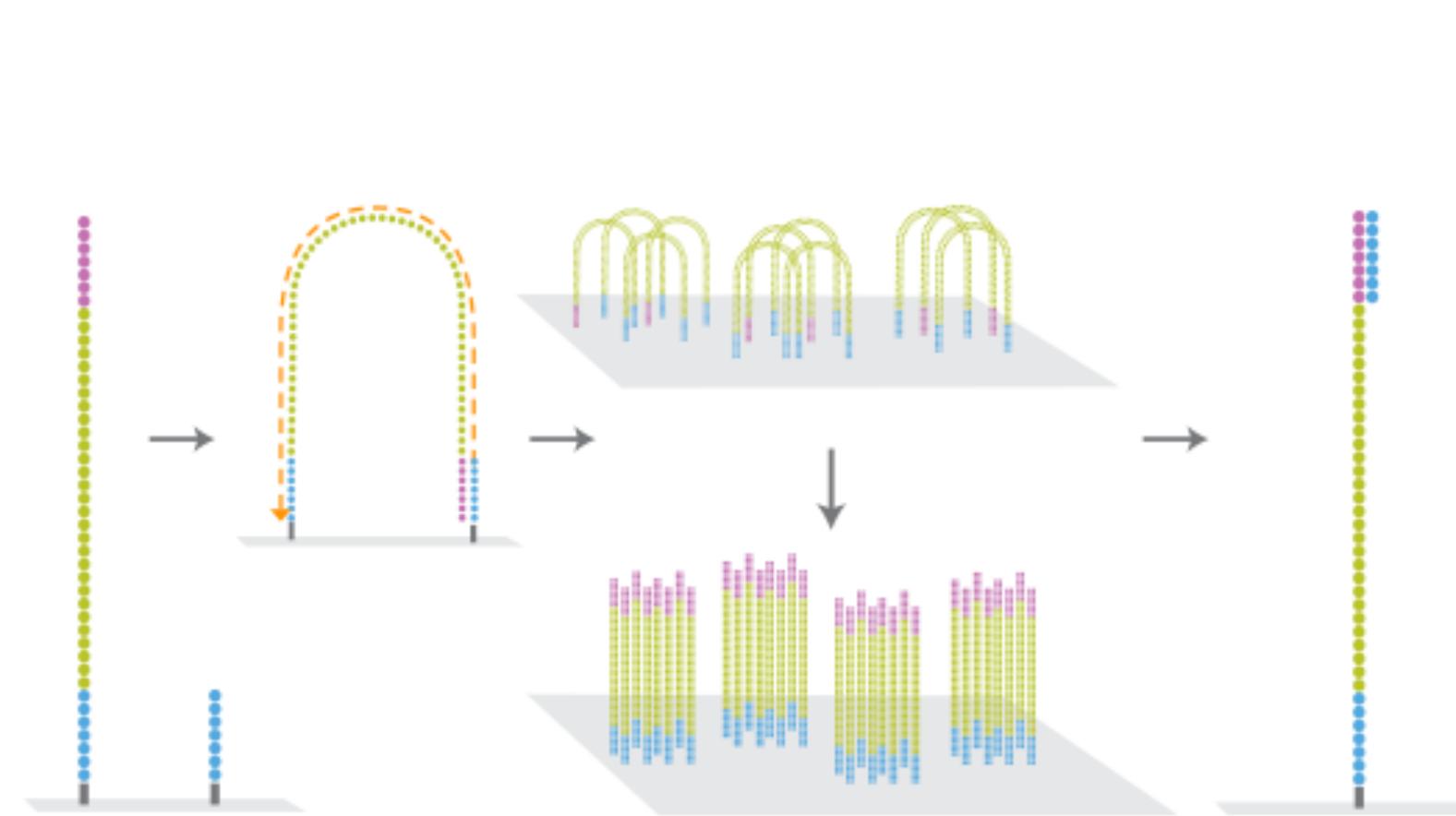
Oxford Nanopore MinION



PacBio RS-II

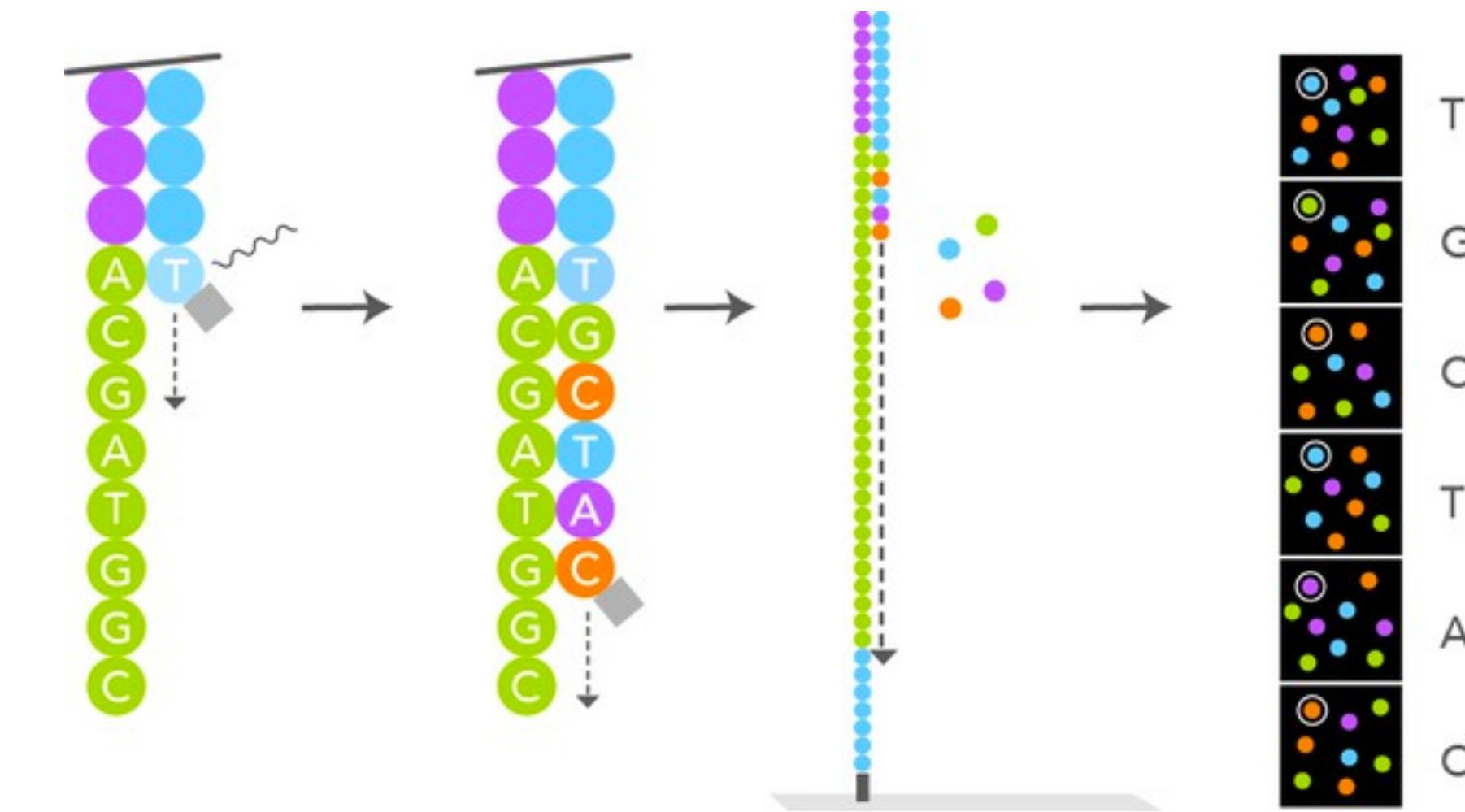


Illumina instruments use sequencing by synthesis (SBS)



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

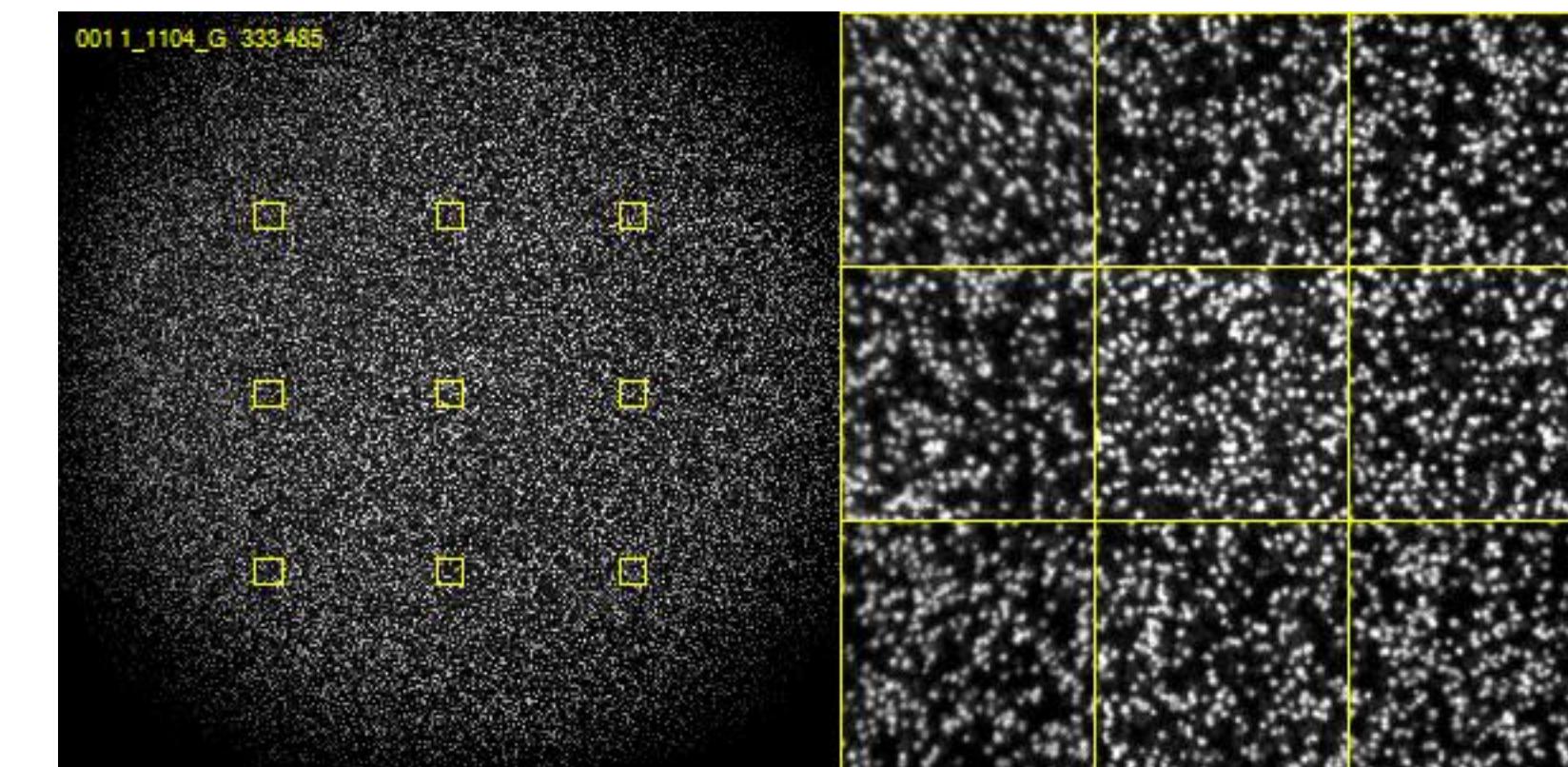
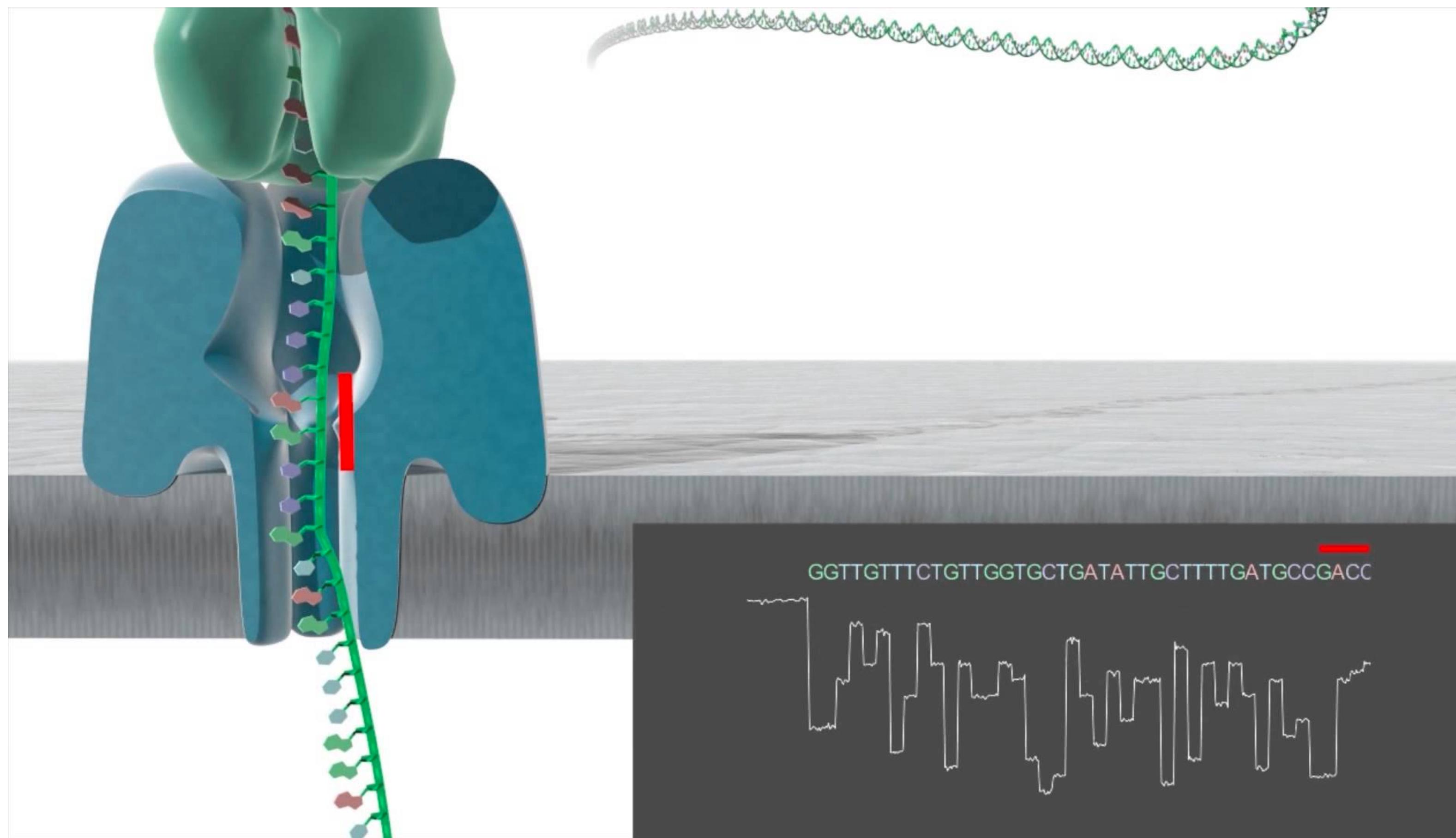


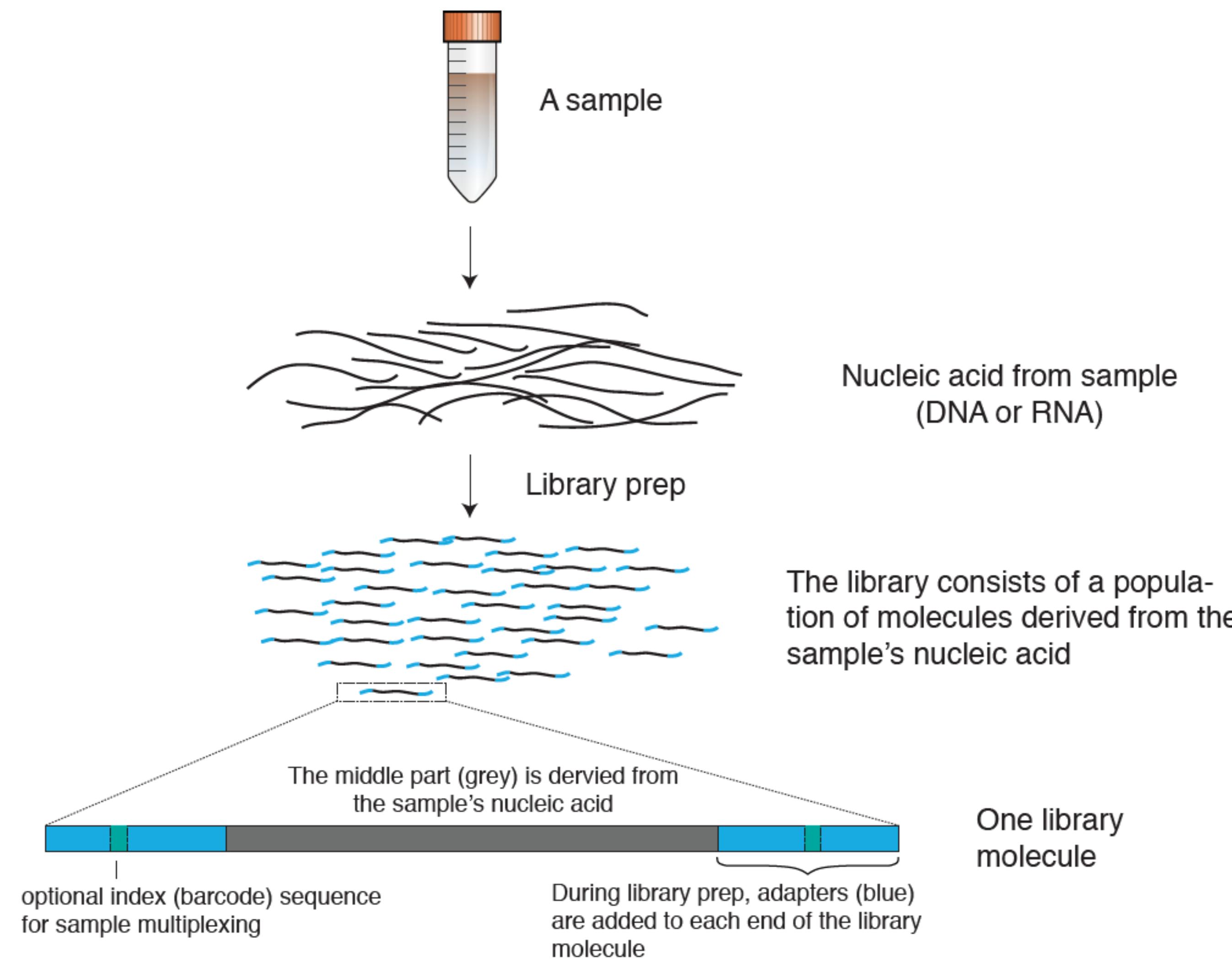
Image credit: Illumina

Long read sequencers sequence single molecules

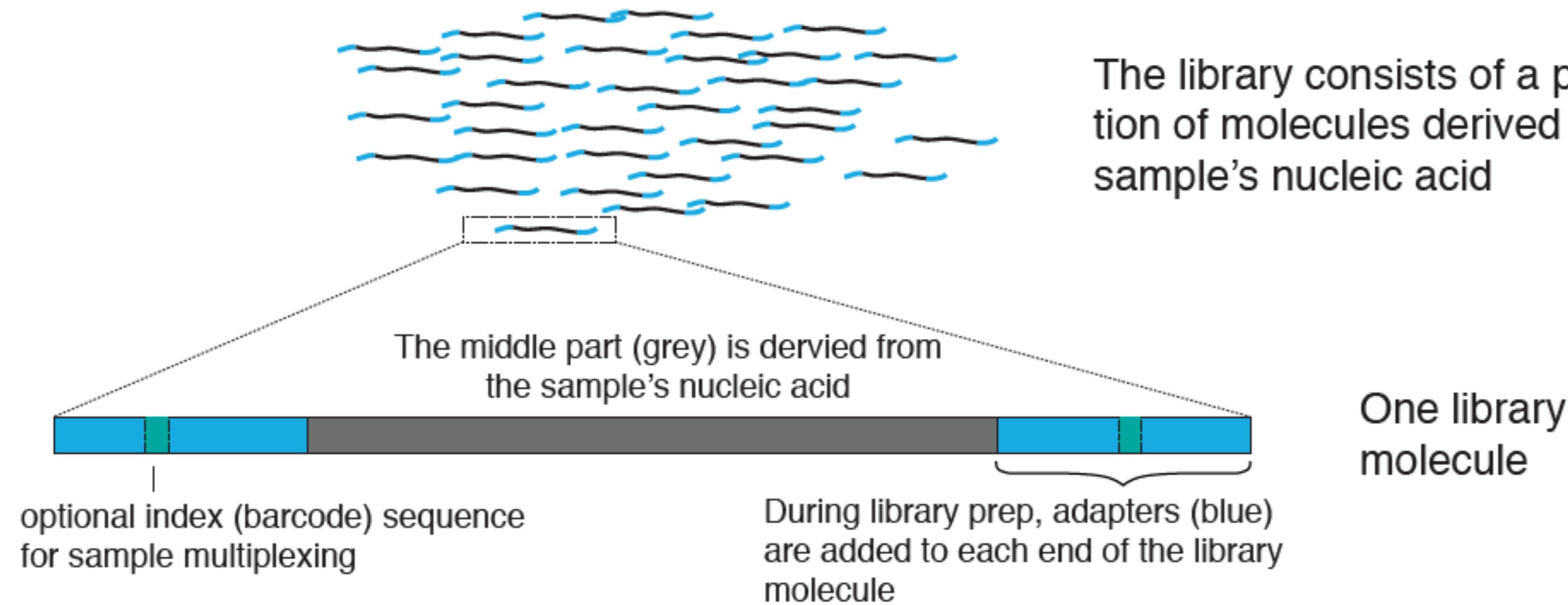


Much longer reads, but with much higher error rates

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

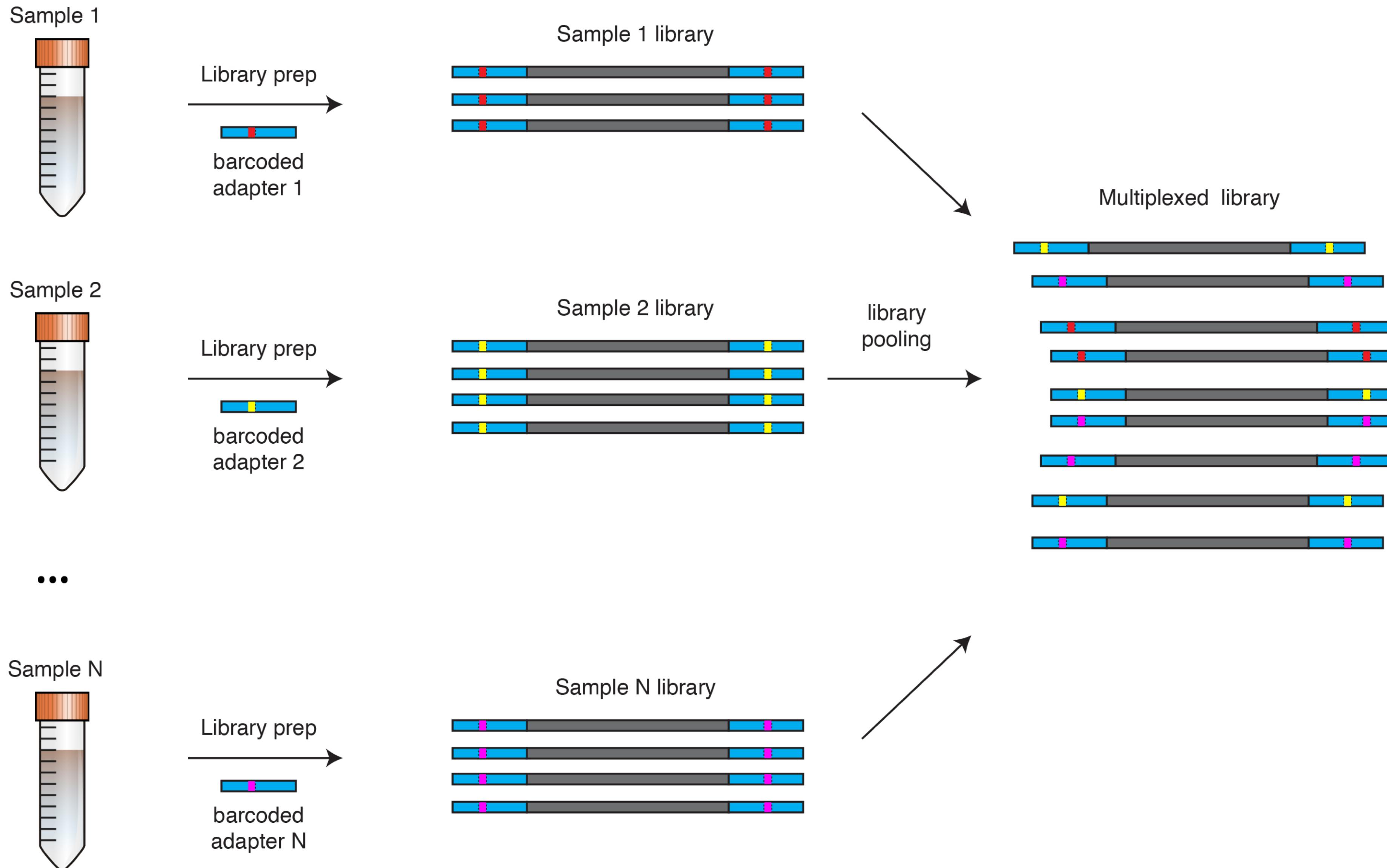


Library prep converts nucleic acids into a form suitable to be sequenced



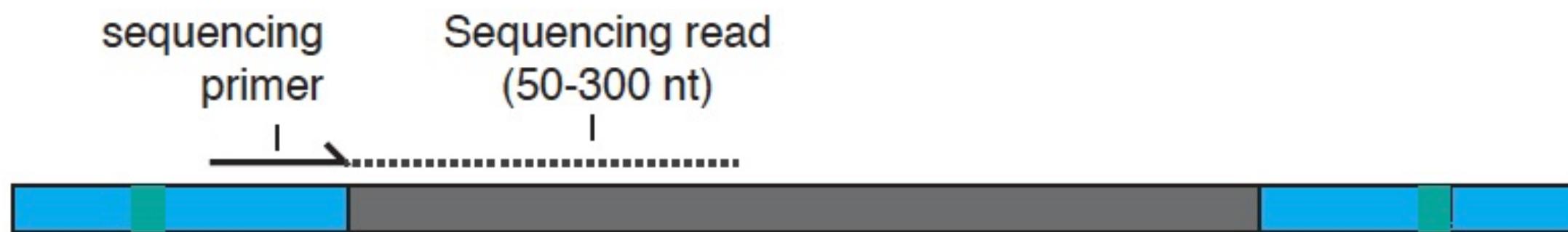
An example Illumina library molecule

Barcodes (also called indexes) allow sample multiplexing



Illumina sequencing produces 1-4 reads per library molecule

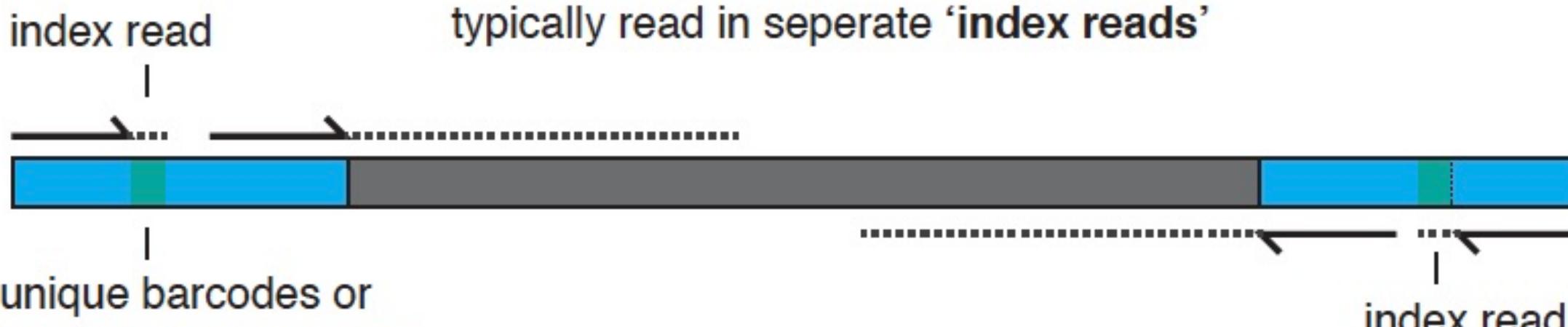
In **single end sequencing**, a library molecule is sequenced from one end



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'



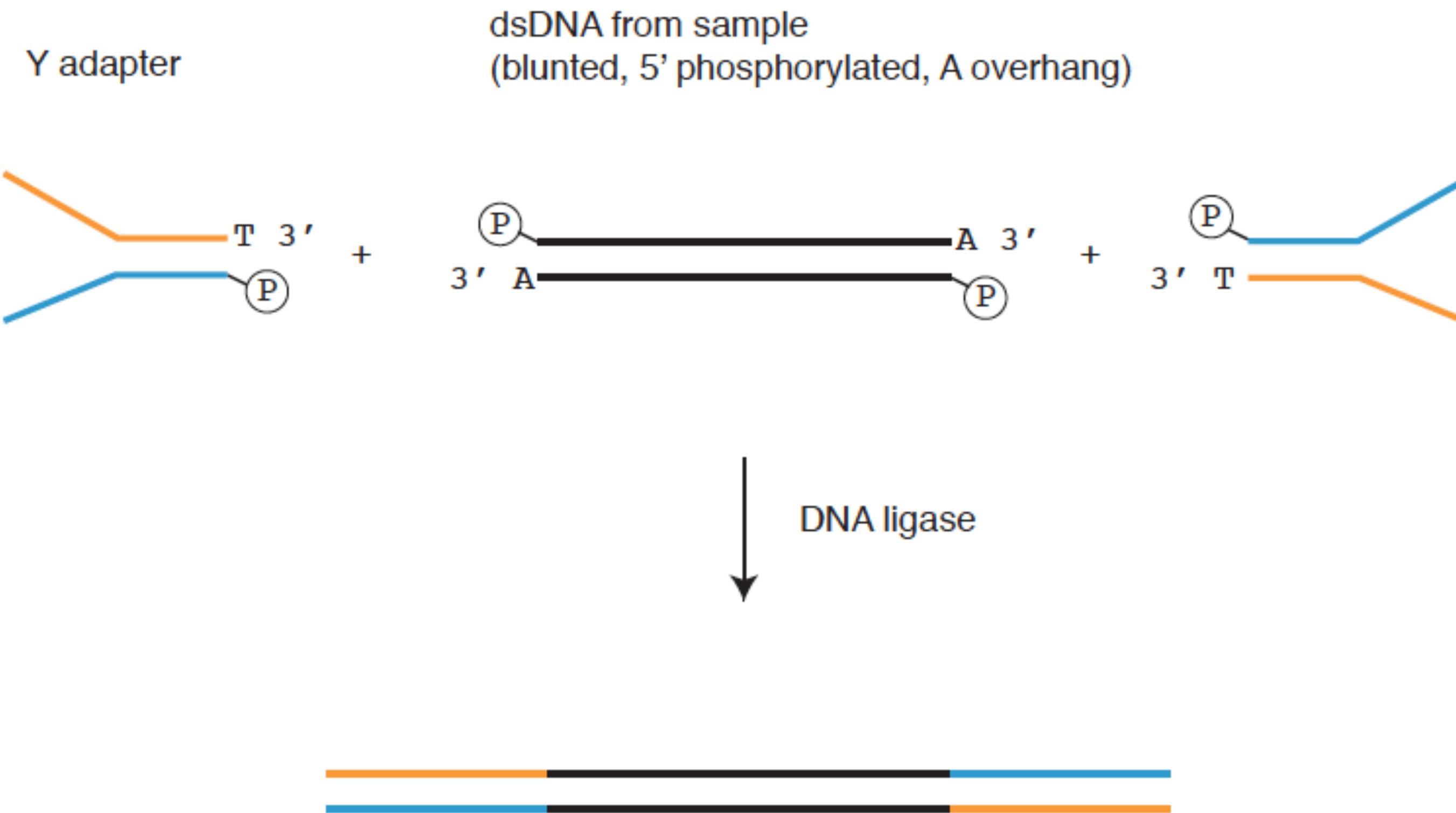
unique barcodes or barcode pairs can be used to differentiate multiplexed samples

index read

Reads are sub-sequences of the starting nucleic acid that often contain errors



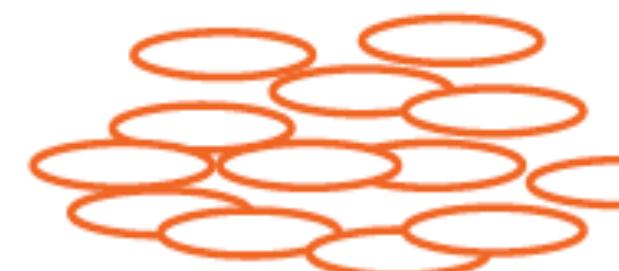
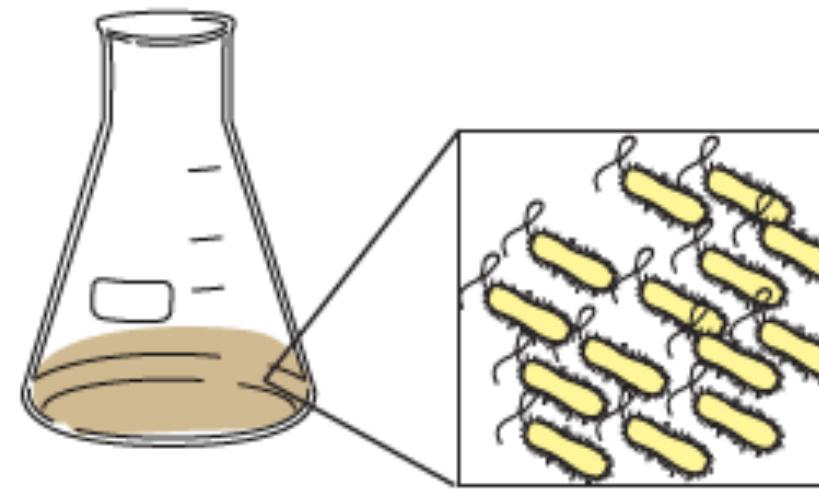
There are many good ways to make sequencing libraries



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

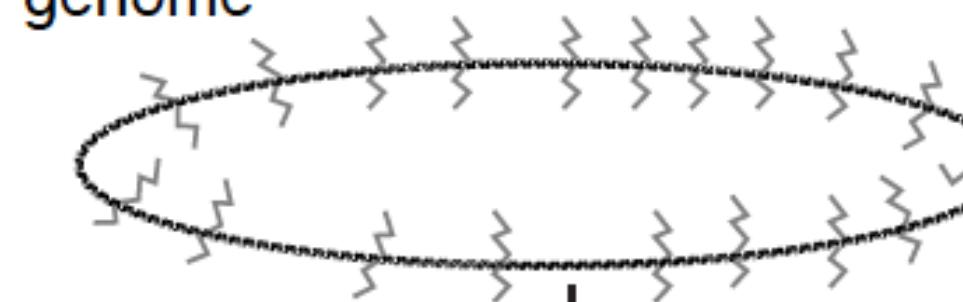
For instance, if you make a 'shotgun library' from a single organism, you're doing
whole genome sequencing (WGS)

bacterial isolate



bacterial genome

bacterial genome



random fragmentation and adapter addition

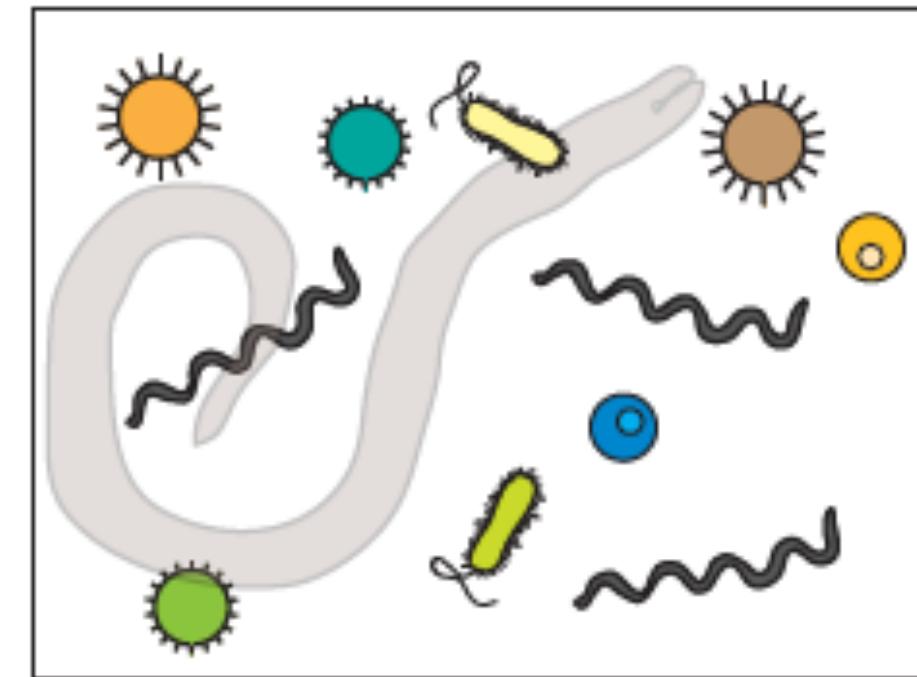


shotgun library molecules contain random bits of the genome

shotgun sequencing samples the entire genome

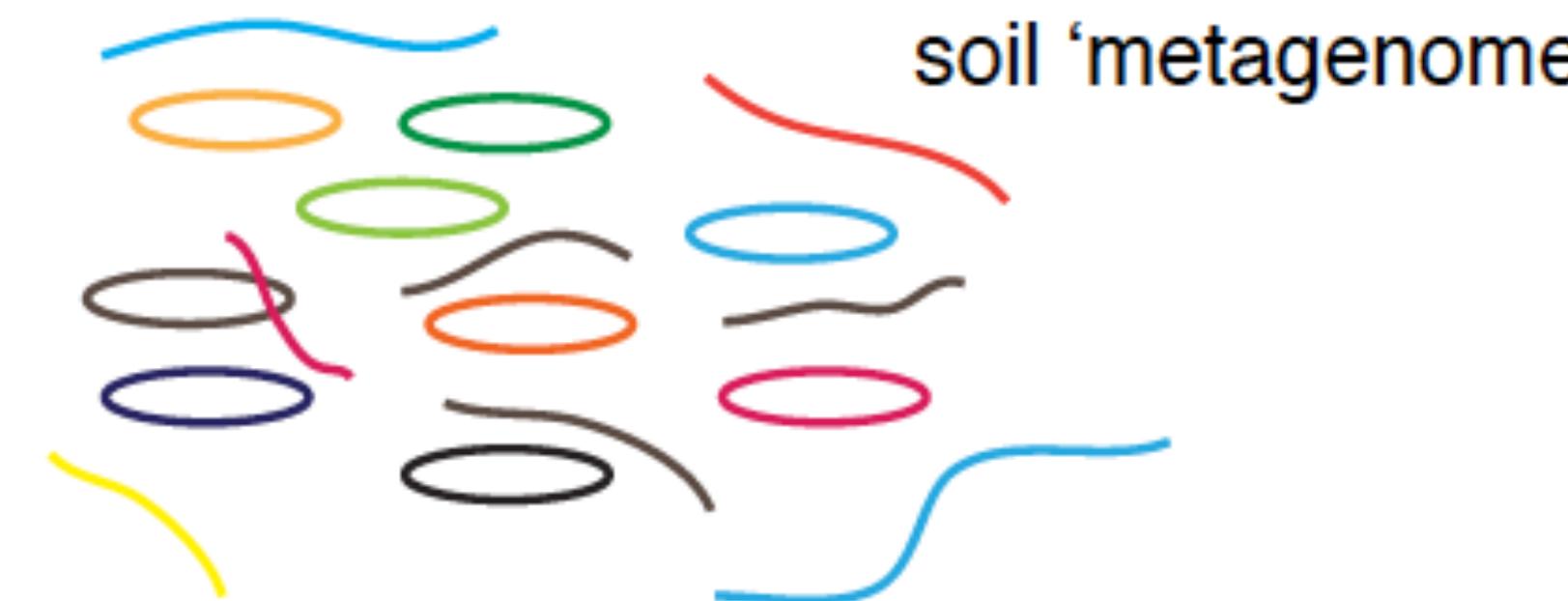
Metagenomic sequencing involves sequencing of genomes from more than one organism

soil community



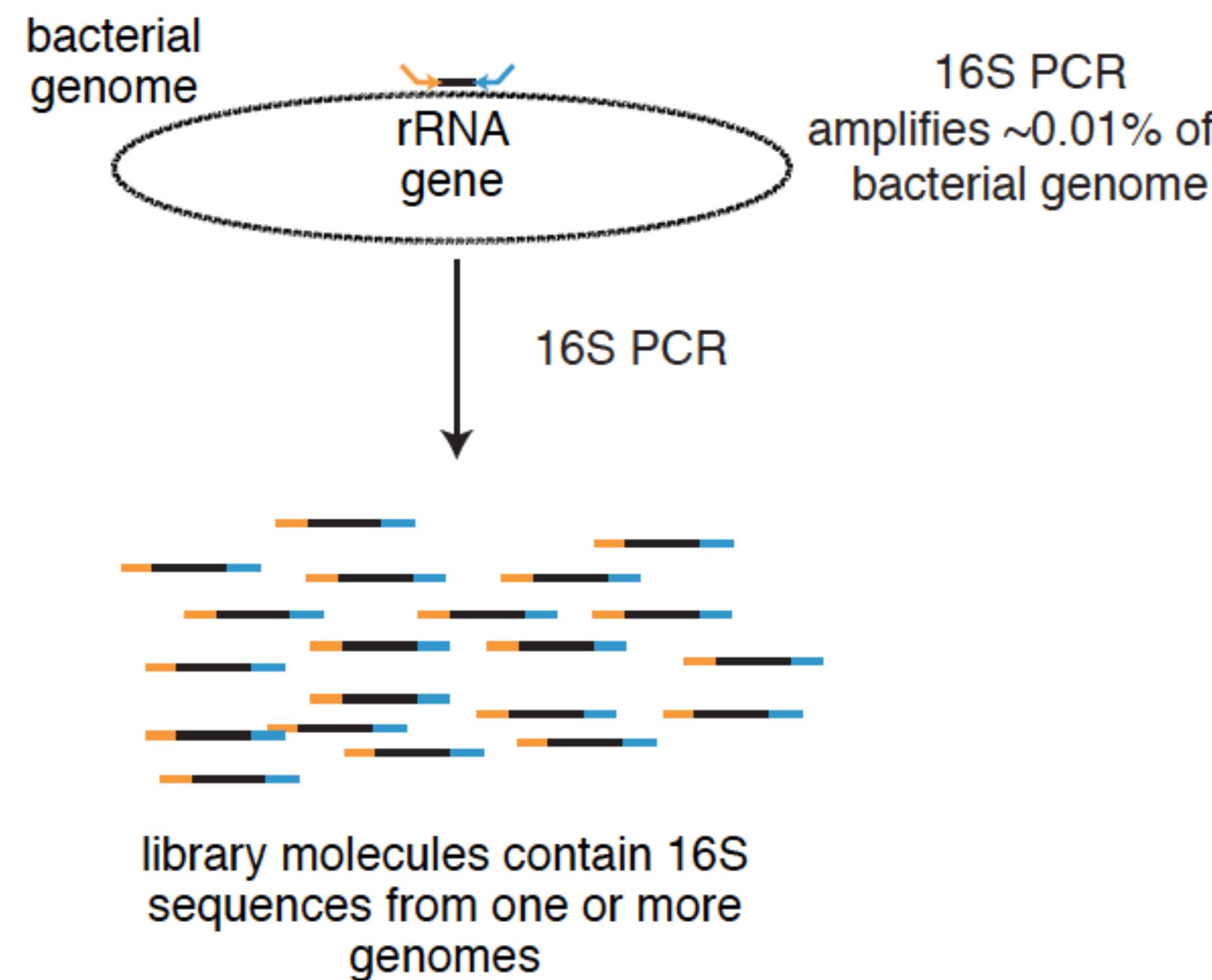
Could make a 16S or a shotgun library from these genomes

Sequencing of RNAs from a complex sample like this is metatranscriptomics

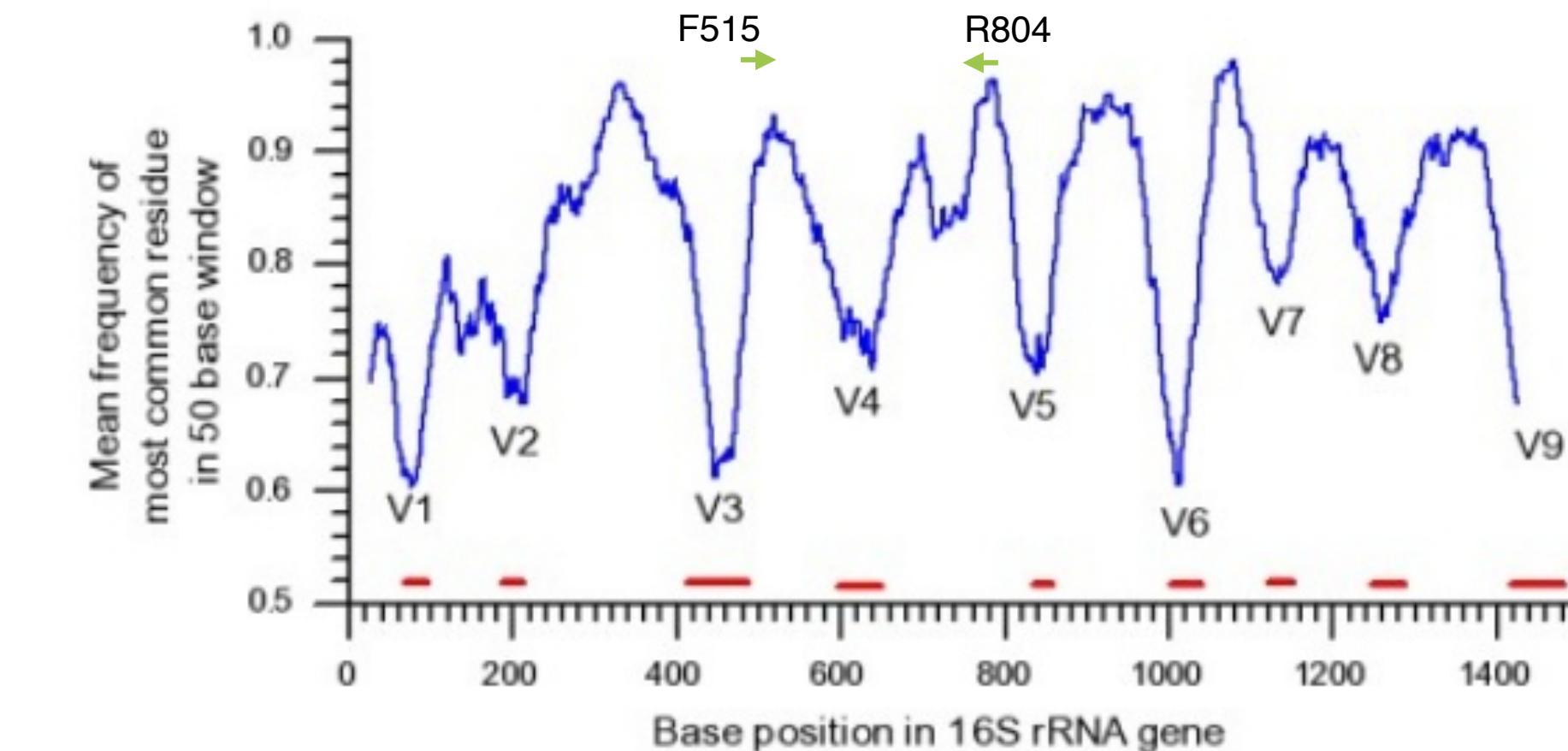


Microbiome sequencing often means **16S rRNA** sequencing

16S sequencing is one type of
'amplicon sequencing'

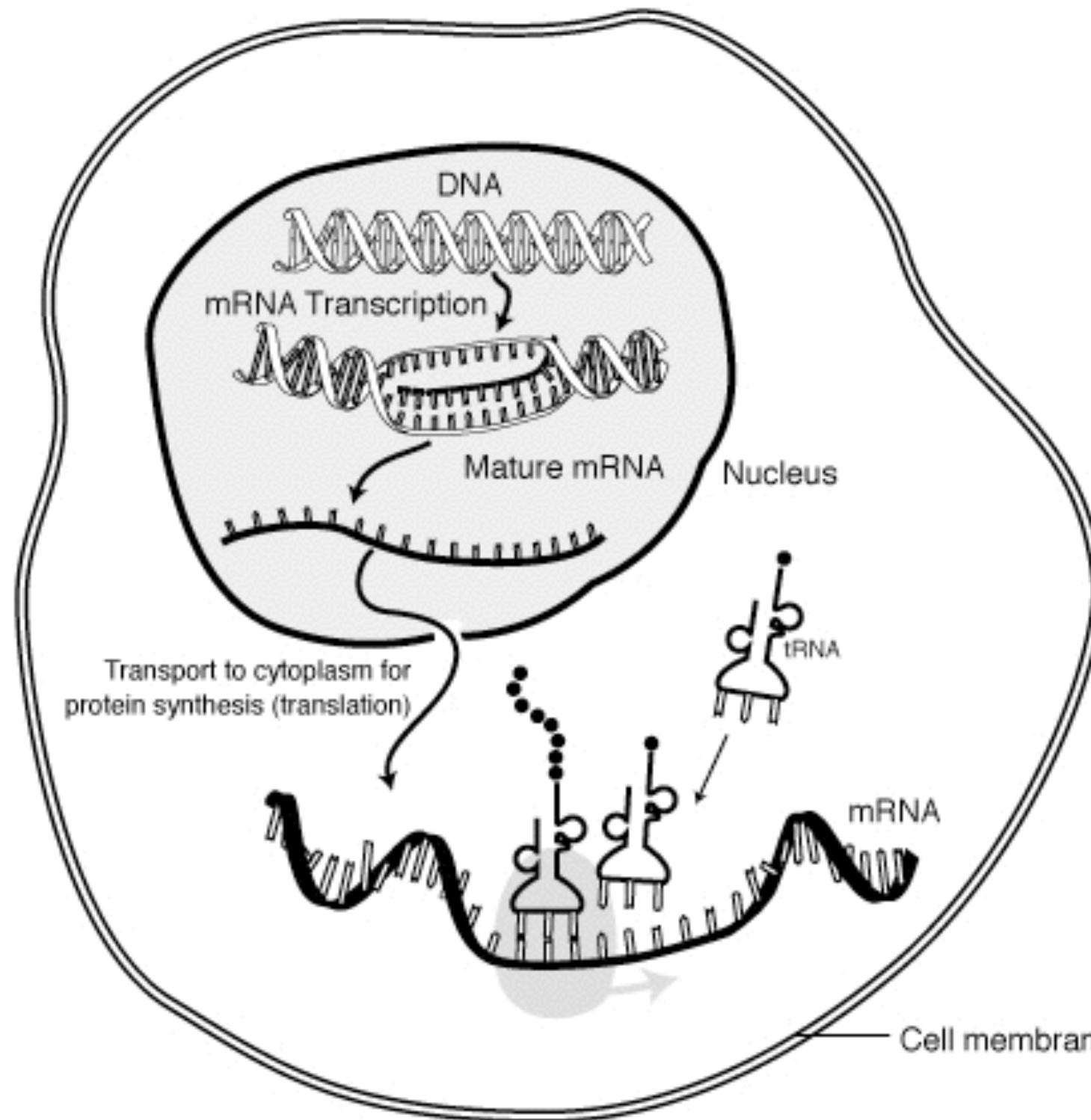


16S rRNA genes have highly conserved regions flanking variable regions



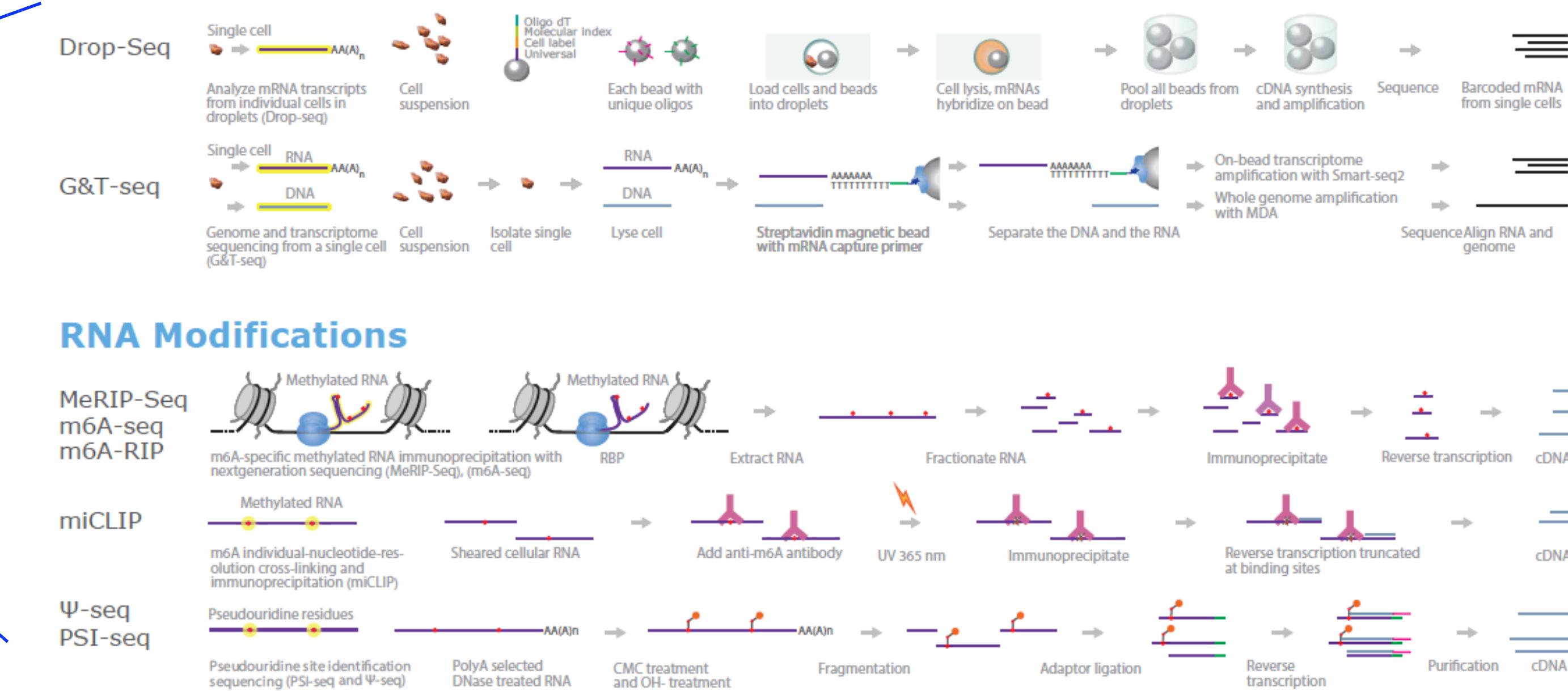
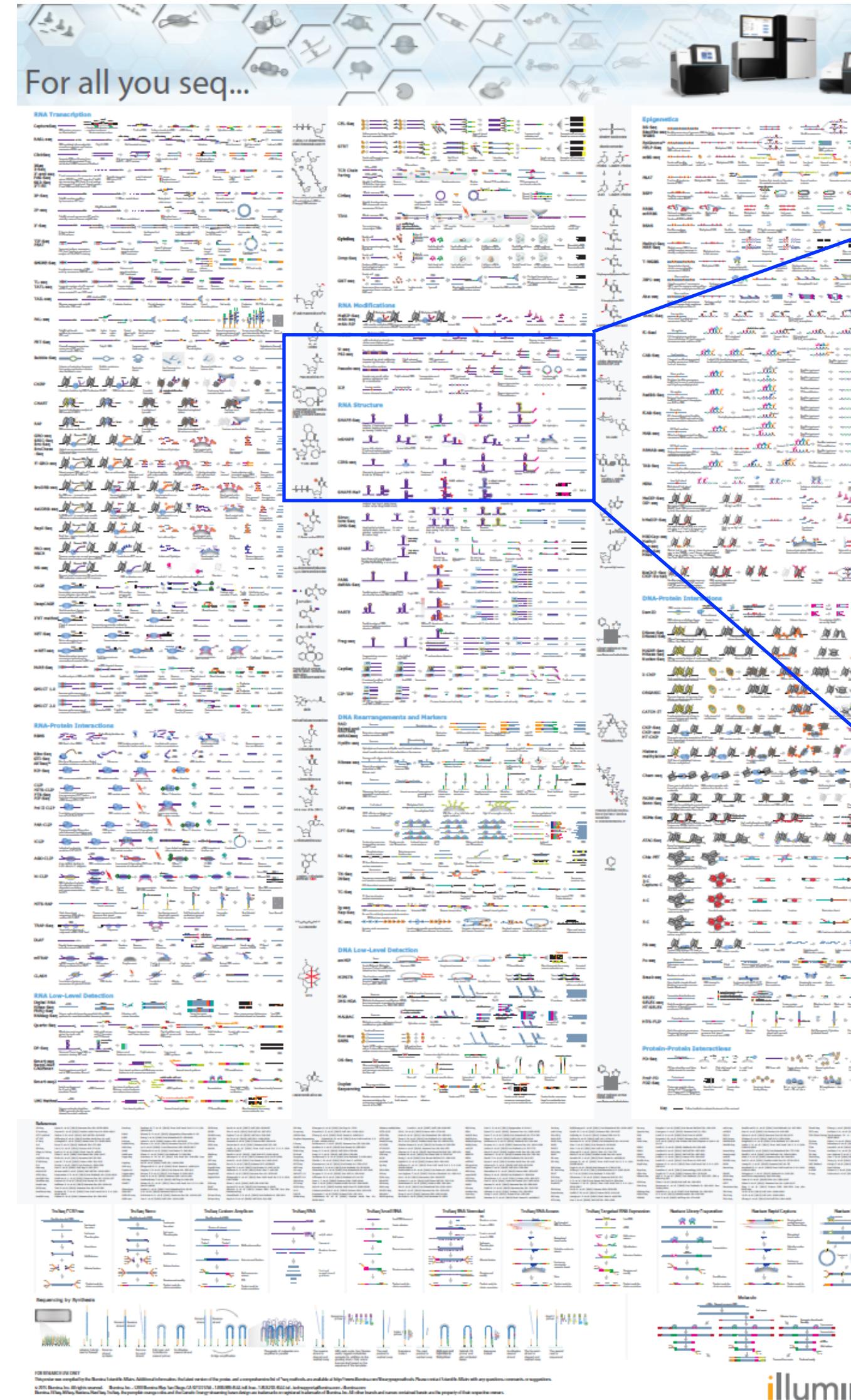
How you make a library determines what type of sequencing you're doing

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



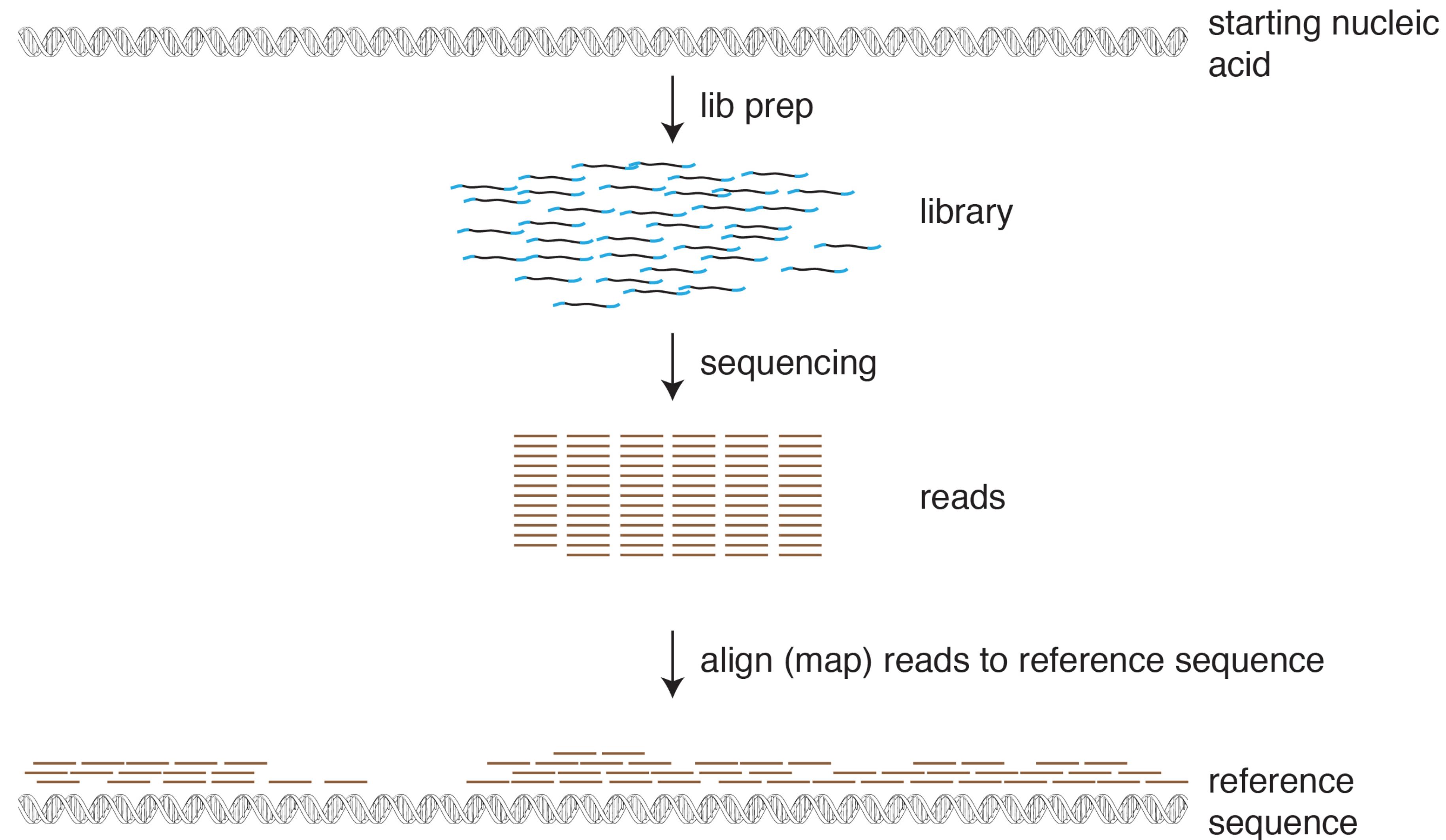
The abundance of reads from a particular mRNA is proportional to that mRNA's abundance in the cell

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

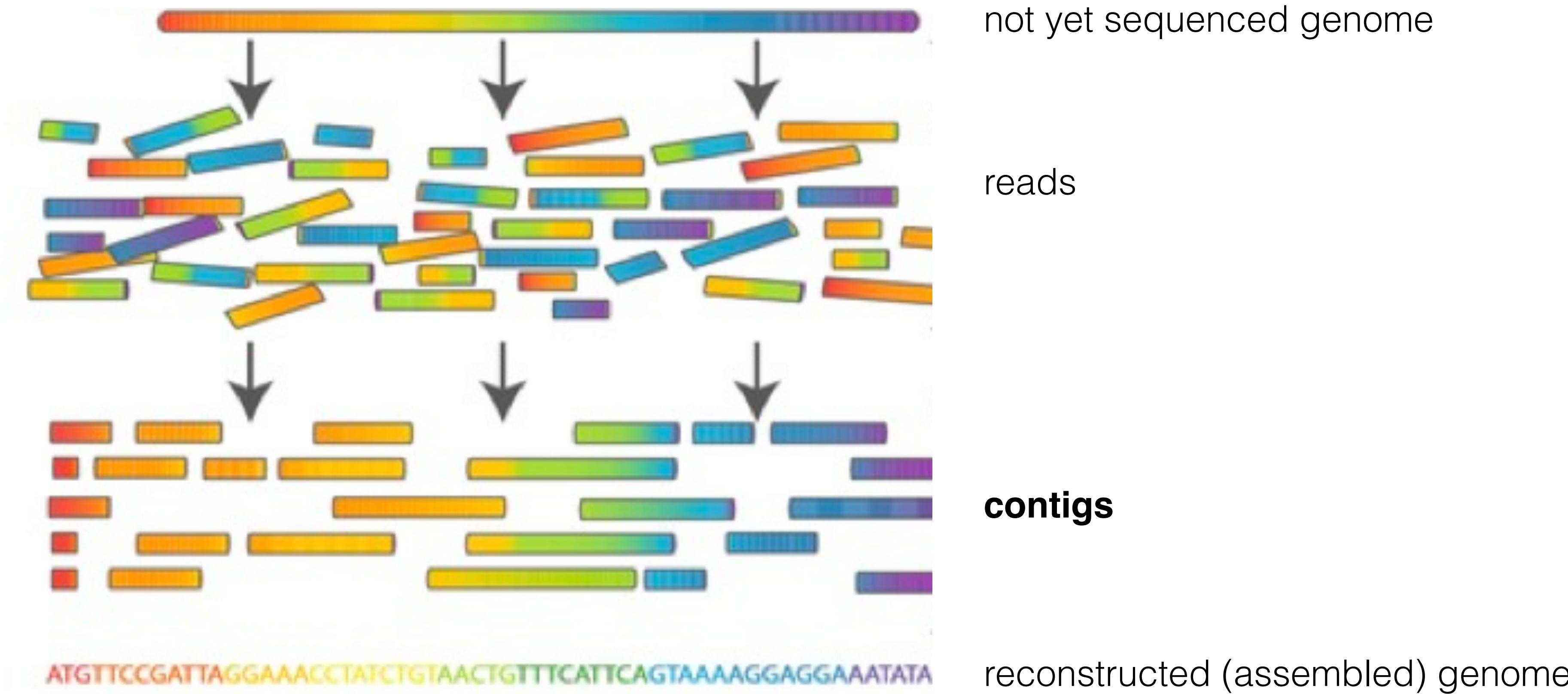


They're all variations on a few themes. Don't let it overwhelm you. Most sequencing is of a few simple types, and it's better to focus on the Biology and experimental design.

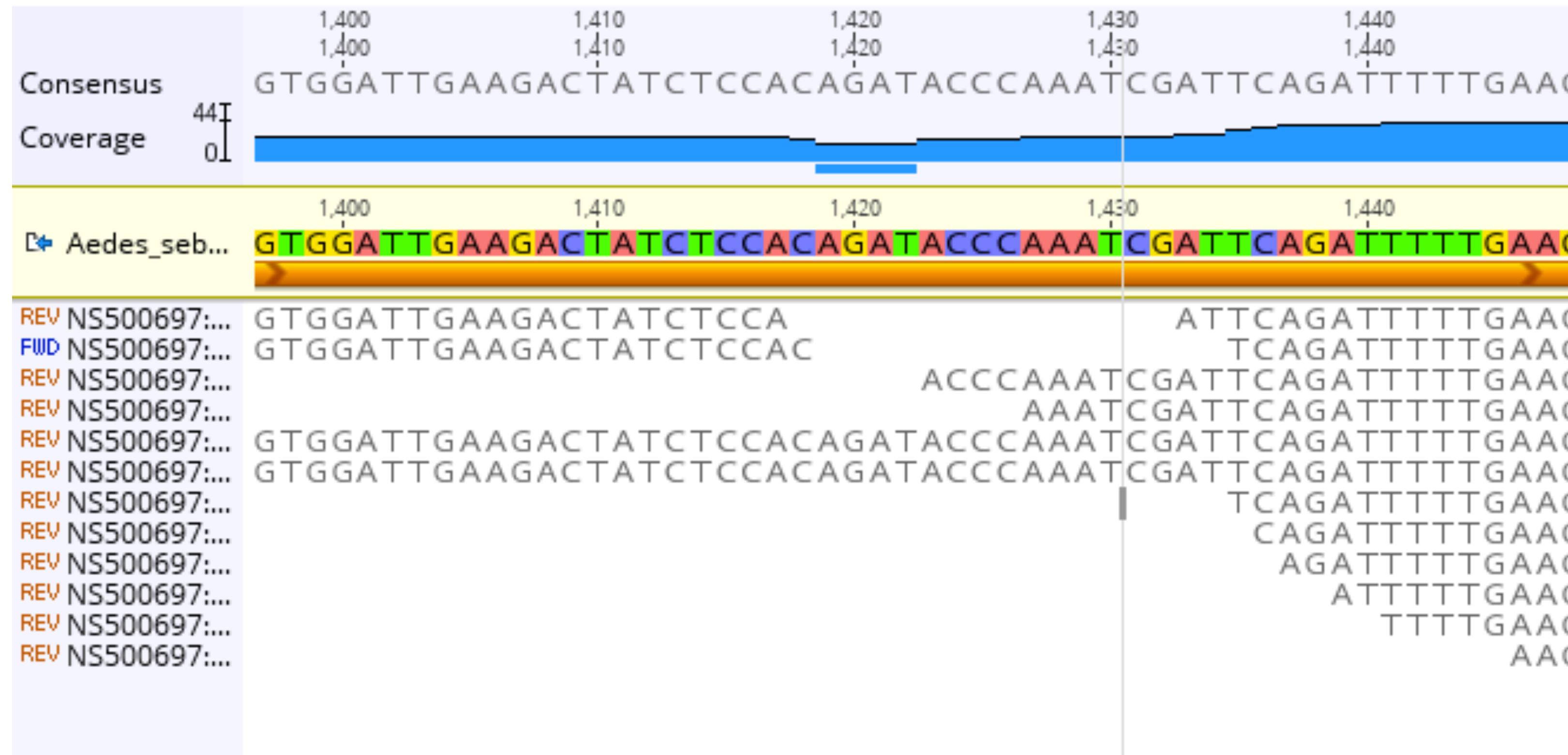
Mapping is the process by which sequencing reads are aligned to the region of a genome from which they derive.



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



Coverage is the number of individual reads that support a particular nucleotide in an assembled (reconstructed) sequence or that align to a particular nucleotide in a reference sequence



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

Questions?

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