An overview of genomics and sequencing terminology and practices

Mark Stenglein, GDW

Genomics of Disease in Wildlife

Math undergrad → 7 years as a software engineer → PhD in mol. biology / biochem. → Postdoc using microarrays, NGS, and bioinformatics → Assoc. Professor at CSU

1999, Bankgok, Thai Airways test facility

Department of MICROBIOLOGY, IMMUNOLOGY & PATHOLOGY
COLLEGE OF VETERINARY MEDICINE & BIOMEDICAL SCIENCES

CENTER FOR VECTOR-BORNE INFECTIOUS DISEASES

UCSF 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

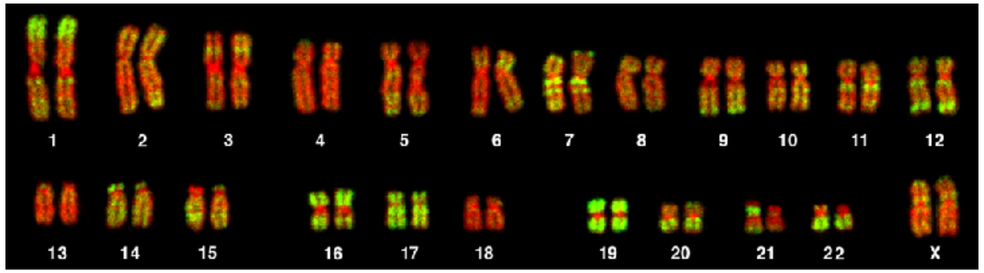
The jargon and terminology association with genomics and ‘next gen’ sequencing can be confusing and intimidating

Word cloud: Illumina

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

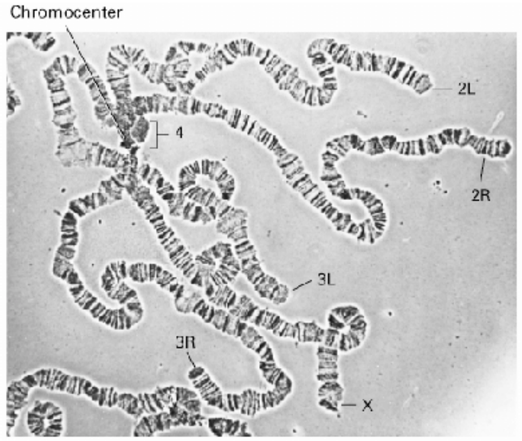
Non-sequencing genomic techniques

FISH

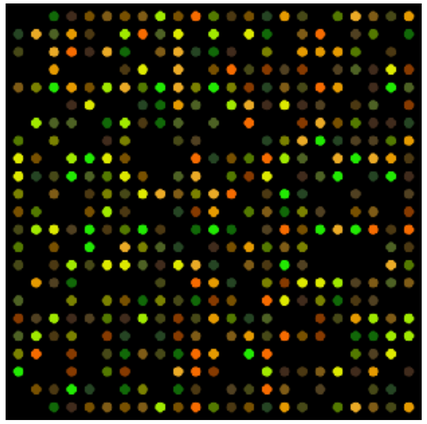


Bolzer et al (2005) PLoS Biol

Polytene chromosomes

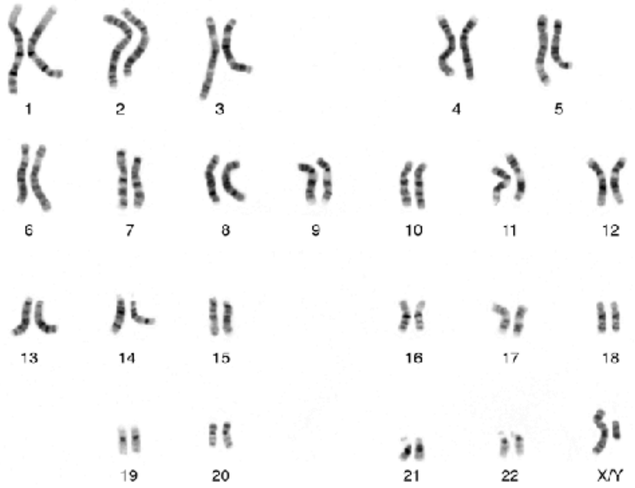


Microarray

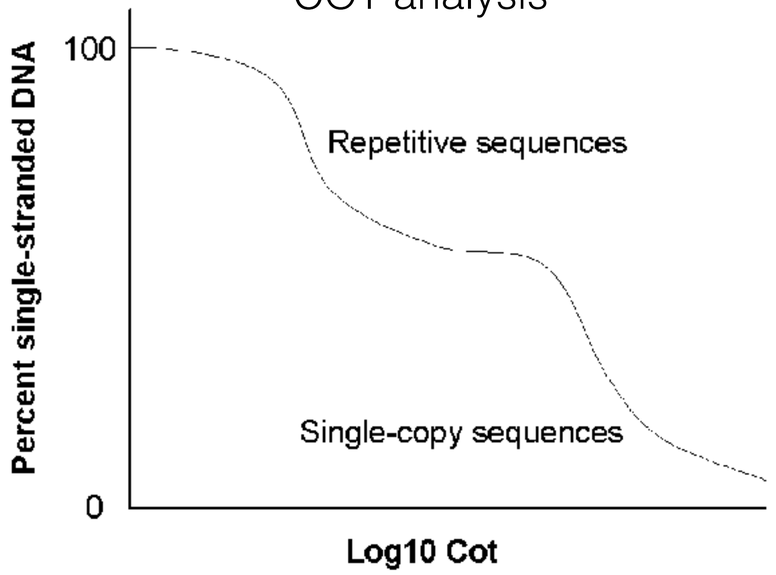


Wikimedia commons

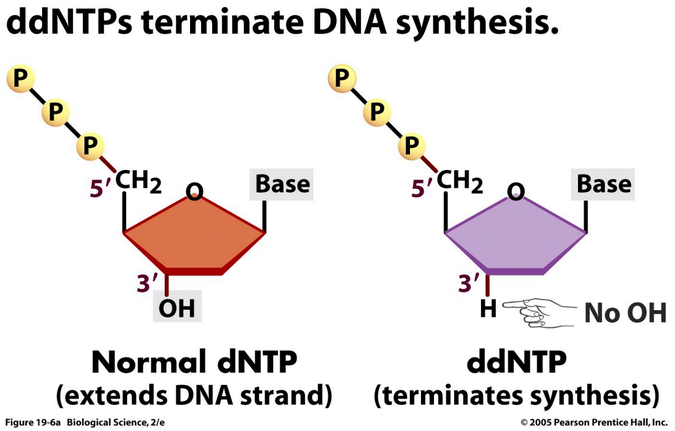
Karyotype

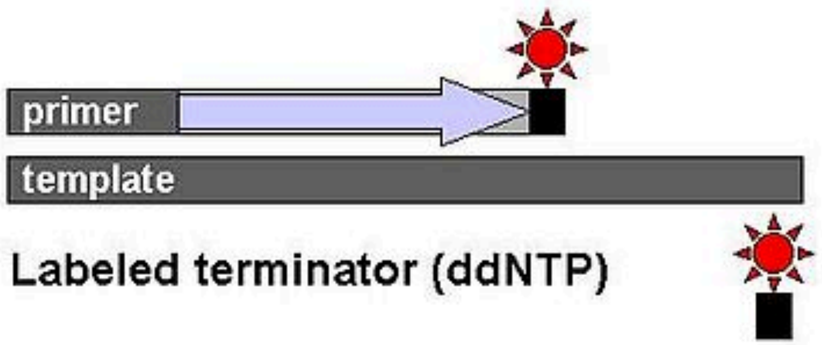


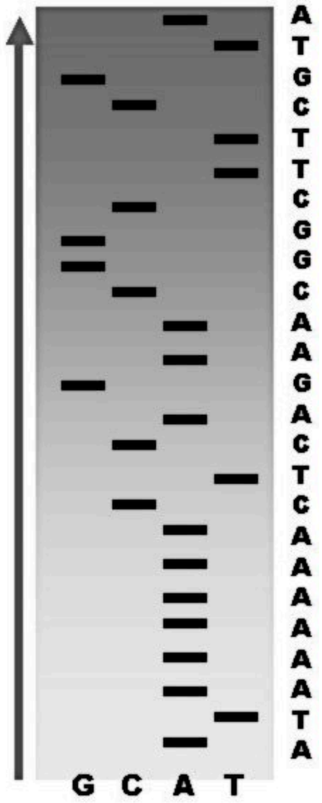
COT analysis

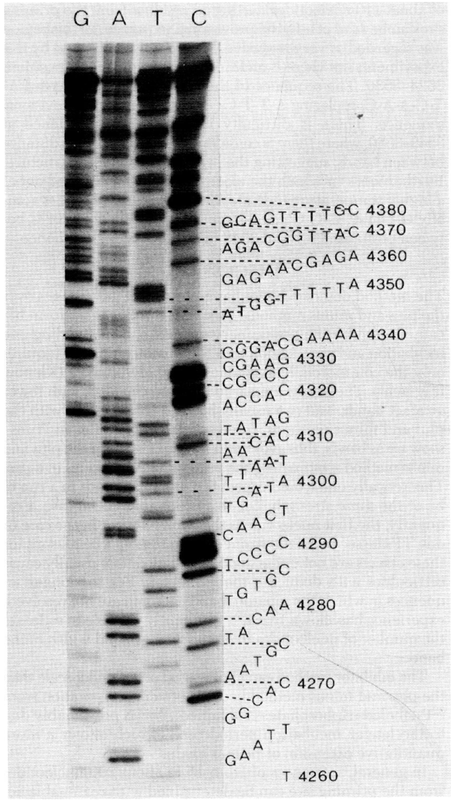


Sanger Sequencing (1977): sequencing 1 target at a time









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 (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

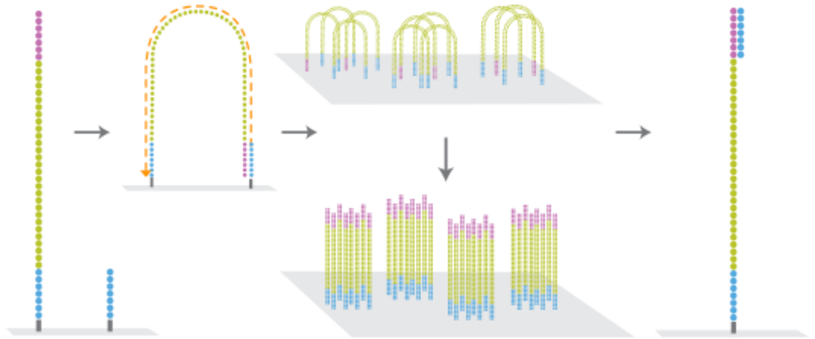
- Fewer, longer reads
- >1 kb (PacBio), up to 100s of kb (Oxford Nanopore)
- Relative high error rates

Oxford Nanopore MinION

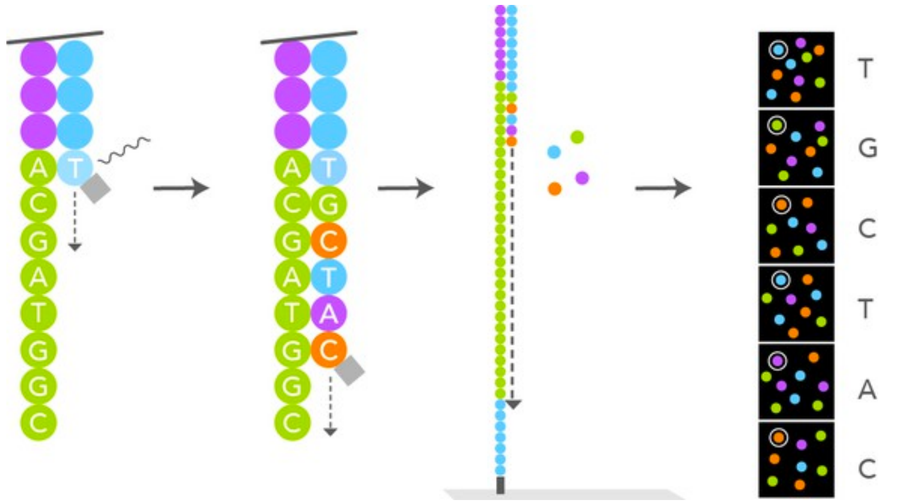
$1000

PacBio RS-II

Illumina 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

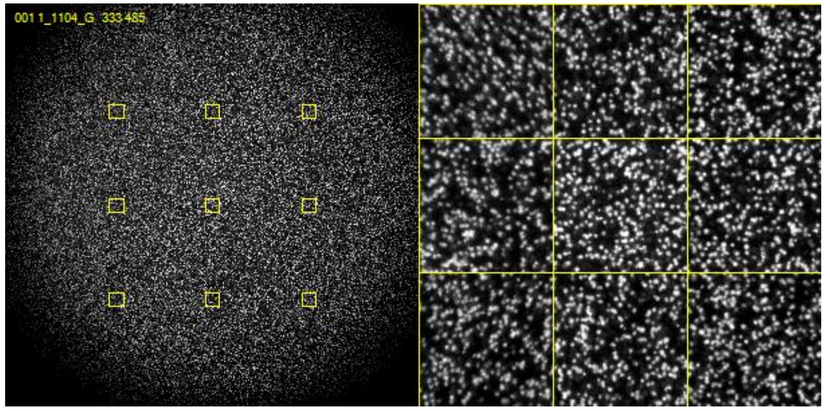
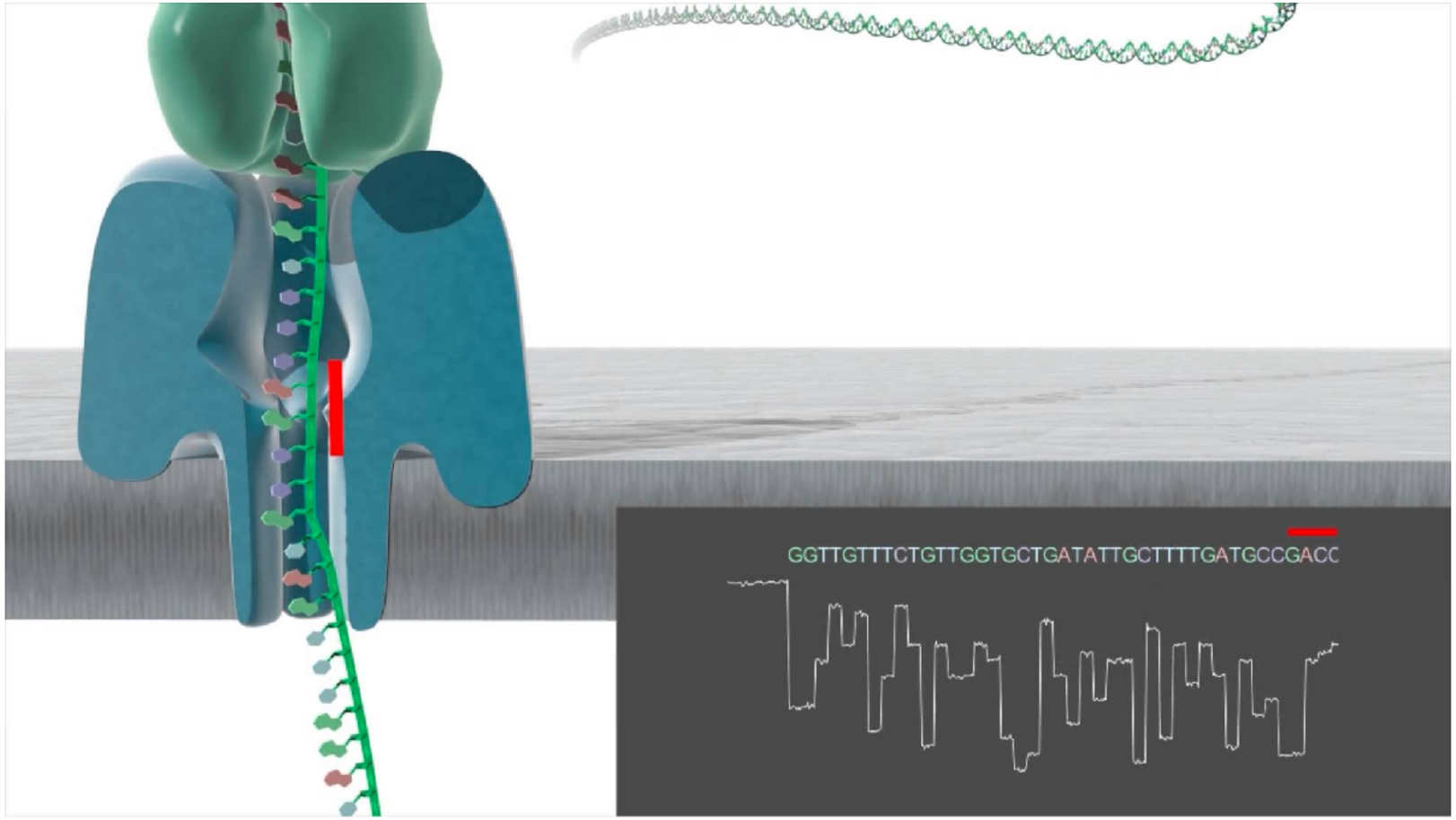


Image credit: Illumina

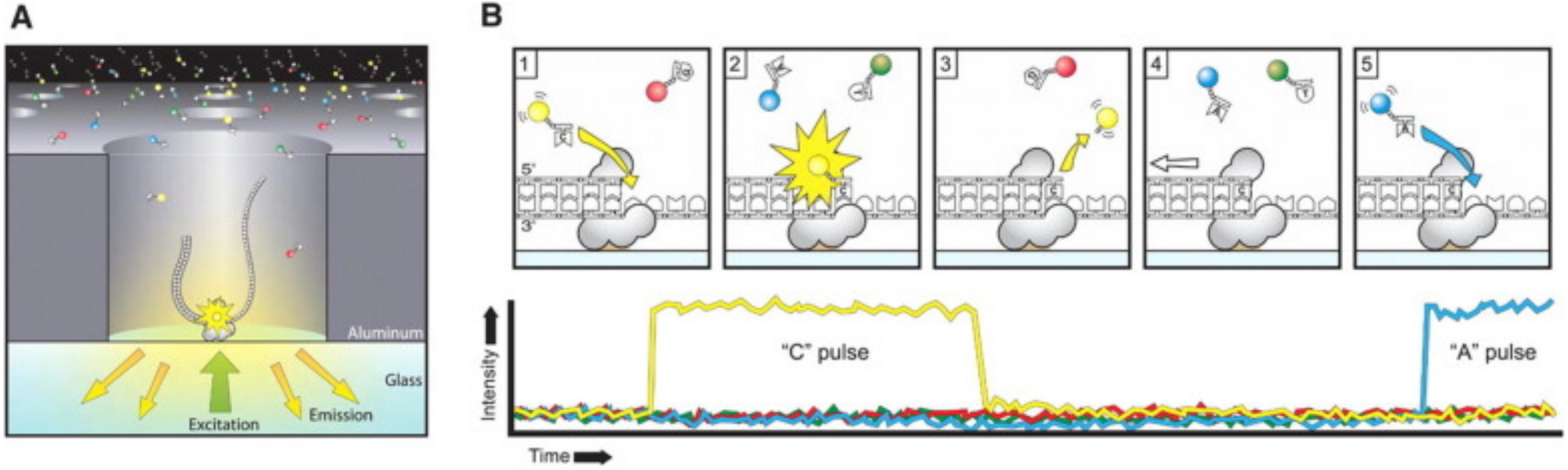
Long read sequencers sequence single molecules



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

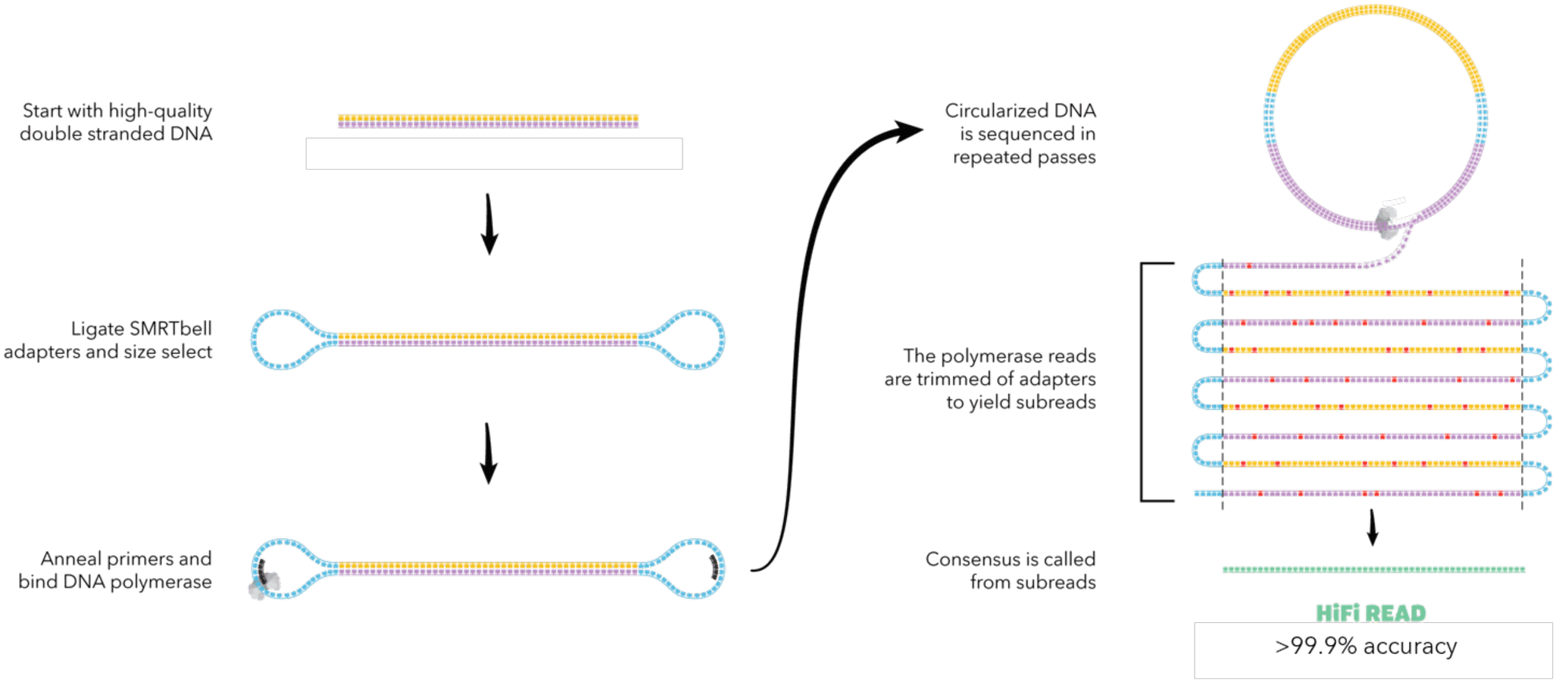
Image: Oxford Nanopore

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



Rhoads and Au, Genomics Proteomics Bioinformatics 13 (2015)

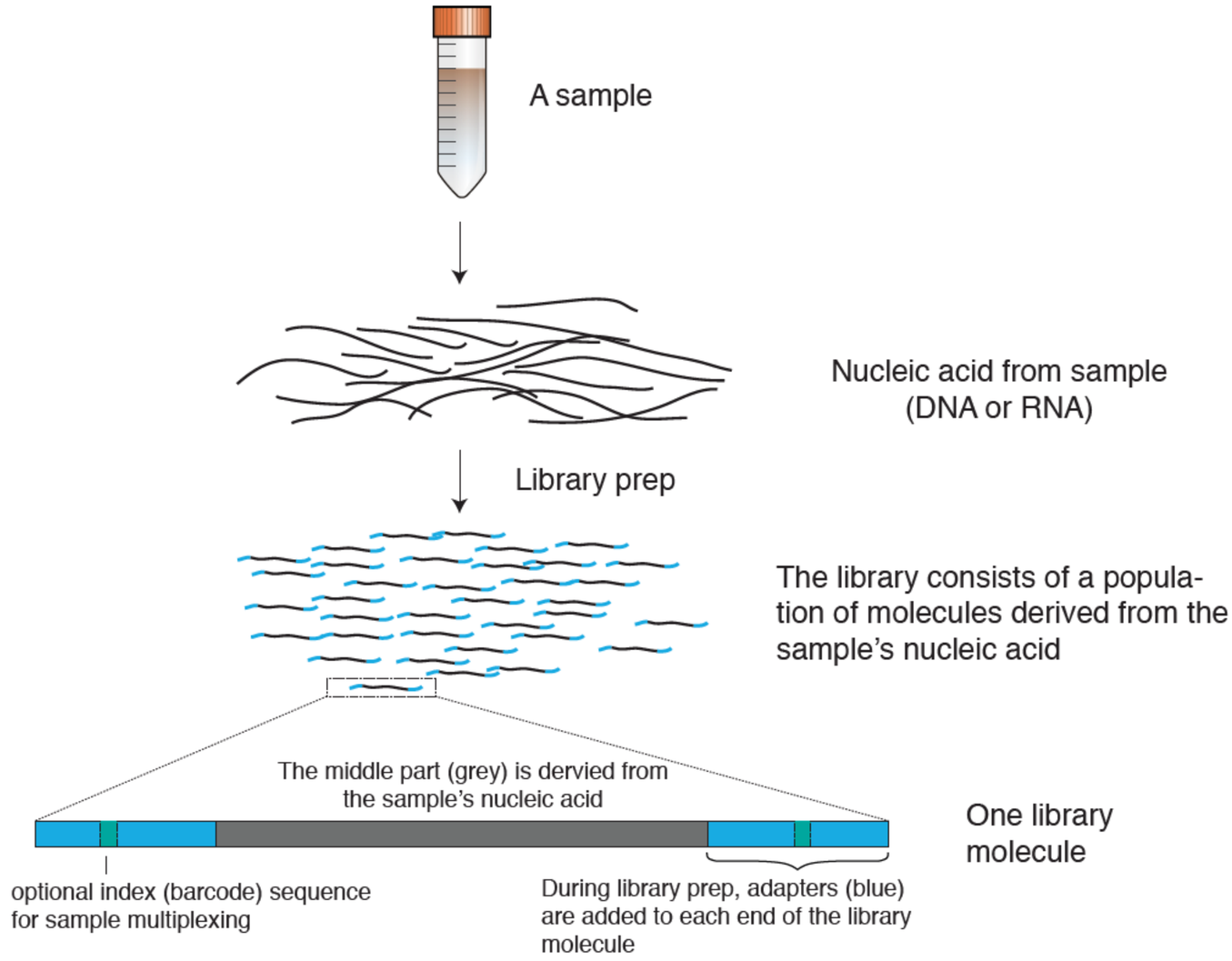
PacBio HiFi sequencing sequences the same molecule many times to reduce error rate



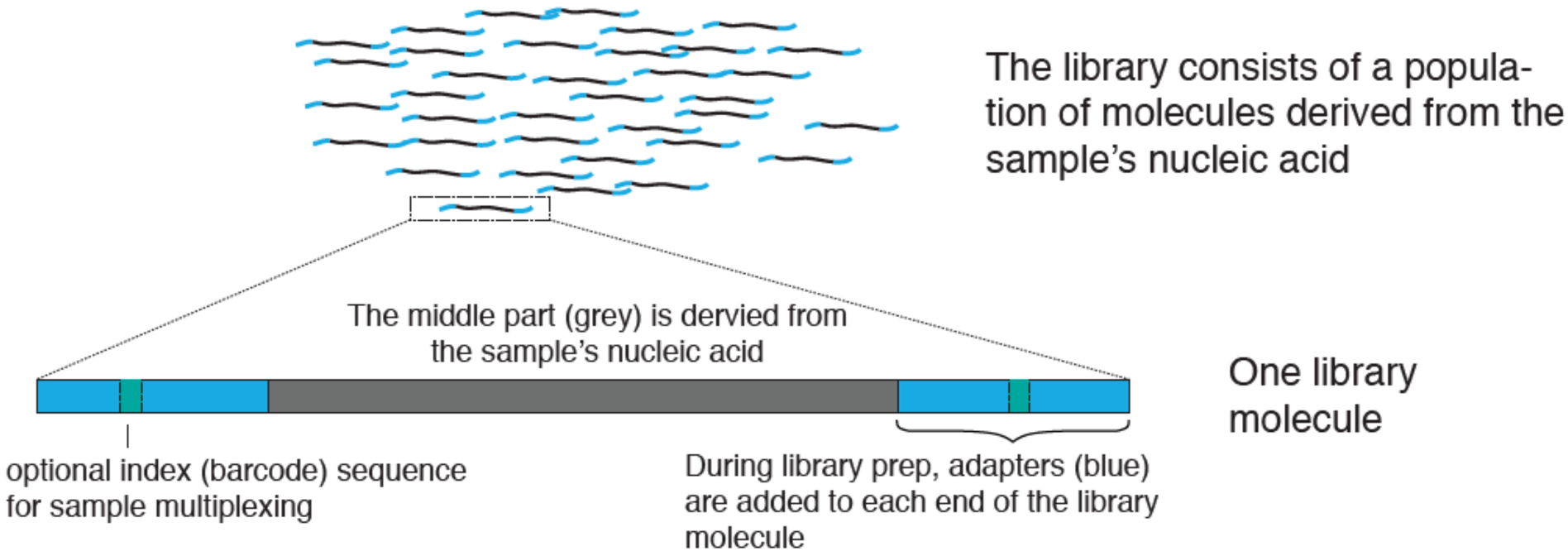
Best of both worlds: long reads with low error rate

Image: PacBio

Library prep converts nucleic acids into a form suitable to be sequenced
Details differ by sequencing platform

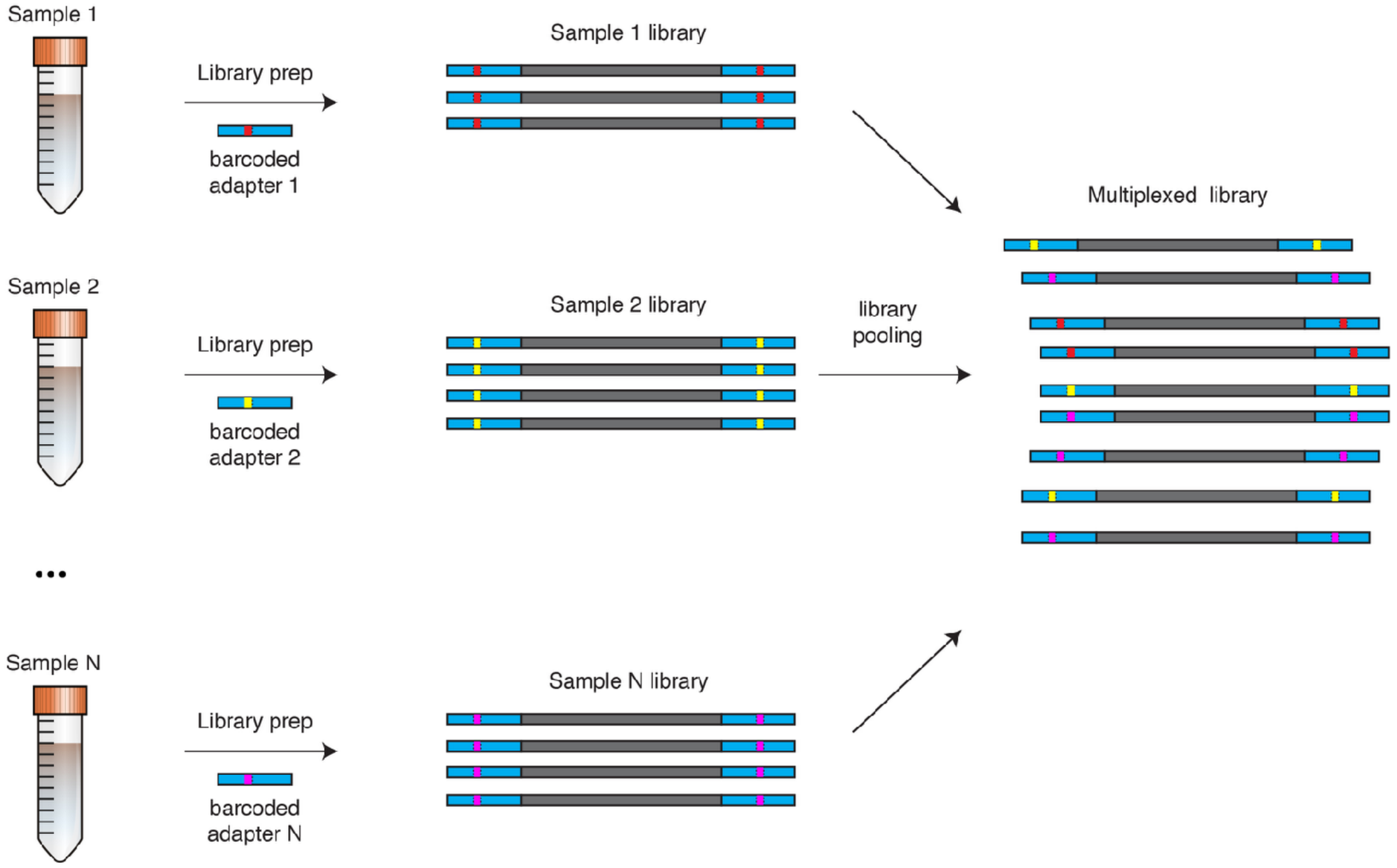


Library prep converts nucleic acids into a form suitable to be sequenced



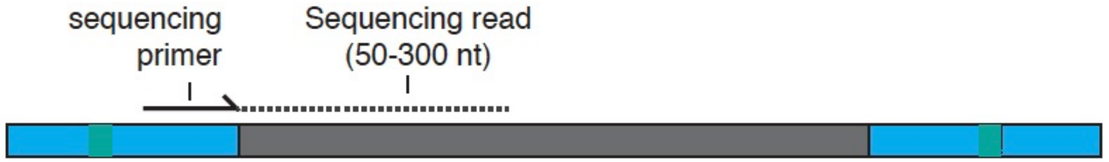
An example Illumina library molecule

Barcodes (or indexes) allow sample multiplexing



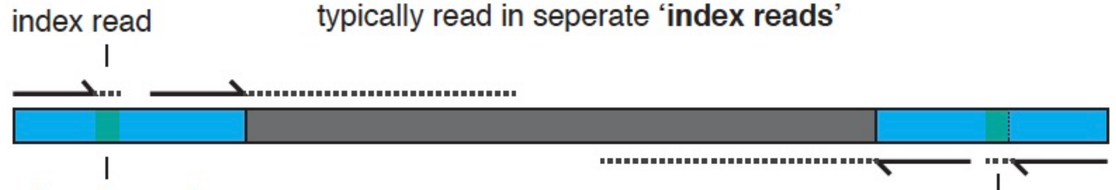
Illumina sequencing produces 1-4 reads per library molecule

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

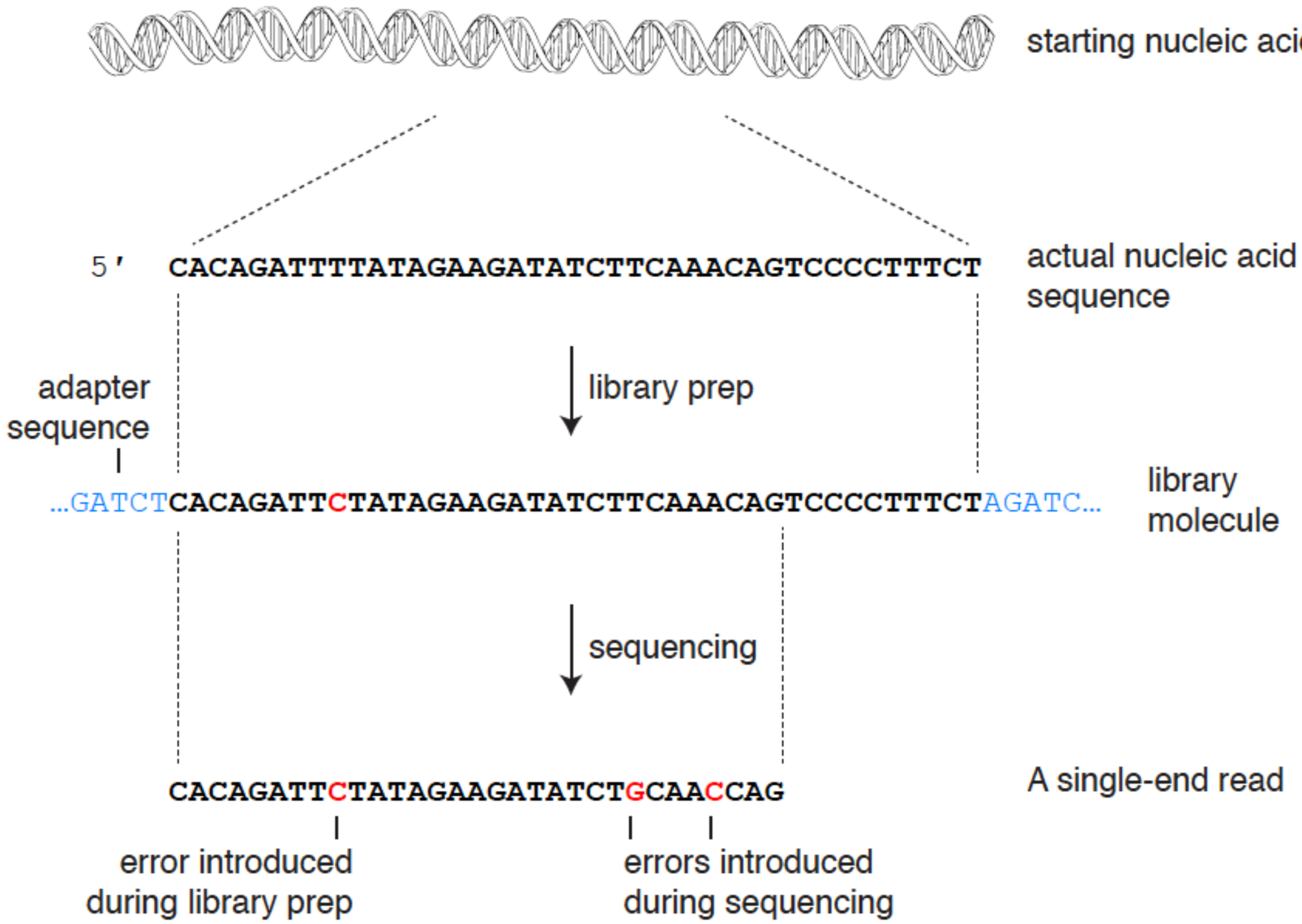


In **paired end sequencing**, a library molecule is sequenced from both end

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



Reads are measurements of the real sequence that often contain errors

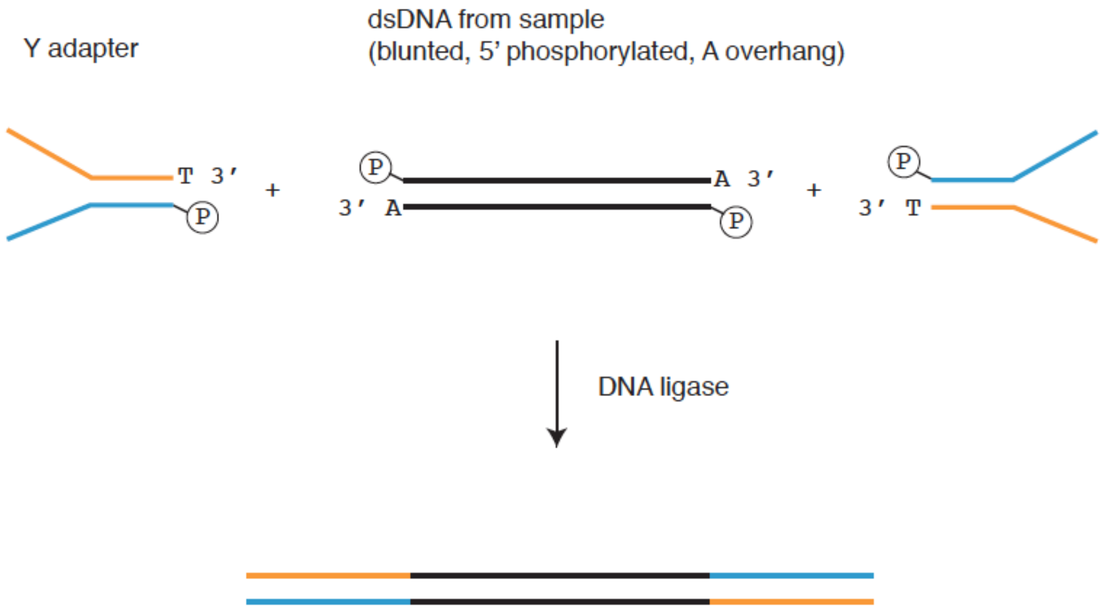


There are many good ways to make sequencing libraries

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

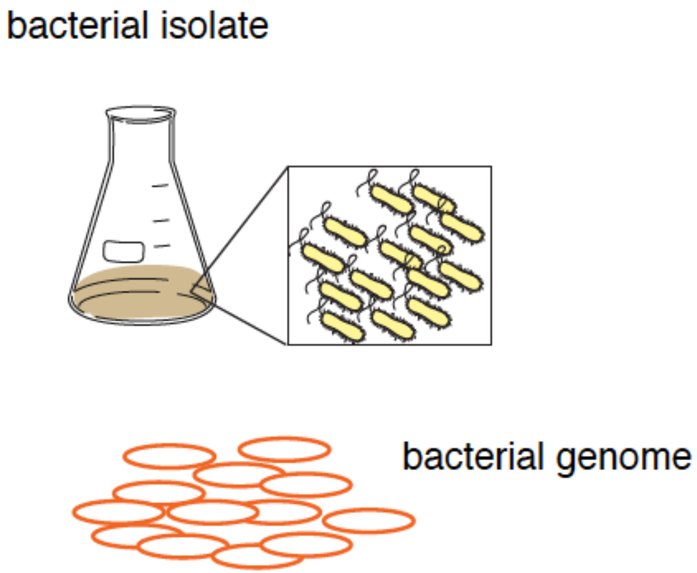
- Nucleic acid isolation
- Enrichment (of nucleic acid subtypes you want) or subtraction (of those you don't want)
- Nucleic acid fragmentation
- Conversion of RNA into dsDNA (for RNA sequencing)
- Addition of adapters to ends of library molecules, possibly with barcodes for multiplexing
- Library amplification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

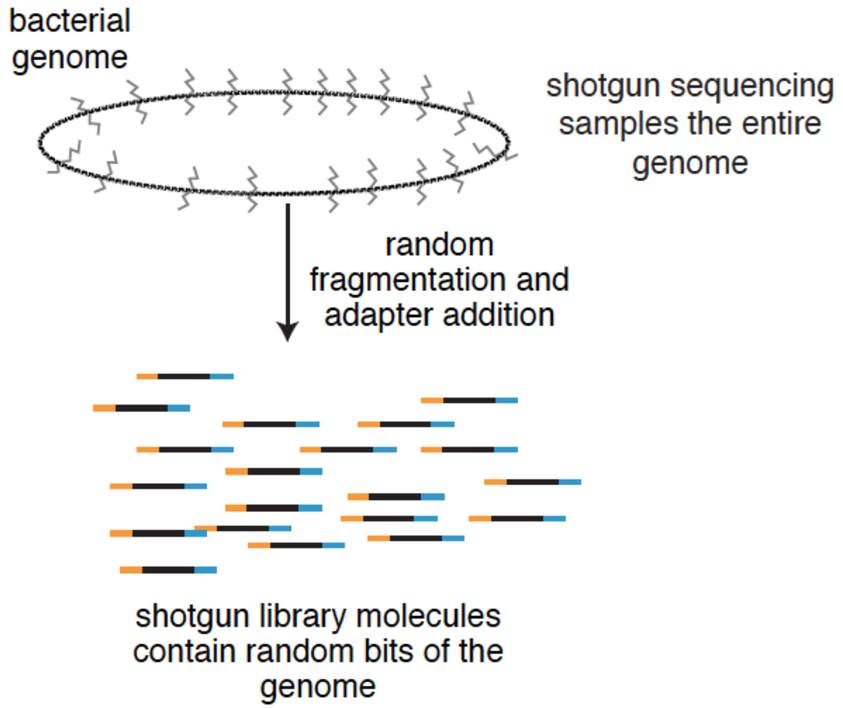
Adapters can be added to sample-derived dsDNA by ligation



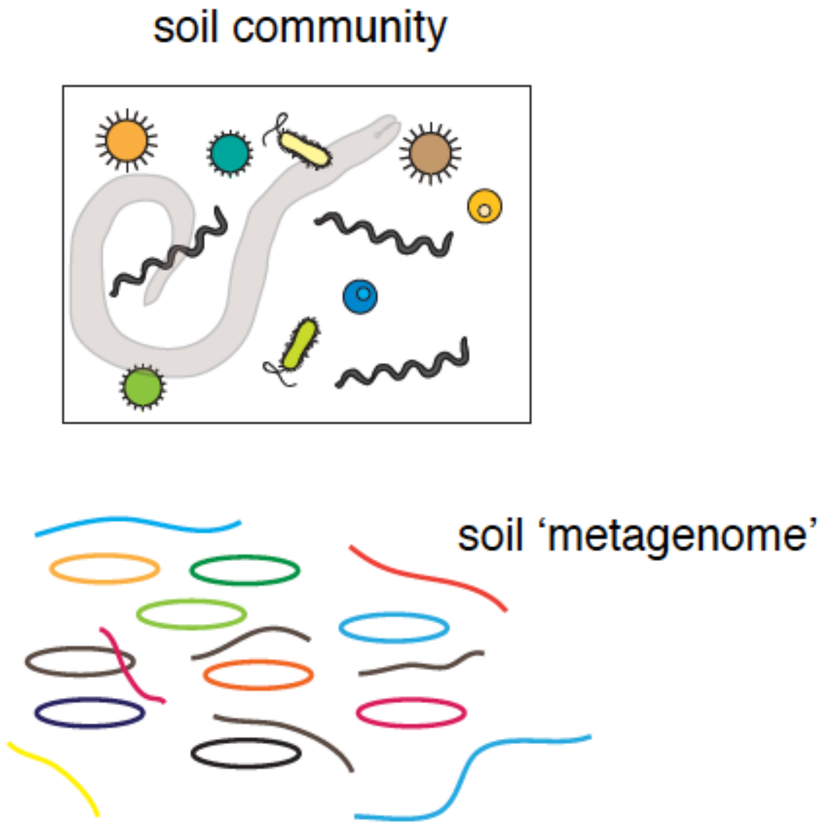
How you make a library determines what type of sequencing you're doing

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

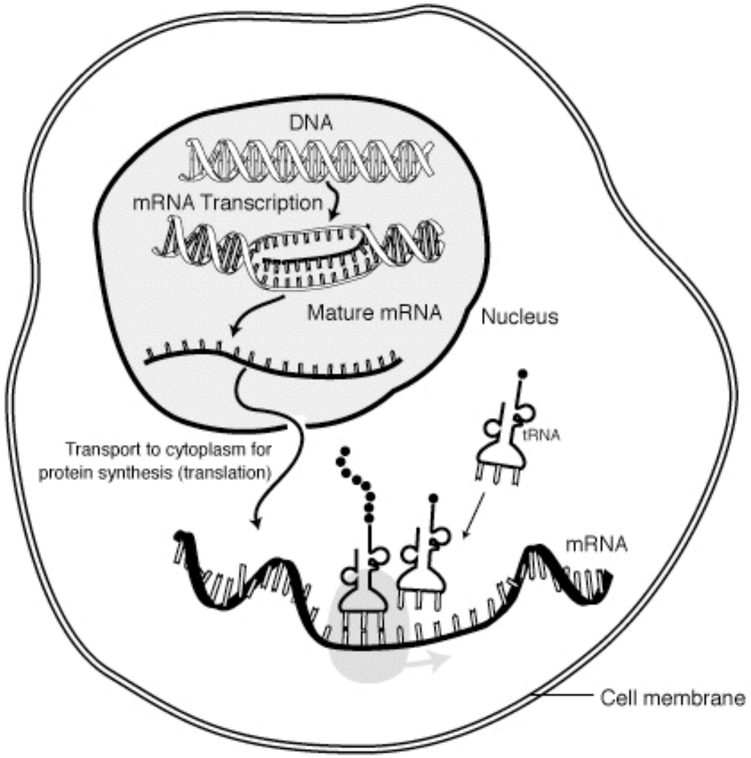


Could make a 16S or a shotgun library from these genomes

Sequencing of RNAs from a complex sample like this is metatranscriptomics

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

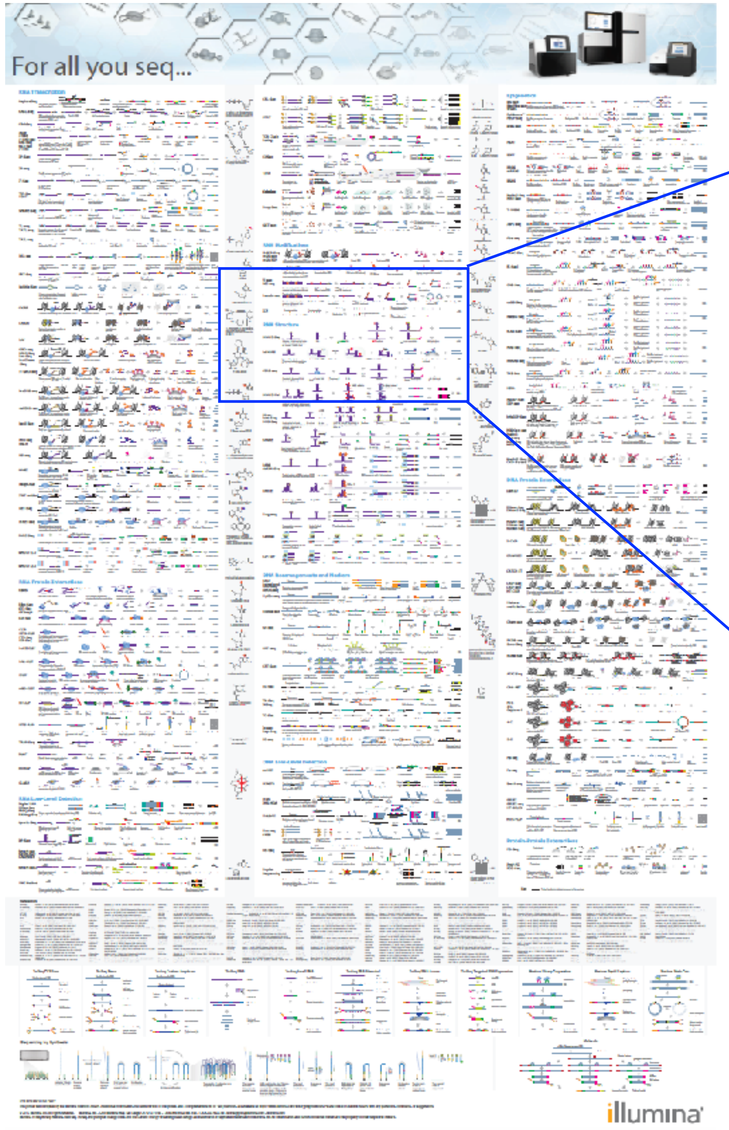
If you make a library from mRNA, that is RNA-Seq (transcriptome sequencing)

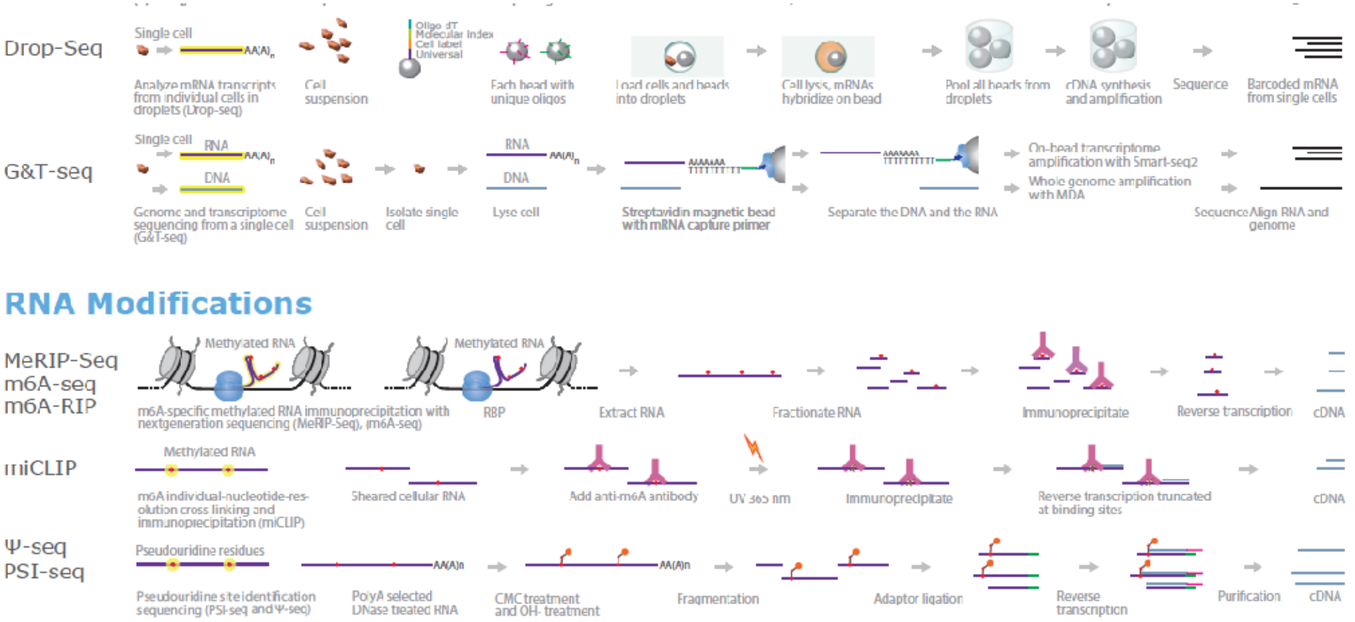


The abundance of reads from a particular mRNA is proportional to that mRNA's abundance in the cell

Image: Wikipedia

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

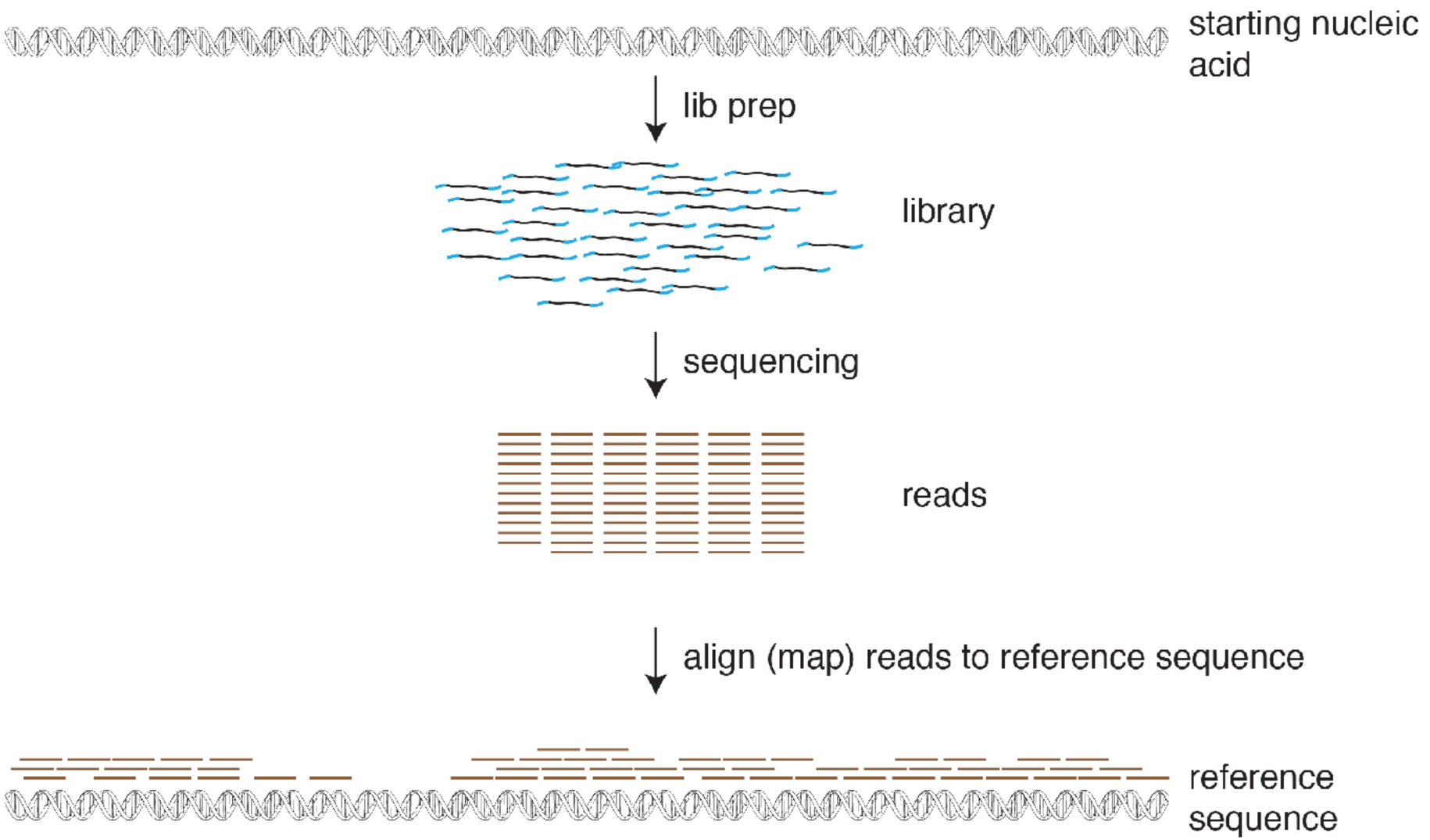




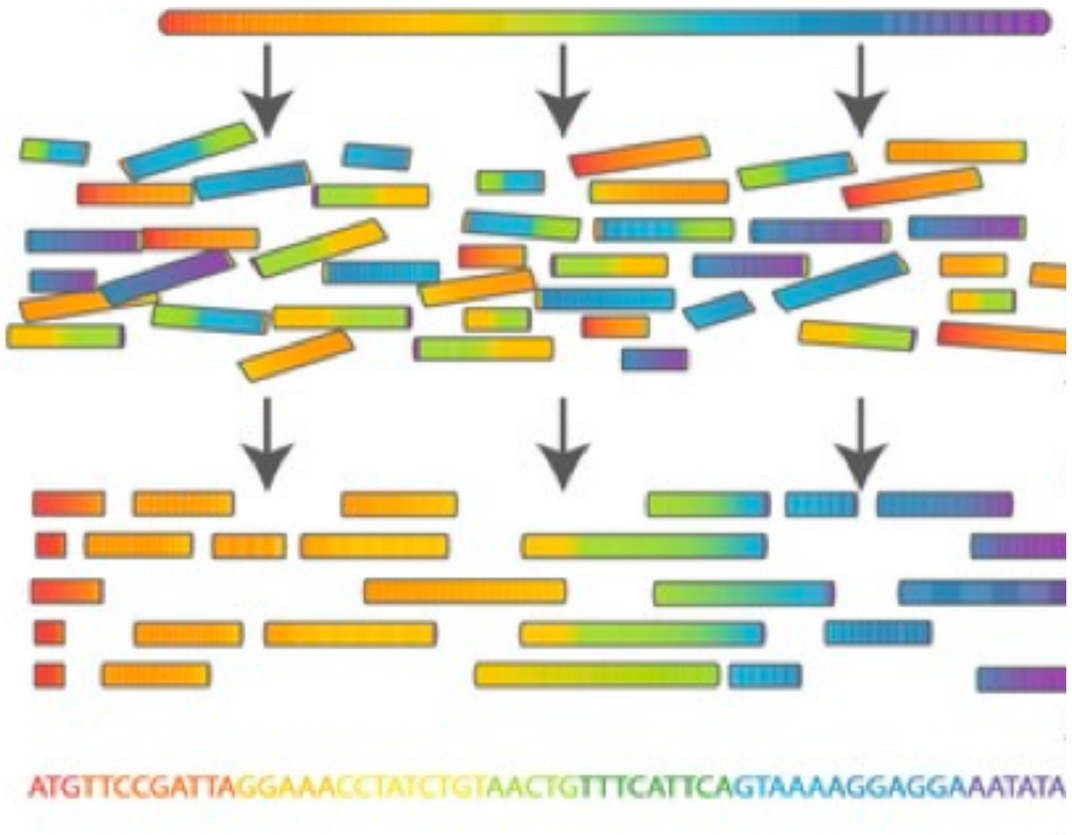
They're all variations on a few themes. Don't let it overwhelm you. Most sequencing is of a few simple types, and it's better to focus on the Biology and experimental design.

https://www.illumina.com/content/dam/illumina-marketing/documents/applications/ngs-library-prep/ForAllYouSeqMethods.pdf

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



not yet sequenced genome

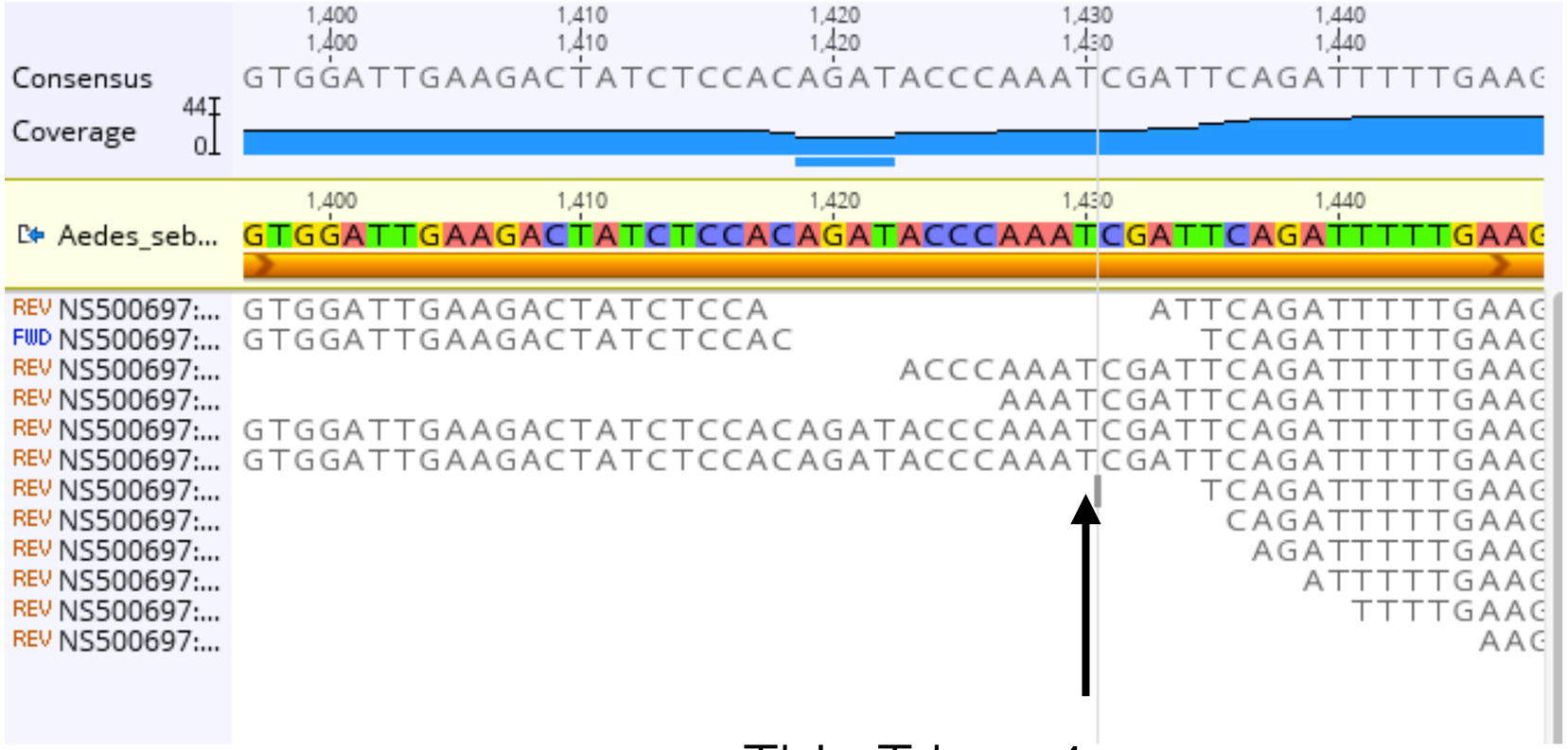
reads

contigs

reconstructed (assembled) genome

Image: Commins, J., Toft, C., Fares, M. A (2009)

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

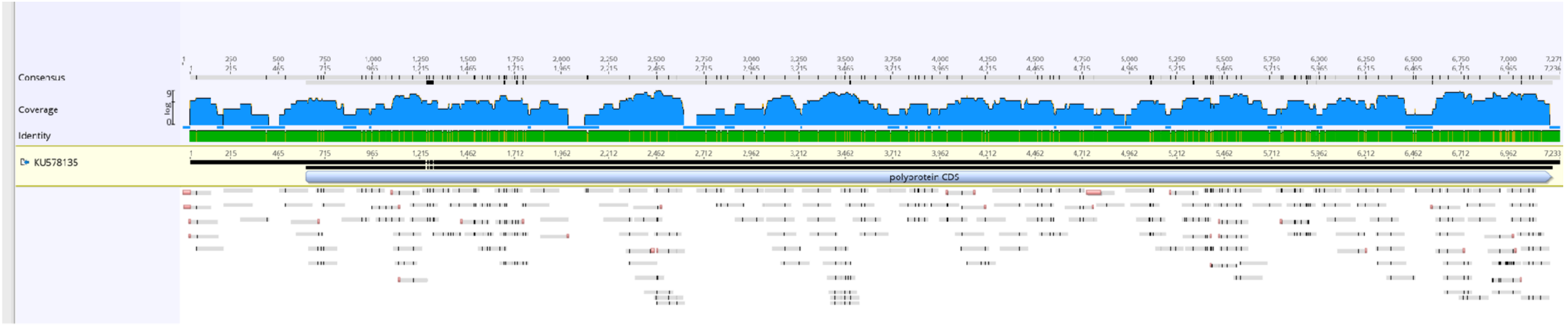


This T has 4x coverage

coverage is often referred to as ‘depth’ or ‘depth of 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)

hand foot and mouth disease

(Mayo clinic)

Questions?

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