An overview of sequencing terminology and practices

Mark Stenglein, BZ/MIP 565

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

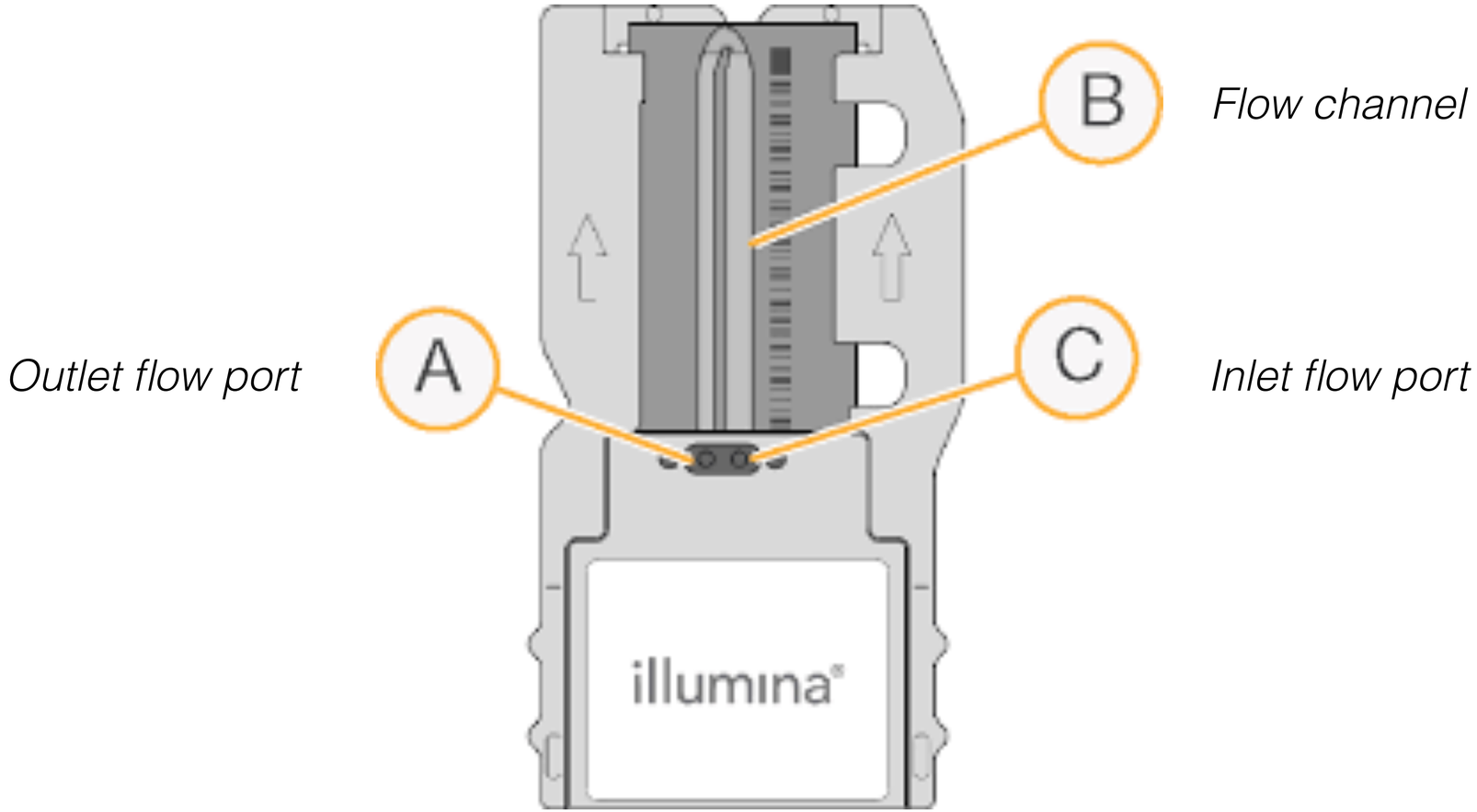
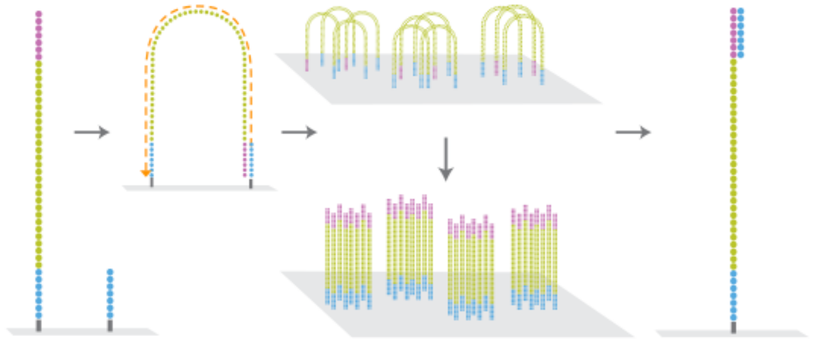
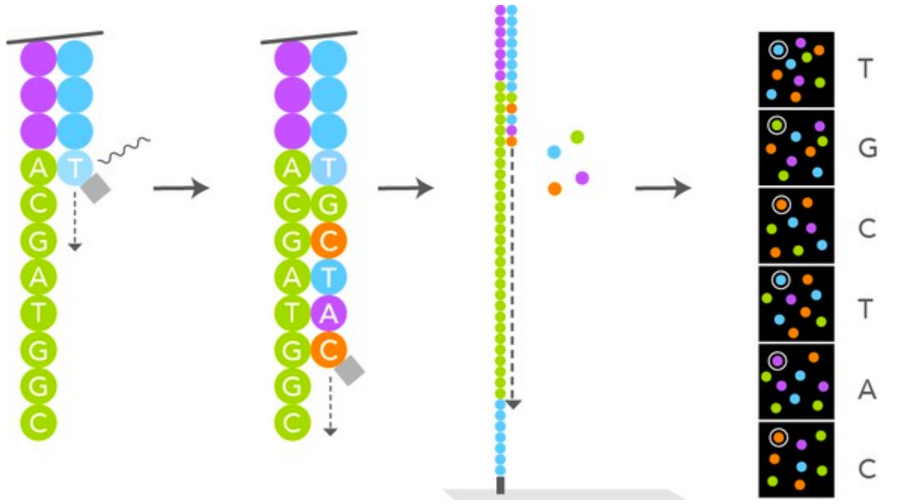


Image credit: Illumina

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

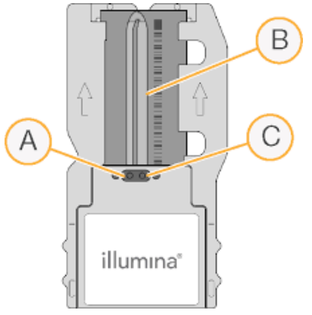
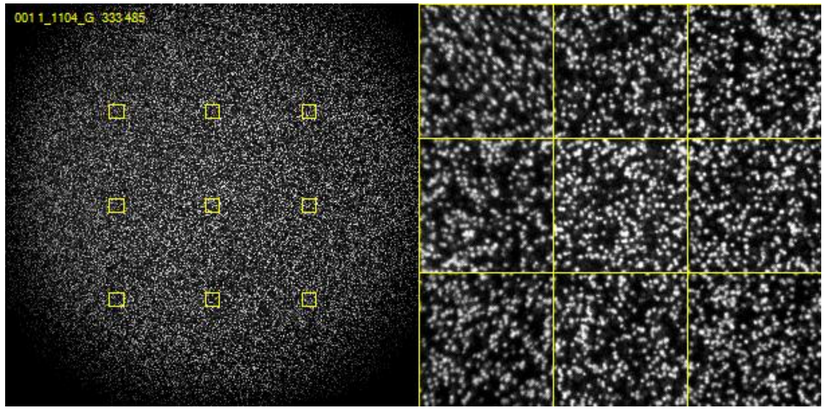


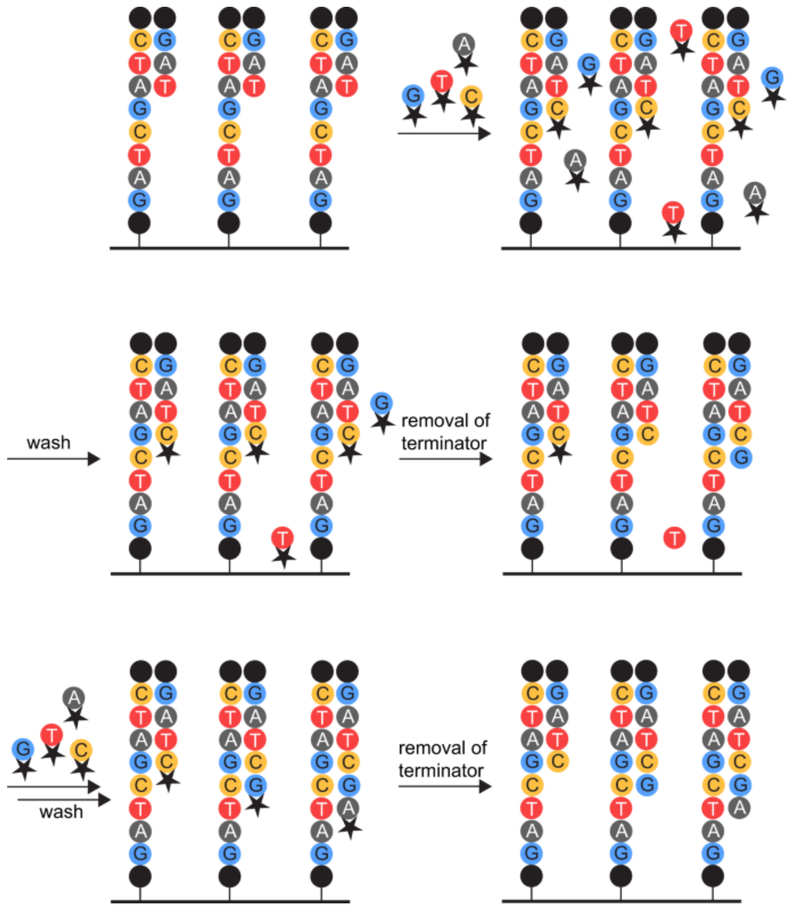
Image credit: Illumina

real raw Illumina sequencing data

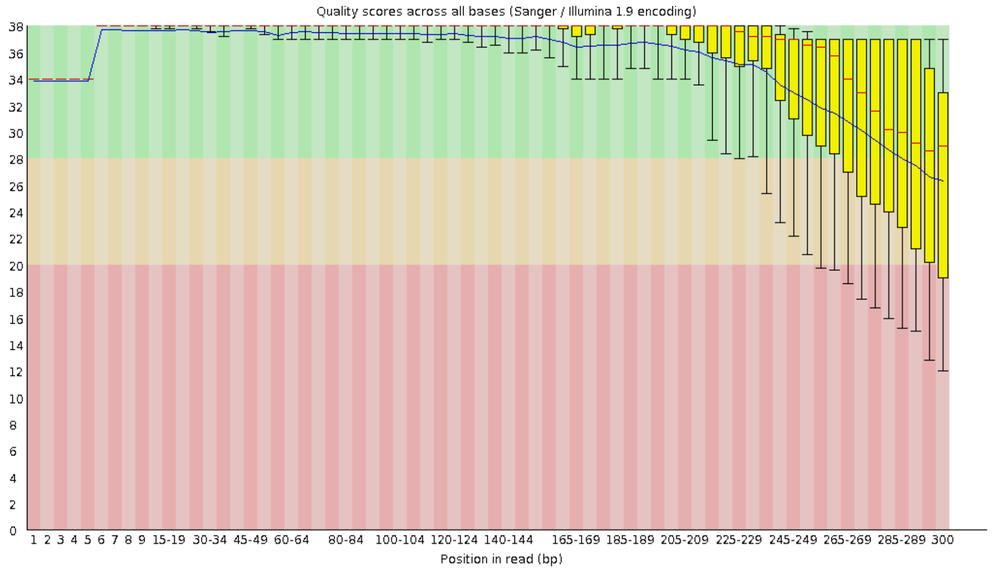


Illumina read length is limited by the “phasing problem”

Incorporating 0 or >1 base in a sequencing cycle can cause the individual copies in a cluster to get out of sync with each other (out of phase), leading to a mixed fluorescent signal at later cycles

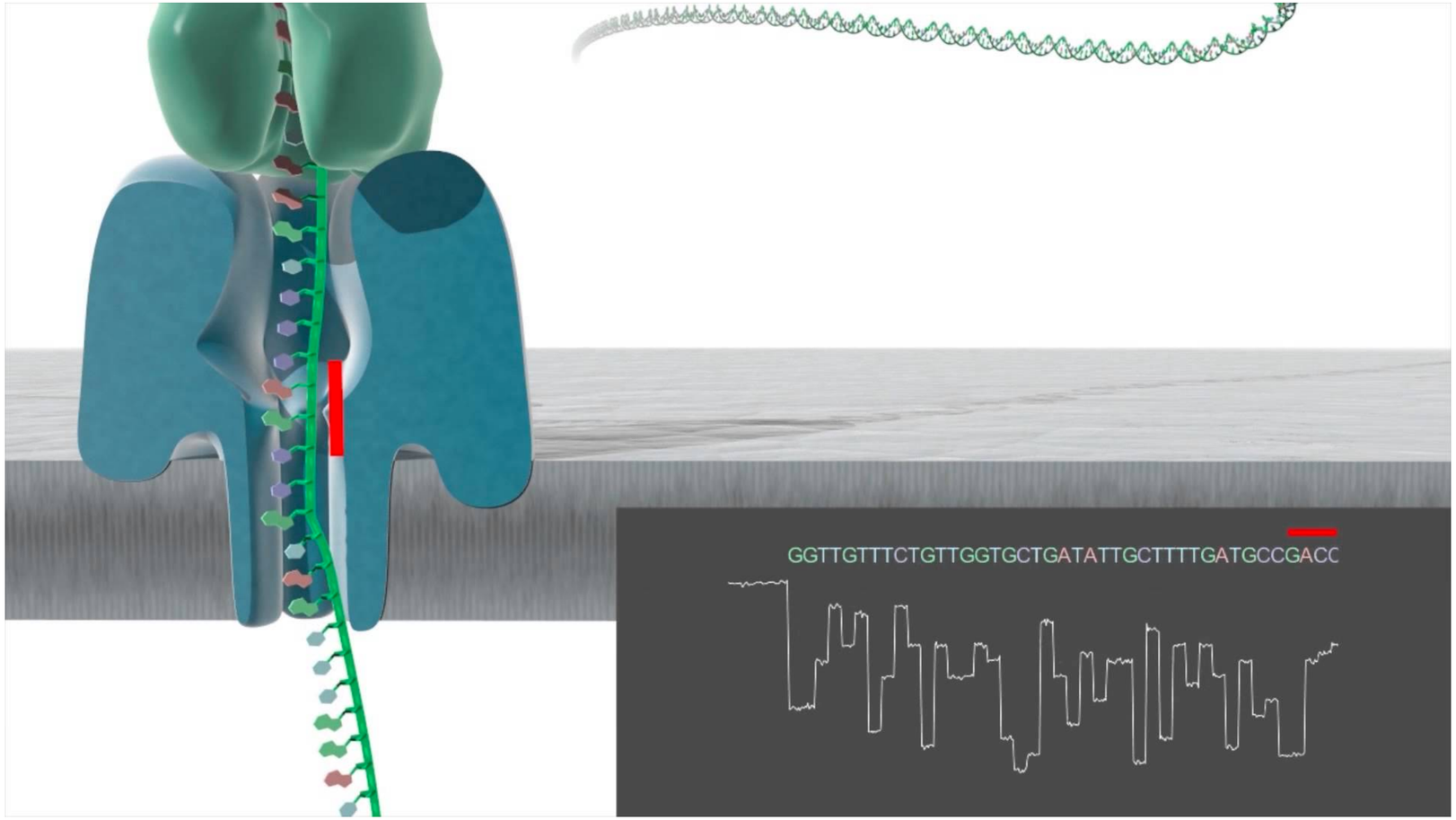


As a consequence, Illumina error rate invariably increases as reads get longer



Pfeiffer et al (2018) Sci Reports

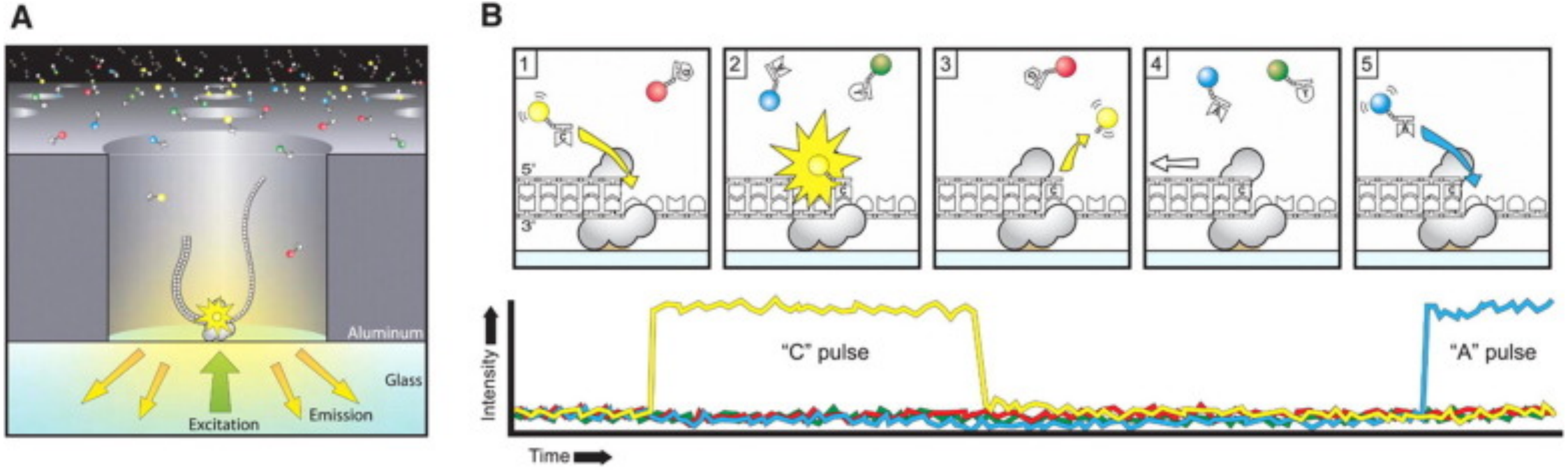
Long read sequencers sequence single molecules



Much longer reads, but with much higher error rates

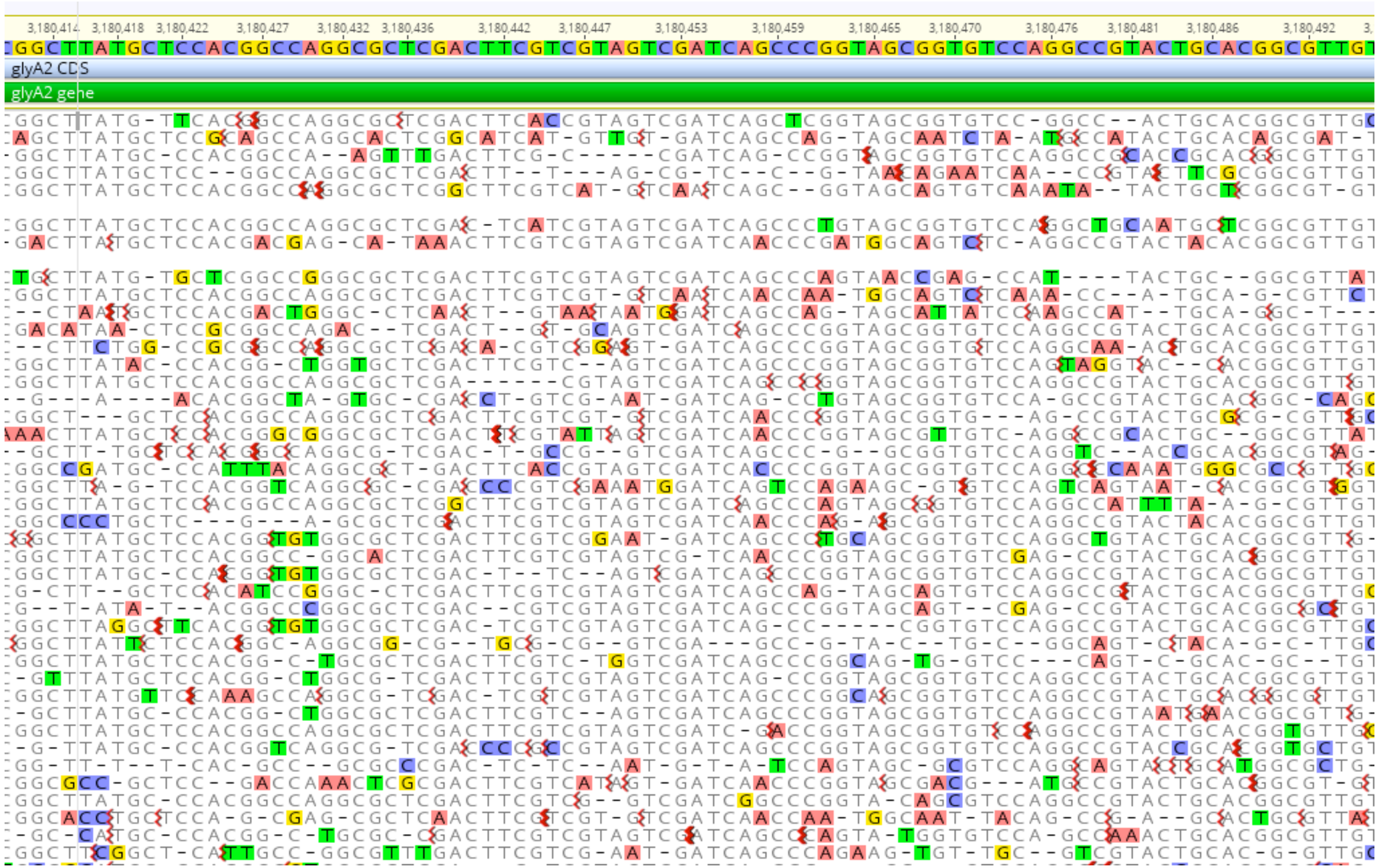
Image: Oxford Nanopore

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

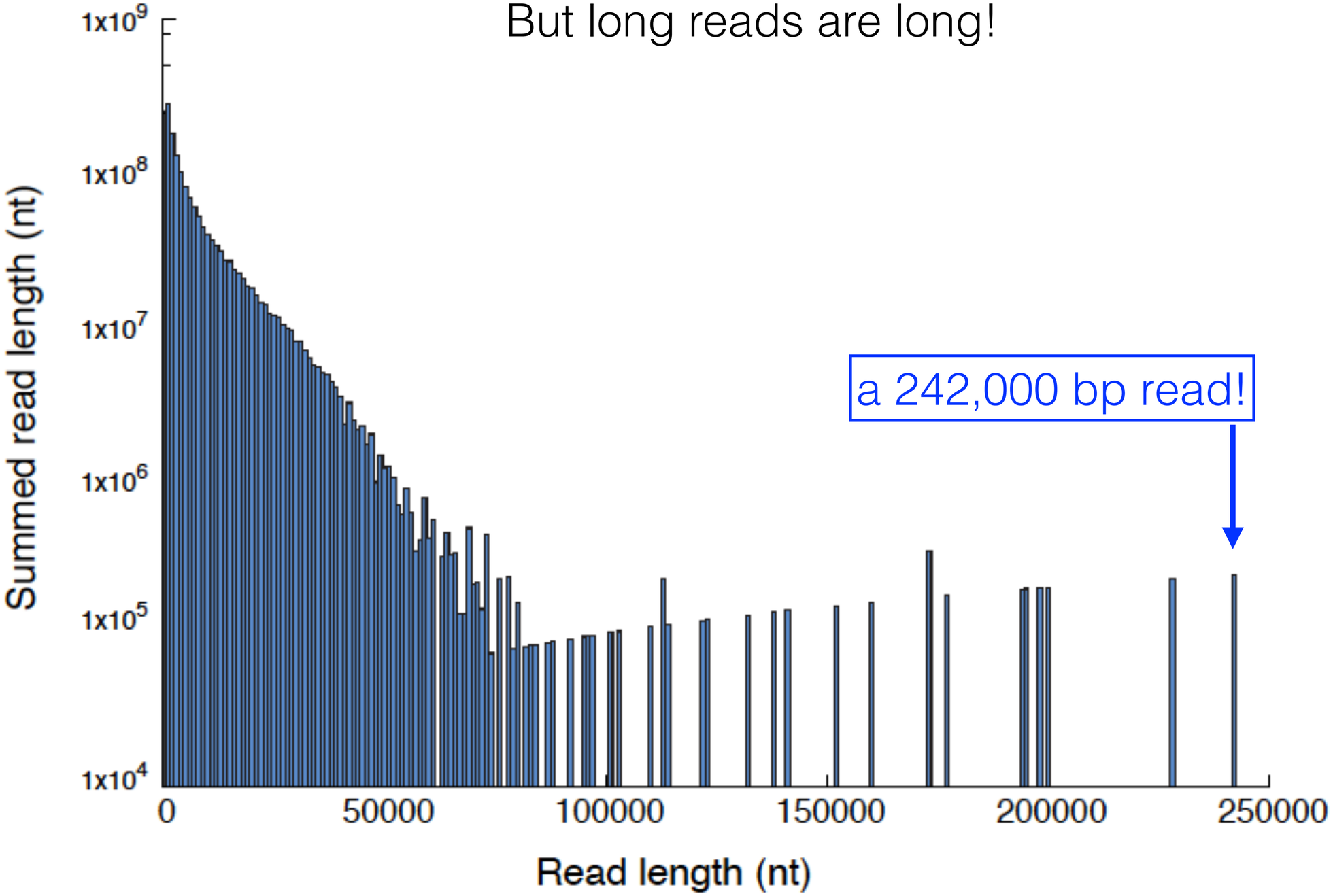


Rhoads and Au, Genomics Proteomics Bioinformatics 13 (2015)

Long reads have very high error rates (up to 10-20%)

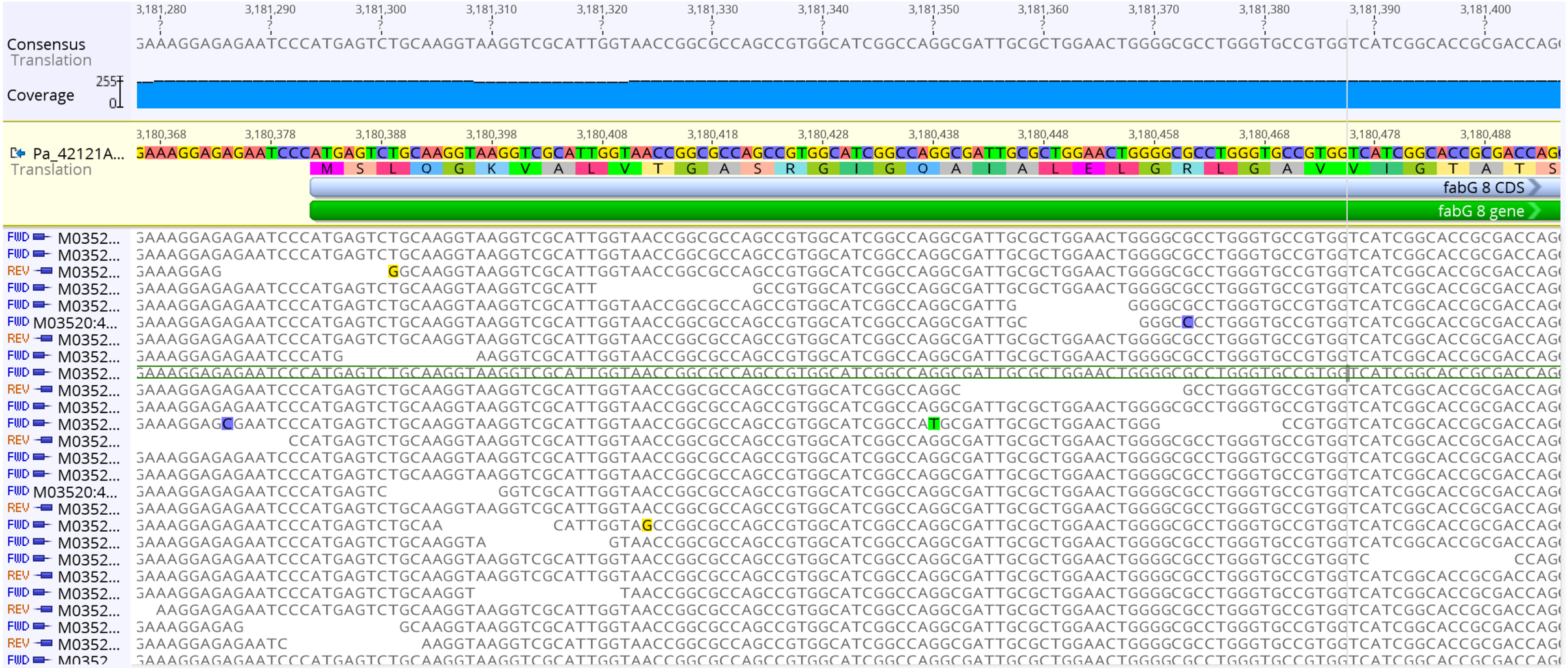


But long reads are long!

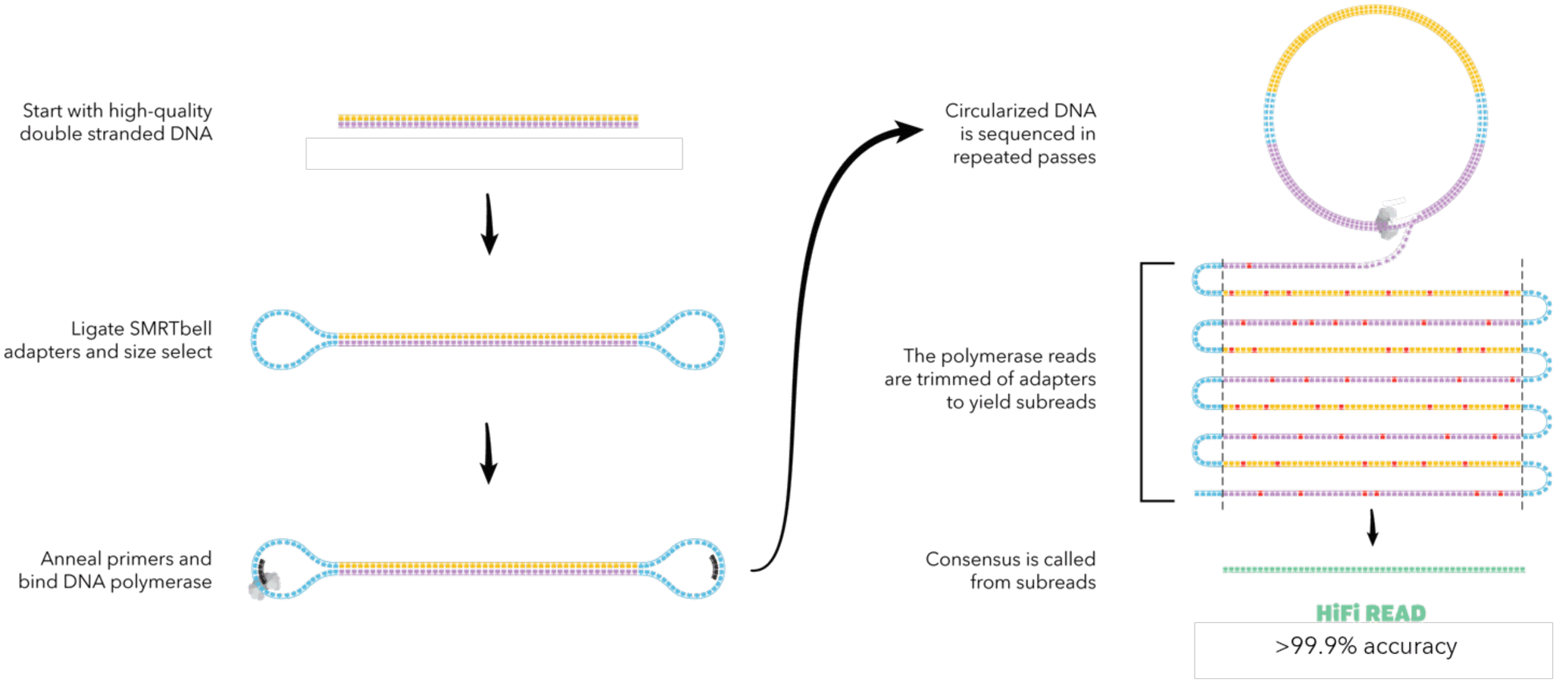


Justin Lee

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



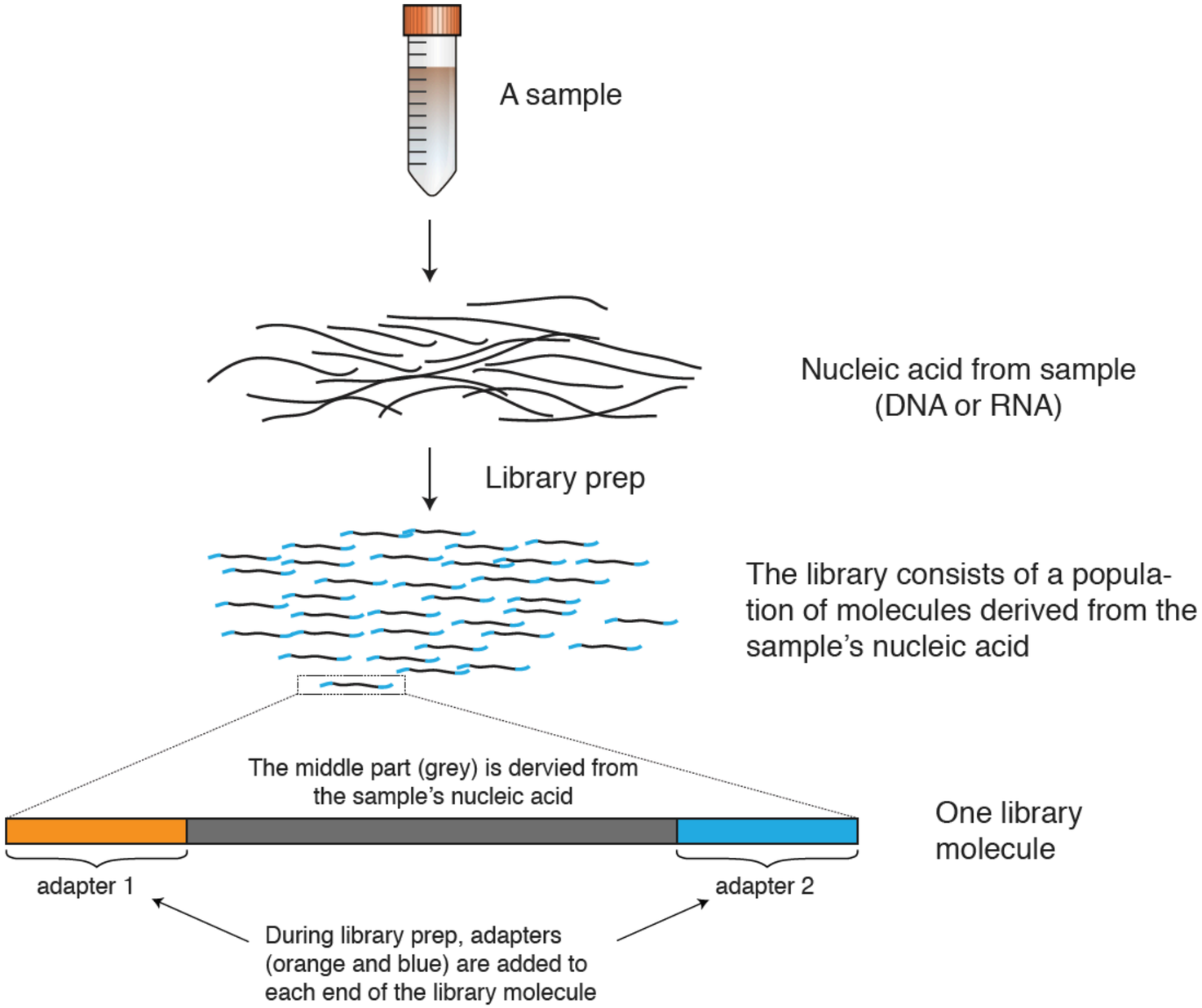
PacBio HiFi sequencing sequences the same molecule many times to reduce error rate



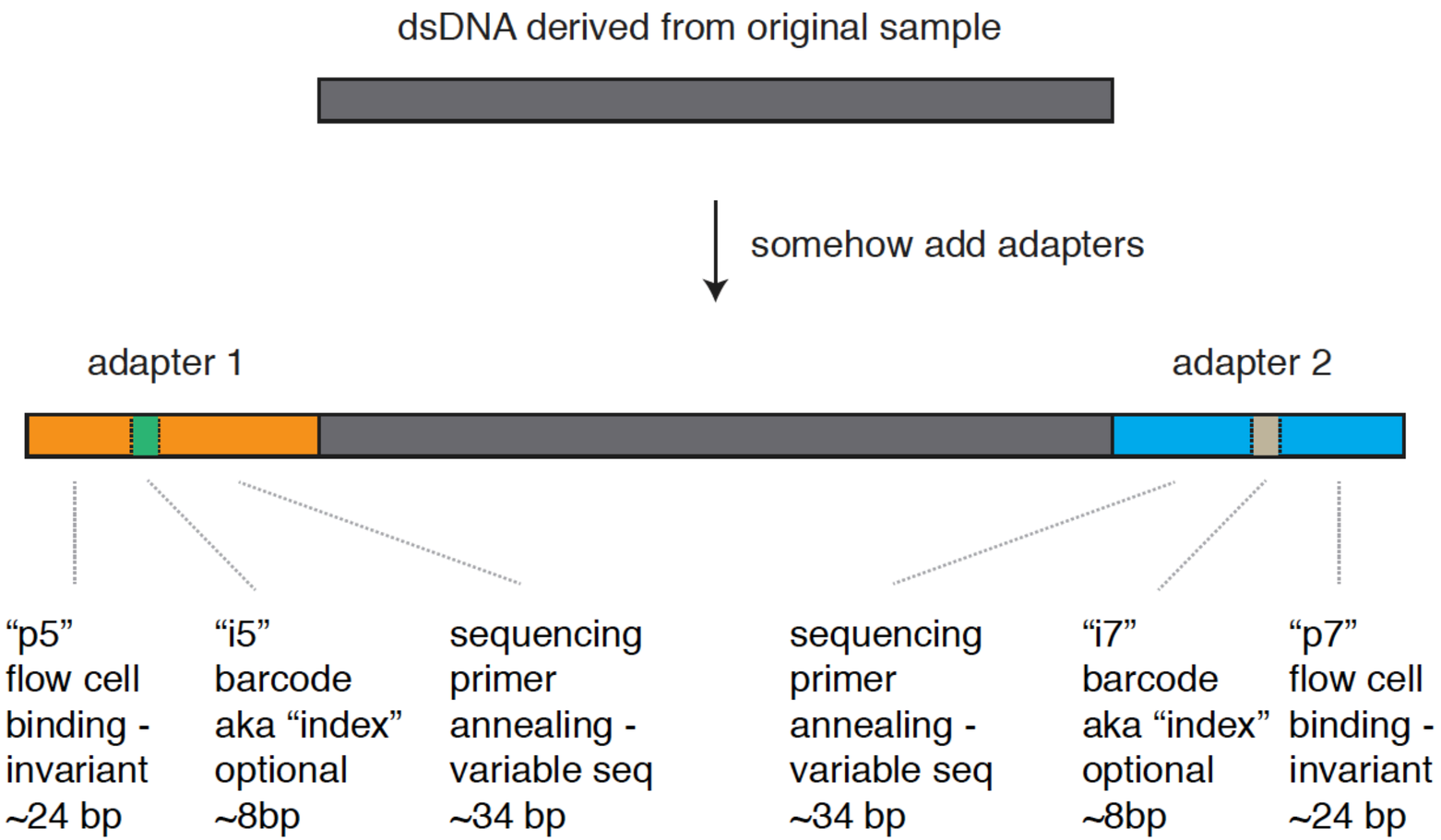
Best of both worlds: long reads with low error rate

Image: PacBio

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

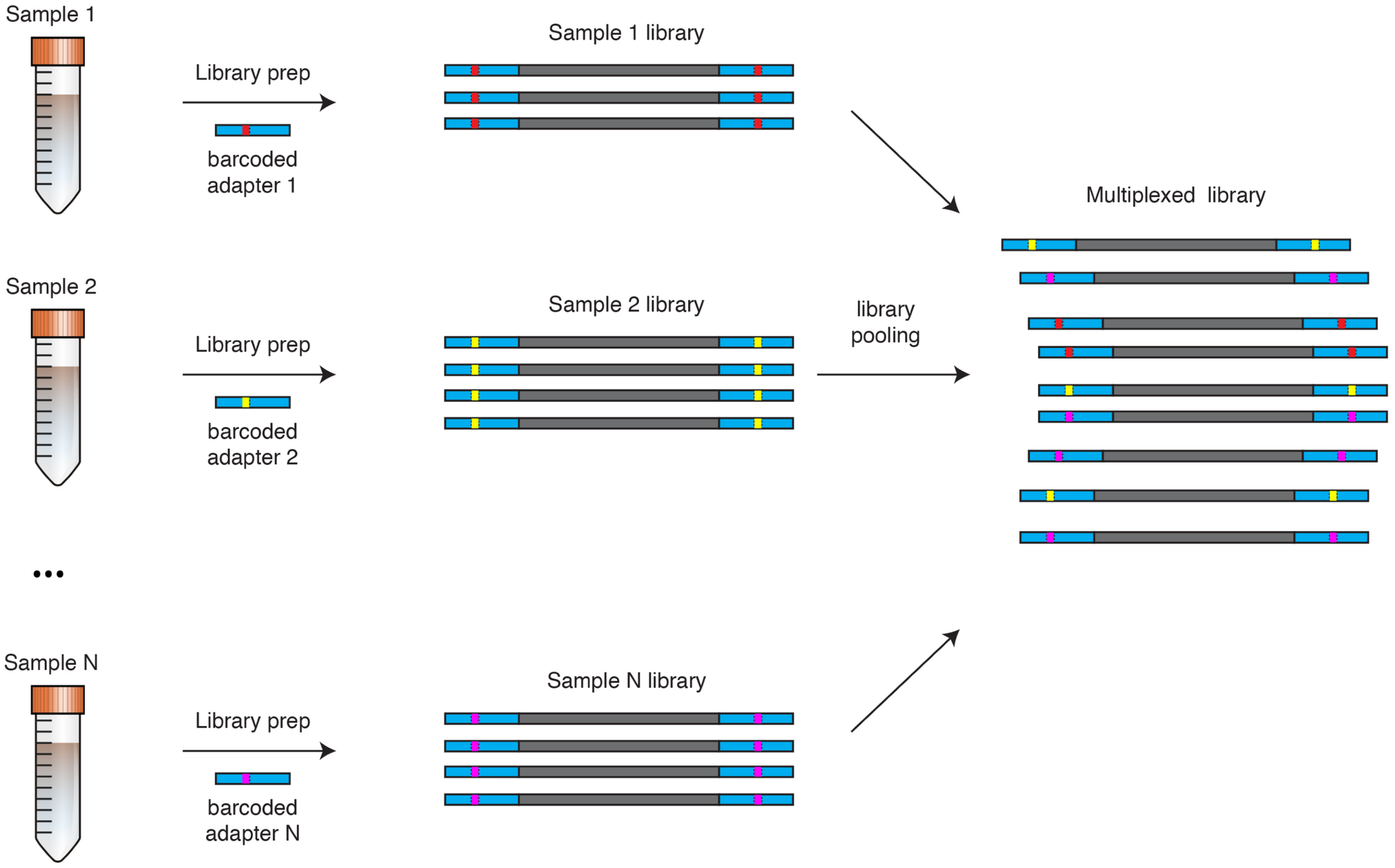


Library prep converts nucleic acids into a form suitable to be sequenced



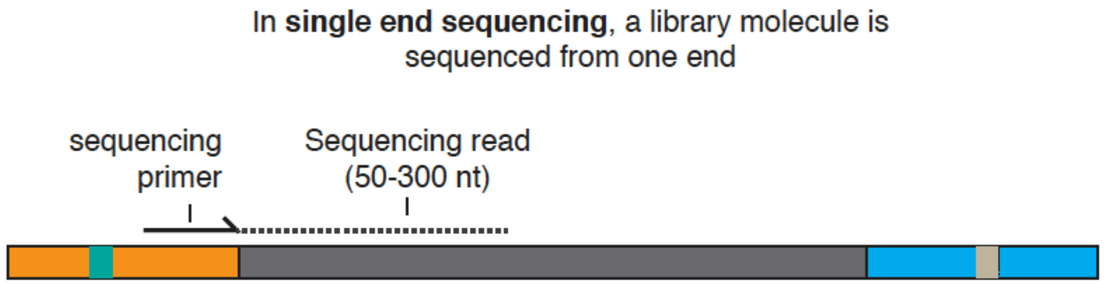
An example Illumina library molecule - the library will be a population of similar molecules

Barcodes (or indexes) allow sample multiplexing

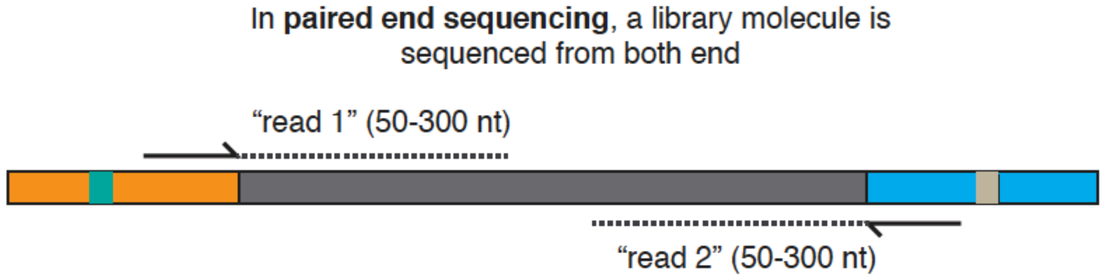


Illumina sequencing produces 1-4 reads per library molecule

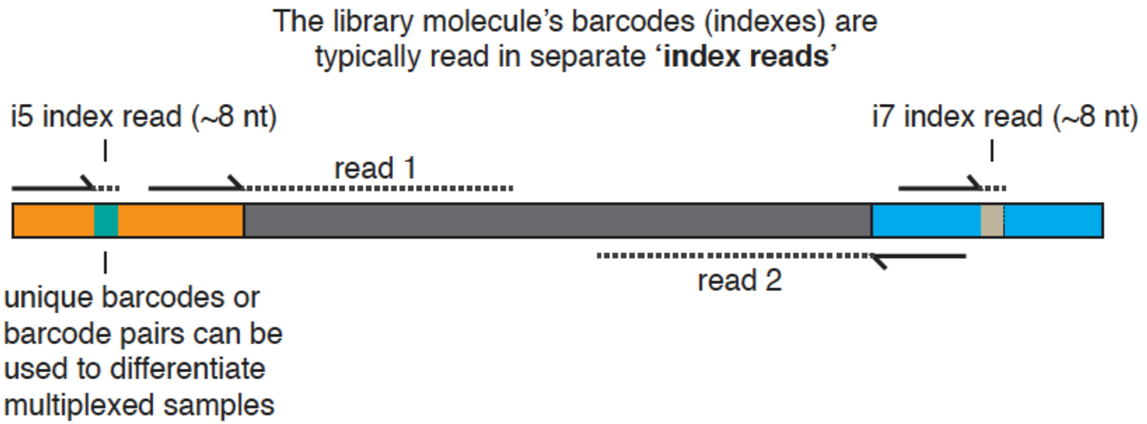
1 read per cluster
single end read
no index read



2 reads per cluster
paired reads
no index reads



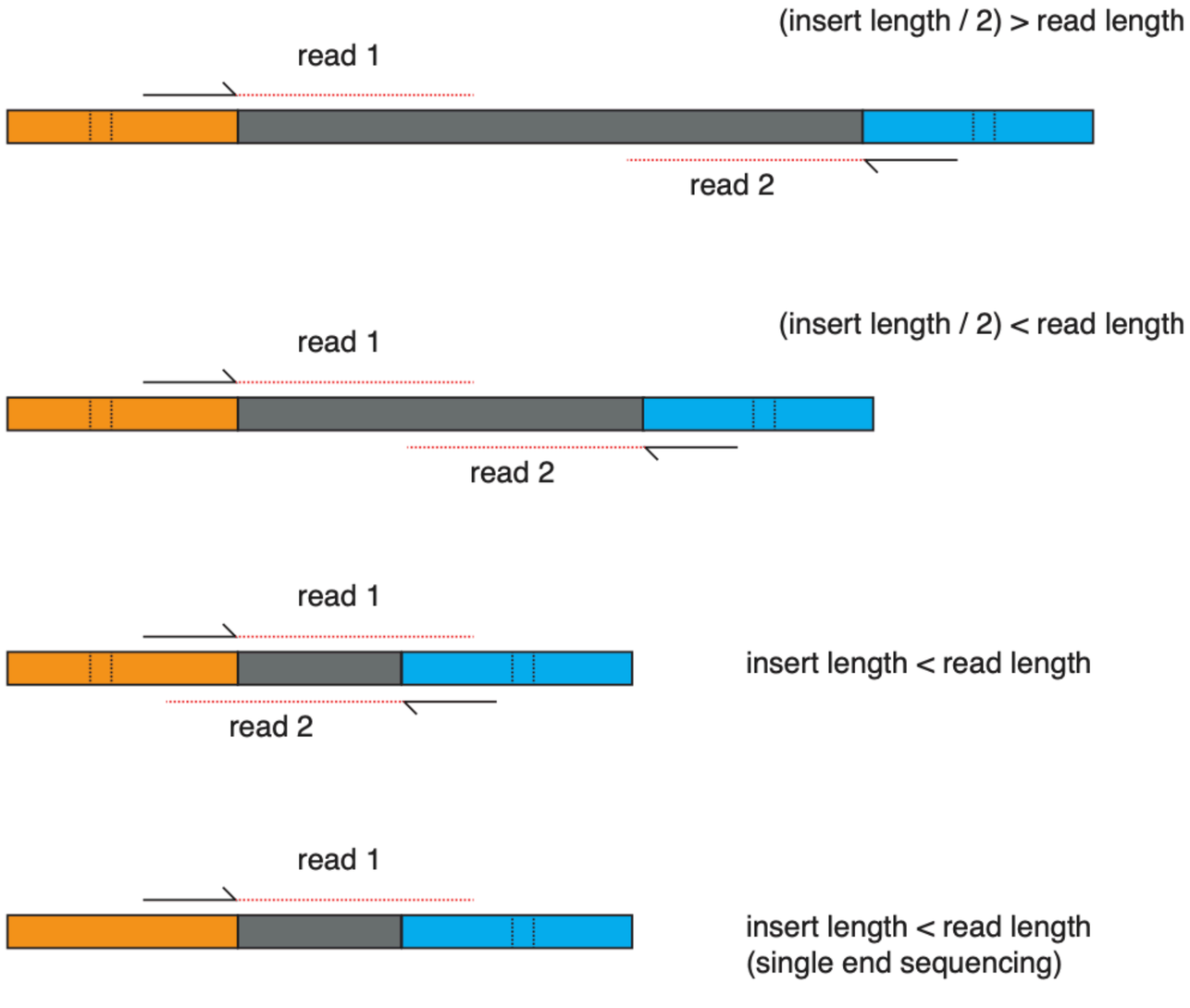
4 reads per cluster
Paired reads
dual index reads



Other combinations possible:
(single end, 2 index reads or
Paired end, 1 index read)

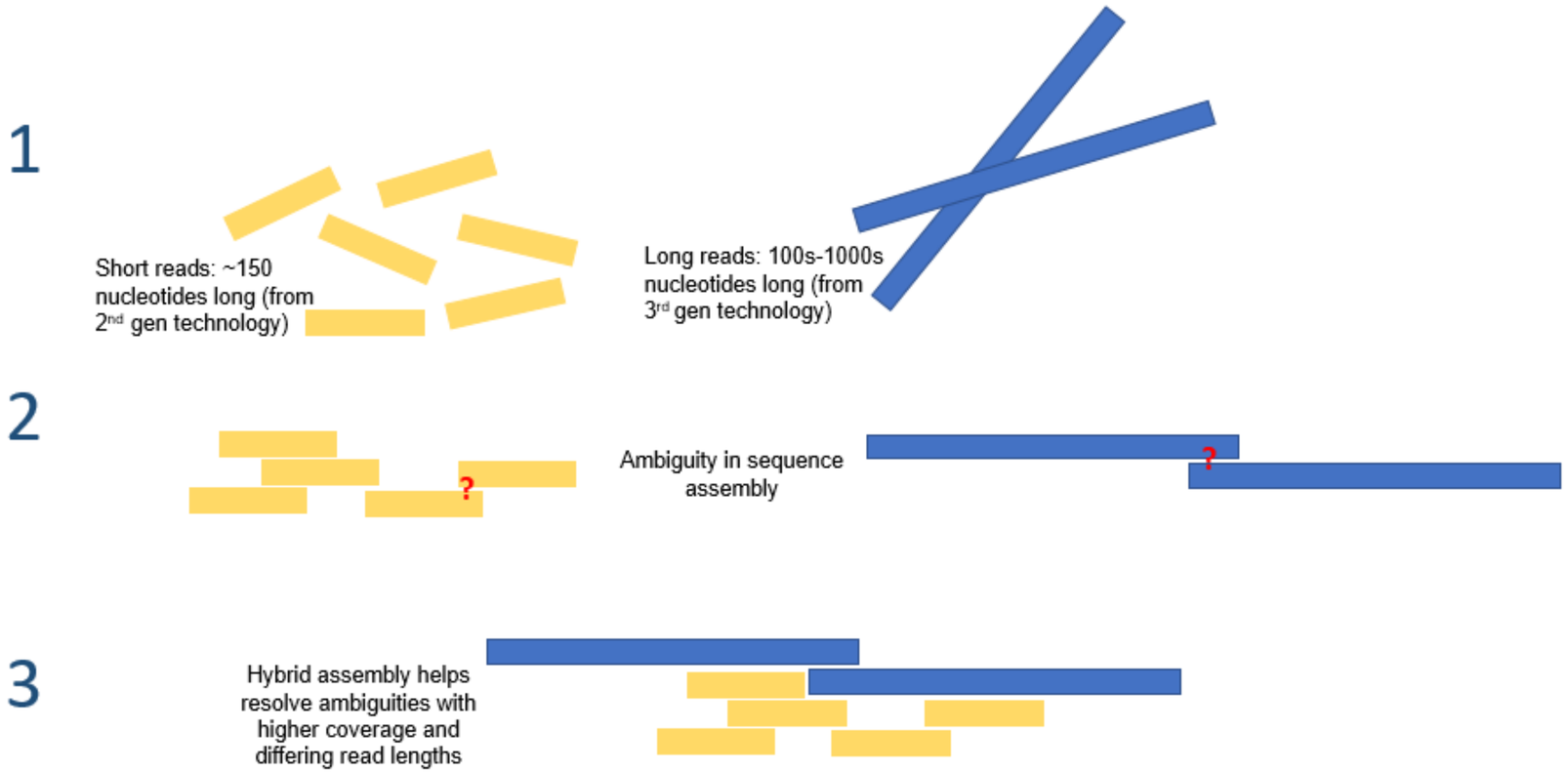
Small technical details:
On different
Illumina platforms
the i5 index read primer
binding site is on different sides
of the indexes

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



Long reads are particularly useful for genome assembly

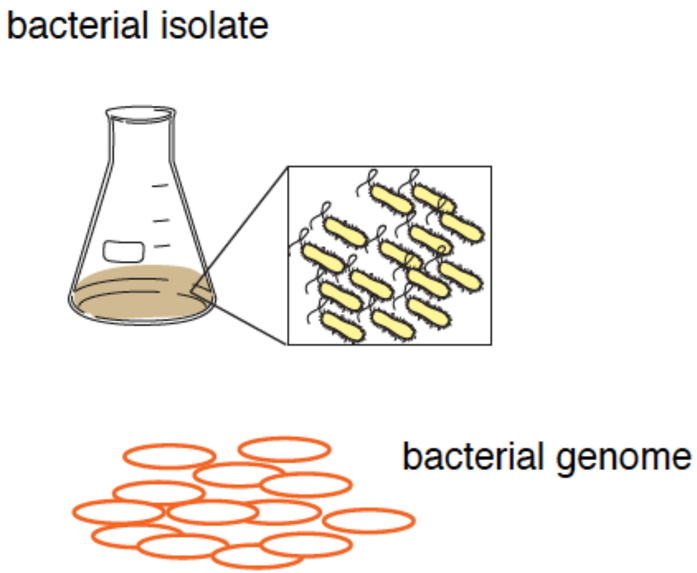
"Hybrid" assembly combines short + long reads and takes advantage of the low error rate of short reads and the long-range structural information of long reads

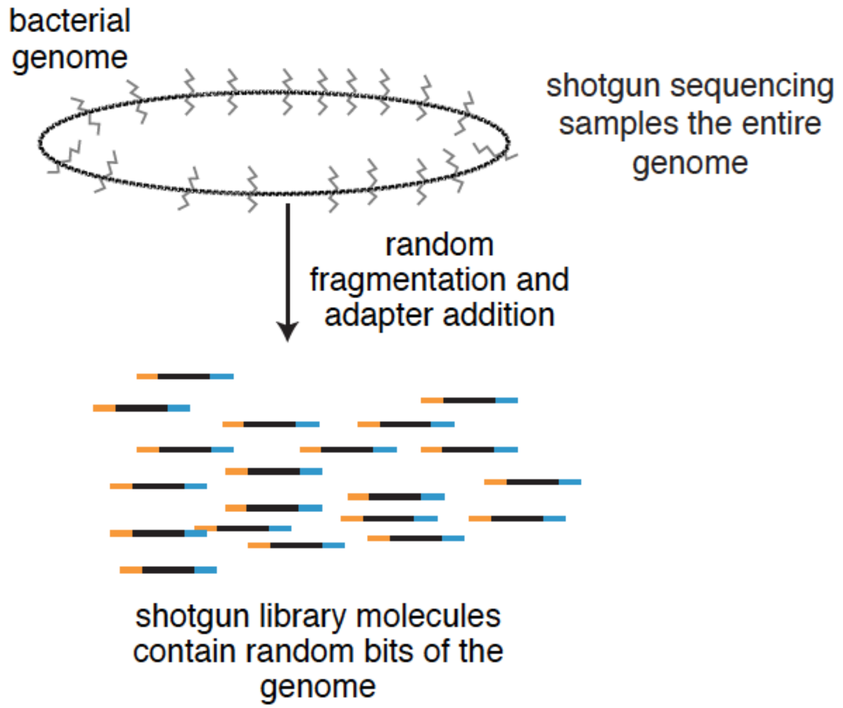


By Nivretta Thatra - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=56772145

The type of nucleic acid you start with and how you do library prep determines what type of sequencing you'll be doing

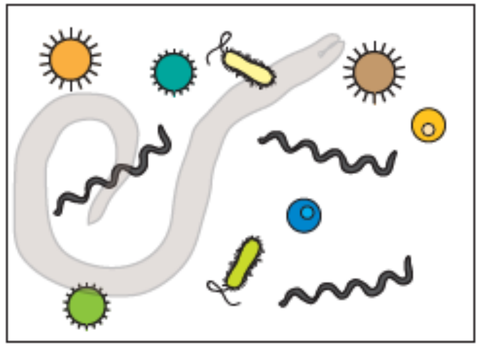
For instance, if you make a 'shotgun library' from a single organism, you're doing whole genome sequencing (WGS)





Metagenomic sequencing involves sequencing of genomes from more than one organism

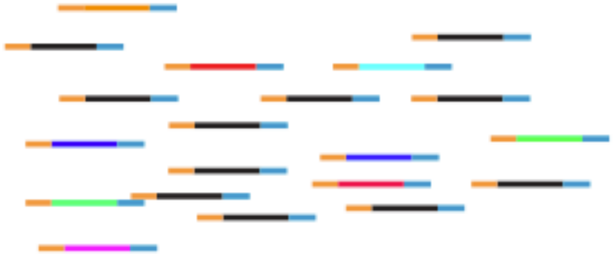
soil community



soil ‘metagenome’

metagenome

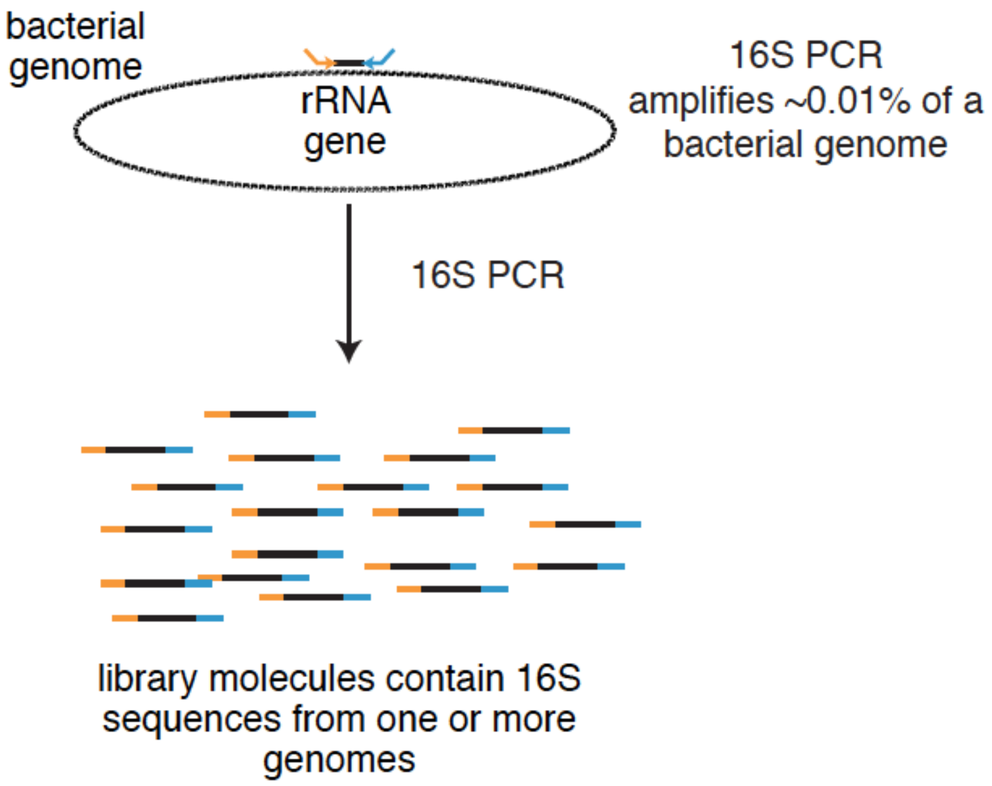
random
fragmentation and
adapter addition



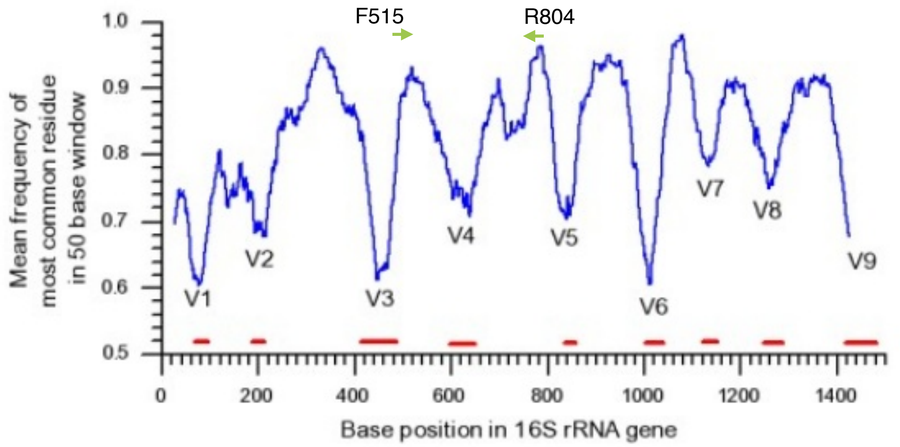
shotgun library molecules
contain random bits of all of
the genomes

Microbiome sequencing often means 16S rRNA gene sequencing

16S sequencing is one type of ‘amplicon sequencing’

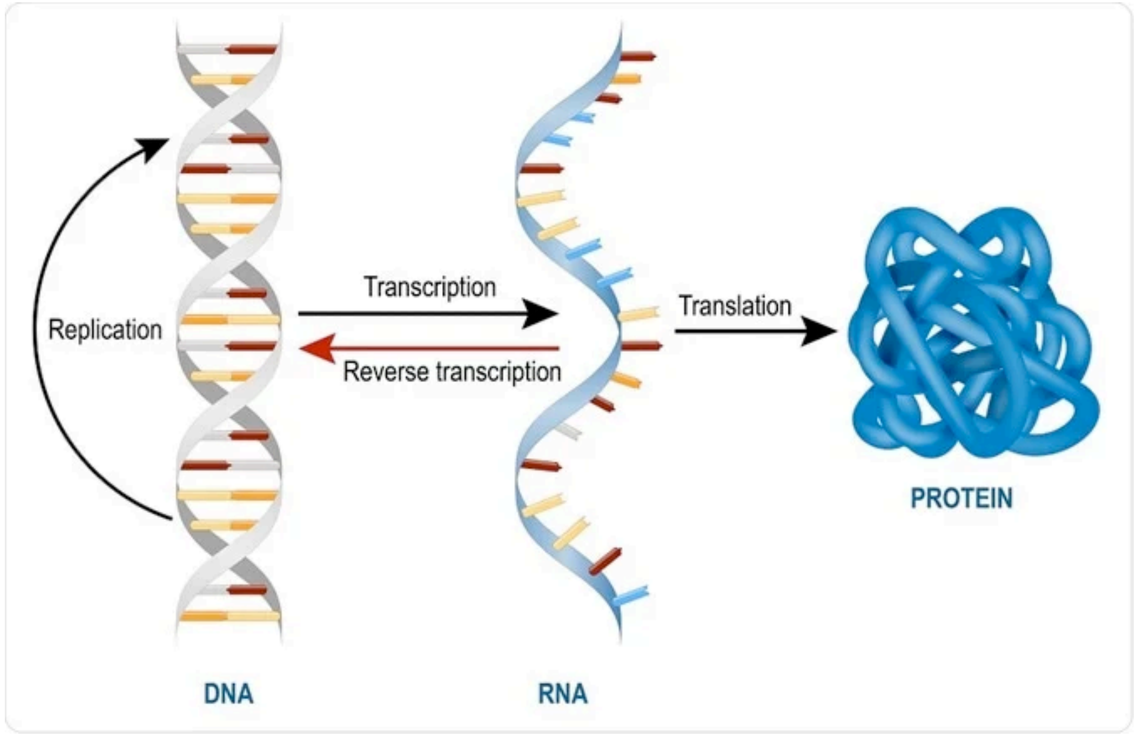


16S rRNA genes have highly conserved regions flanking variable regions



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

The raw data produced by sequencers is in FASTQ format

FASTQ files contain the read sequences and a quality score for each baseball

4 lines for each read:

1. starts with @ followed by sequence ID
2. The sequence
3. Starts with +. Sometimes repeats header
4. The quality score

Header Sequence Quality

```
@HWI-ST227:389:C4WA2ACXX:7:1204:2272:59979
GGAGGAAGGTCCTCGCTCCTCTTTCATATAAGGGAAATGGCTGAAT
+
FFFFHHHHHHJIJJJJJJJJIJJJIGIGIGGIJJIJIJJJJJJIII
@HWI-ST227:389:C4WA2ACXX:7:1205:15214:42893
GAGGATCCCAGGGAGGAAGGTCCTCGCTCCTCTTTCATCTAAGGGA
+
12BAFB?A:3<AE1@<FF;1*@EG*)?0?DBD>9BF9B*?######
@HWI-ST227:389:C4WA2ACXX:8:2208:2467:44624
AAAGAGGAGAGAGGACCATCCTCCCTGGGATCCTCAGAAGTCTACT
+
BDDA:DB?2AA@FC>F?EEGC<FED>GFD;?GBB?<?F99*/9?9?
```

https://biocorecrg.github.io/RNAseq_course_2019/fastq.html

FASTQ quality scores are encoded by a single character

The single character encodes a value from ~0-41

Using 1 character

- Keeps Q scores synchronized with the corresponding base
- Saves disk space

Phred quality score.

- $Q=-10 \log_{10} p$
- p = probability that the corresponding base call is incorrect
- Example: p = 0.001 means a quality of 30

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
0.........................26...31.........41
```

https://biocorecrg.github.io/RNAseq_course_2019/fastq.html

FASTQ quality scores are encoded by a single character

The single character encodes a value from ~0-41

Using 1 character

- Keeps Q scores synchronized with the corresponding base
- Saves disk space

Phred quality score.

- $Q=-10 \log_{10} p$
- p = probability that the corresponding base call is incorrect
- Example: p = 0.001 means a quality of 30

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
0.........................26...31.........41
```

https://biocorecrg.github.io/RNAseq_course_2019/fastq.html

FASTQ quality scores are encoded by a single character

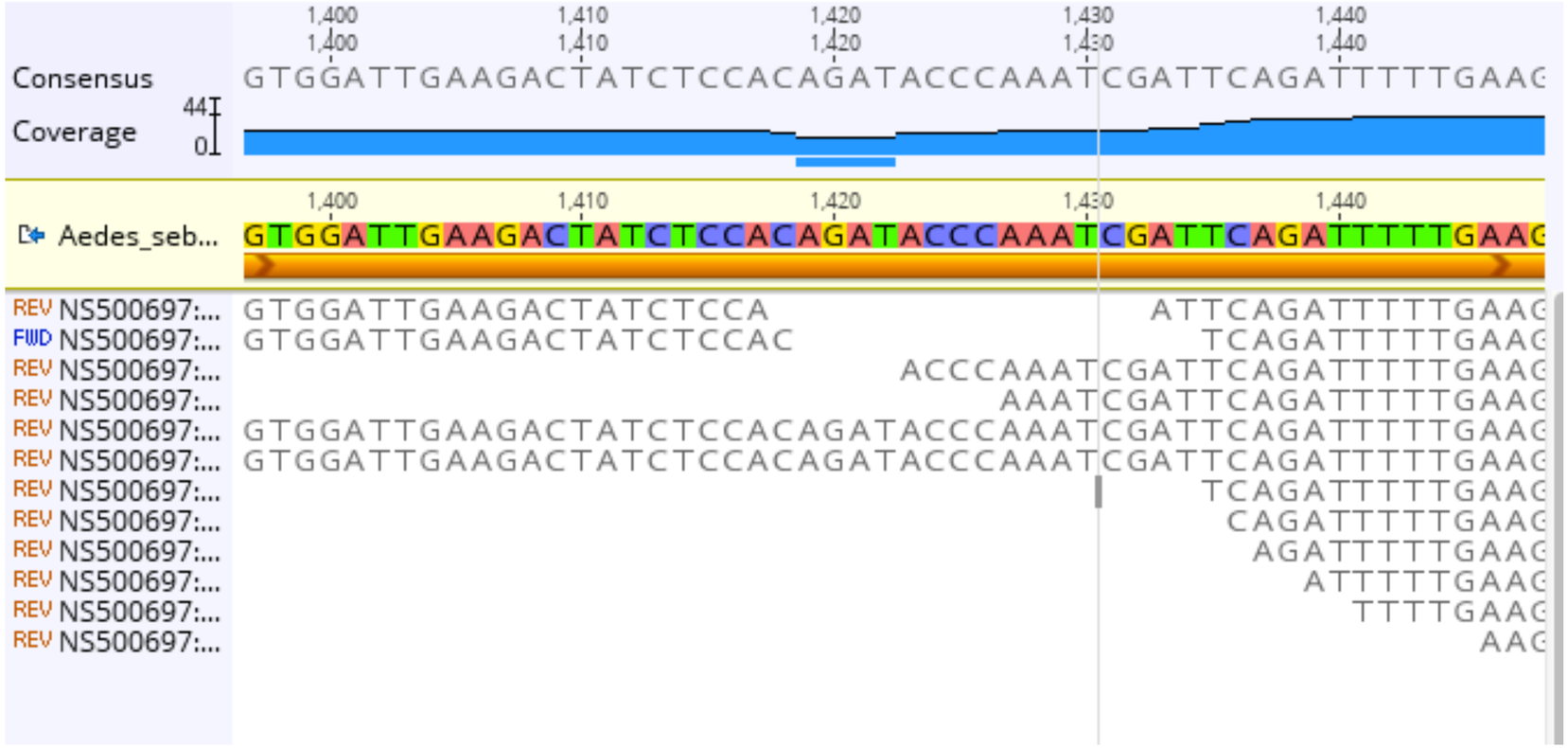
A quality value *Q* is an integer representation of the probability *p* that the corresponding base call is incorrect.

$$Q = -10 \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

https://learn.gencore.bio.nyu.edu/ngs-file-formats/quality-scores/

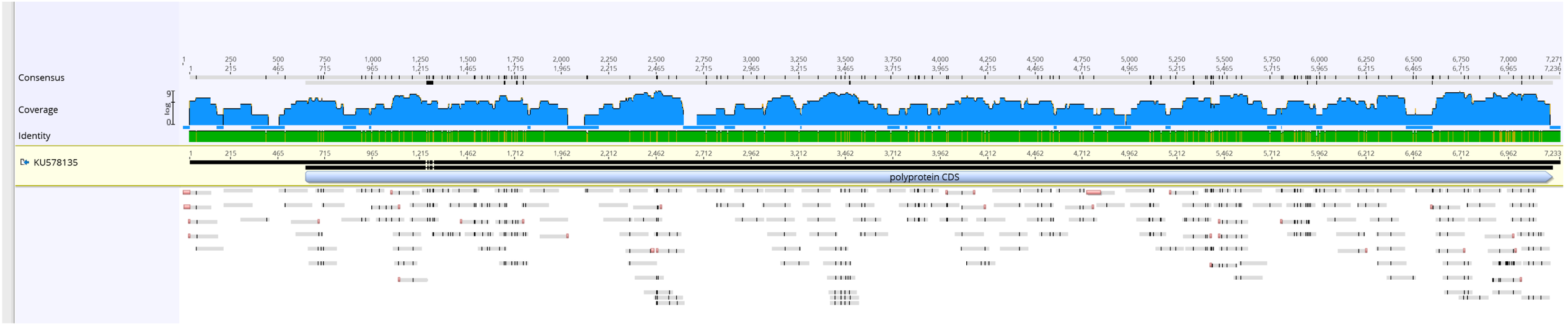
Coverage is the number of individual reads that support a particular nucleotide in an assembled (reconstructed) sequence or that align to a particular nucleotide in a reference sequence



this metric is often referred to as ‘depth’ or ‘depth of coverage’

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

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



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

hand foot and mouth disease

(Mayo clinic)