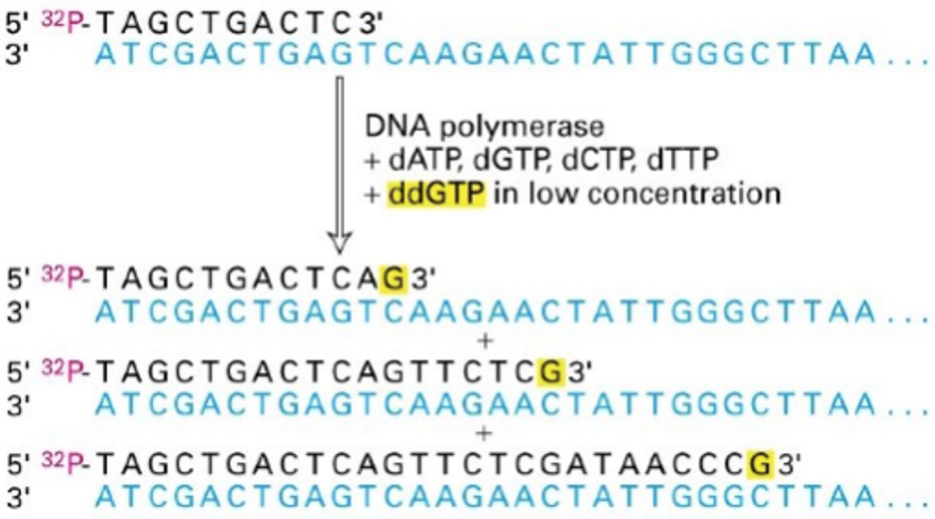
Long Read Sequencing

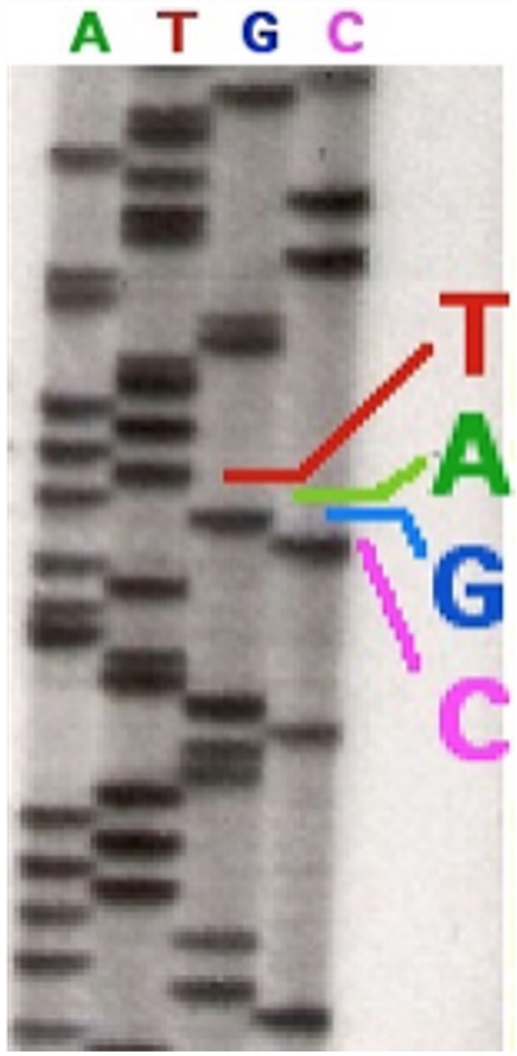
2023-03-07

How to sequence a human genome: Sanger method

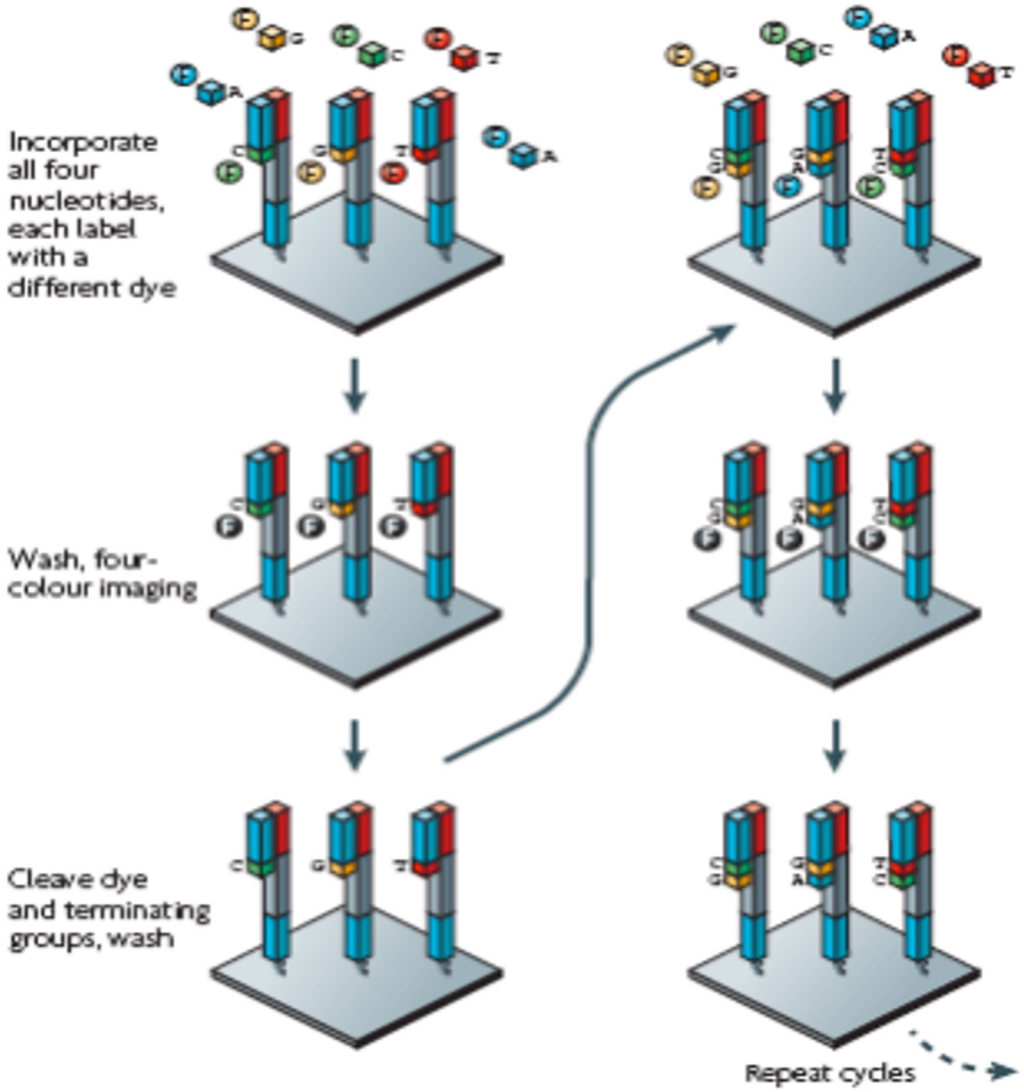
Key points:

1) sequencing by synthesis (not degradation)
2) primers hybridize to DNA
3) polymerase + dNTPS + ddNTP terminators at low concentration
4) 1 lane per base, visually interpret ladder

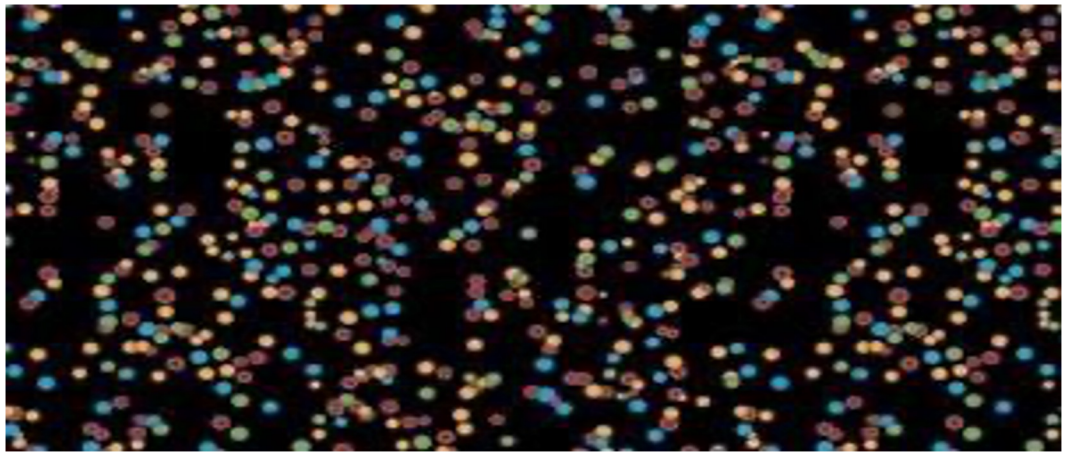




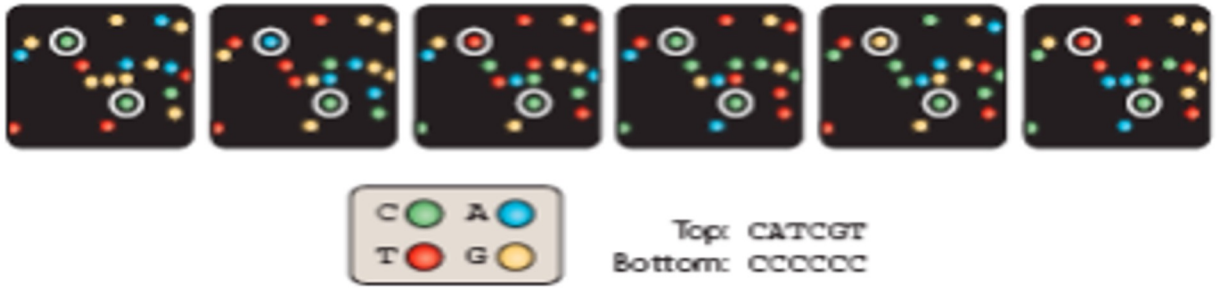
Illumina: Cluster amplification by "bridge" PCR



4 different images merged

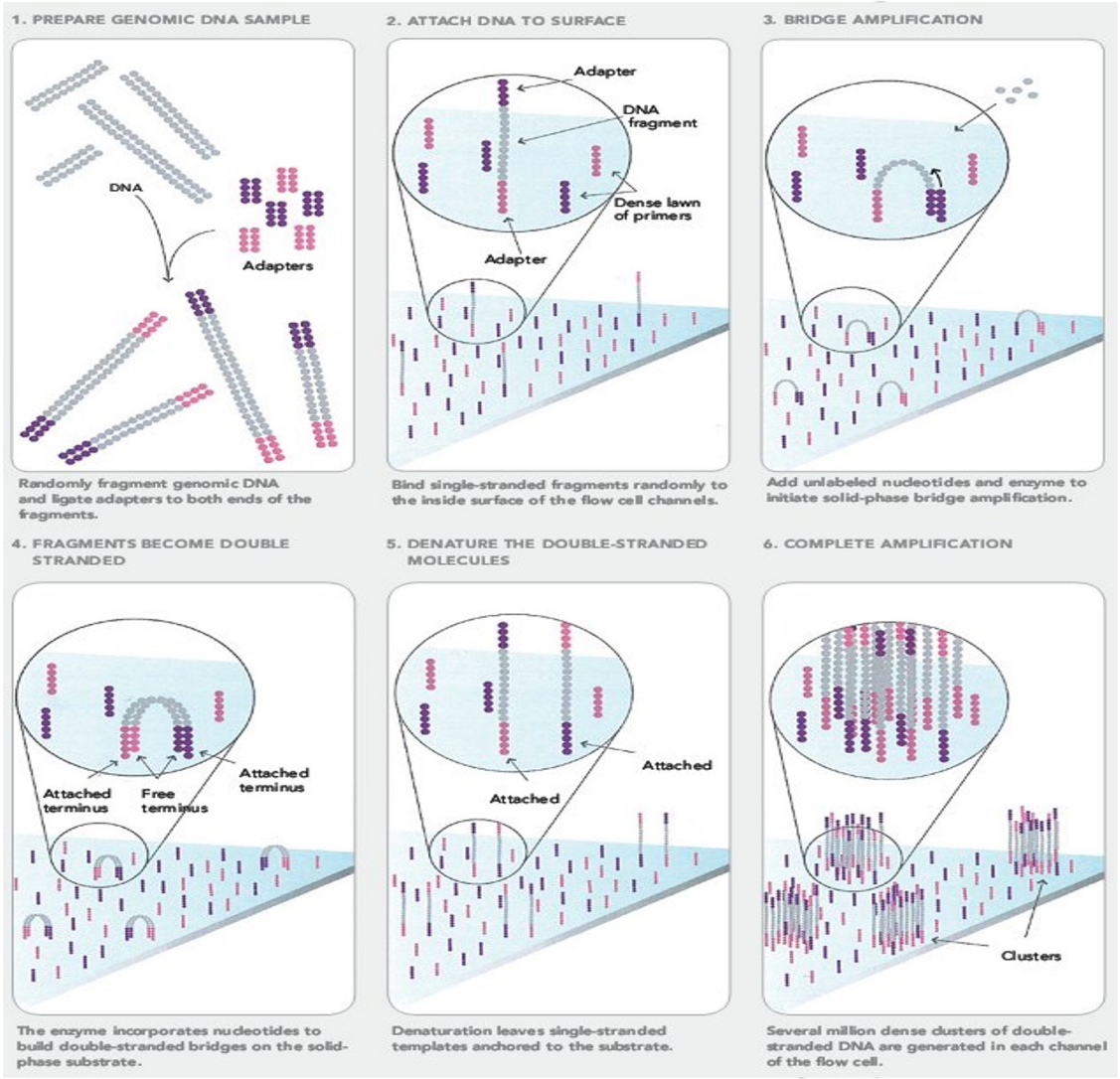


6 cycles w/ base-calling

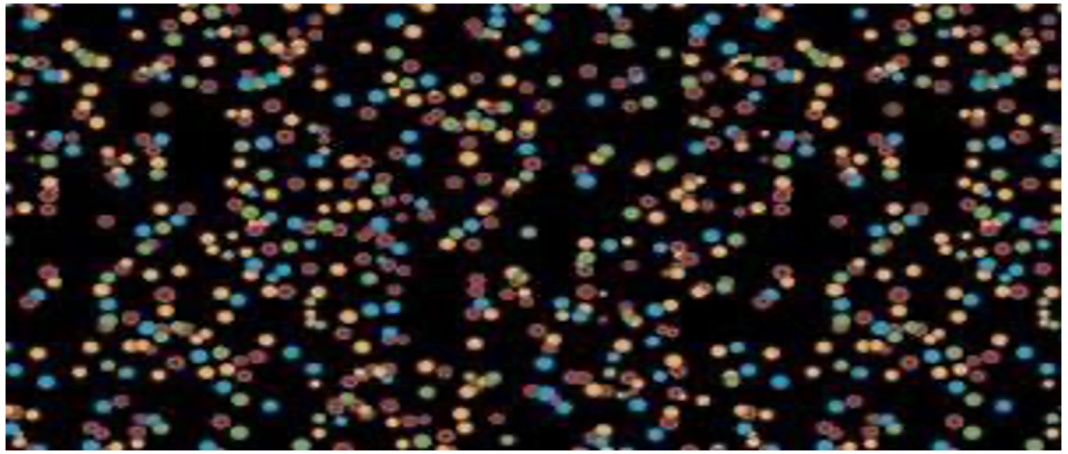


Slide from Ira Hall

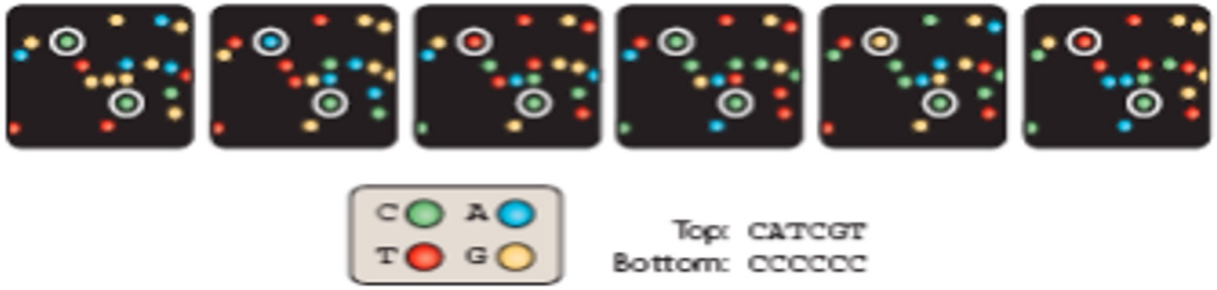
Illumina: Cluster amplification by "bridge" PCR



4 different images merged

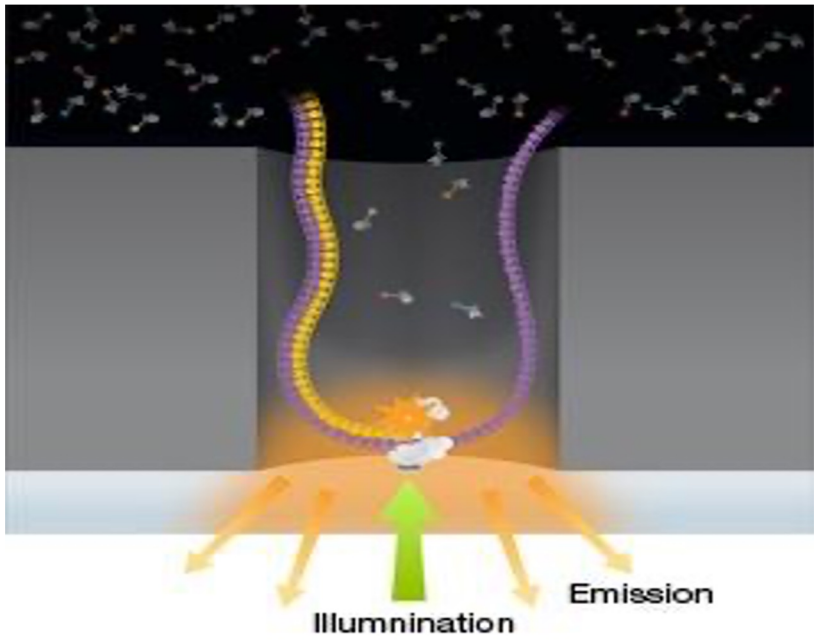


6 cycles w/ base-calling



Slide from Ira Hall

Pacific Biosciences



Key Points:

- 1 DNA molecule and 1 polymerase in each well (zero-mode waveguide)
- 4 colors flash in real time as polymerase acts
- Methylated cytosine has distinct pattern
- No _theoretical_ limit to DNA fragment length

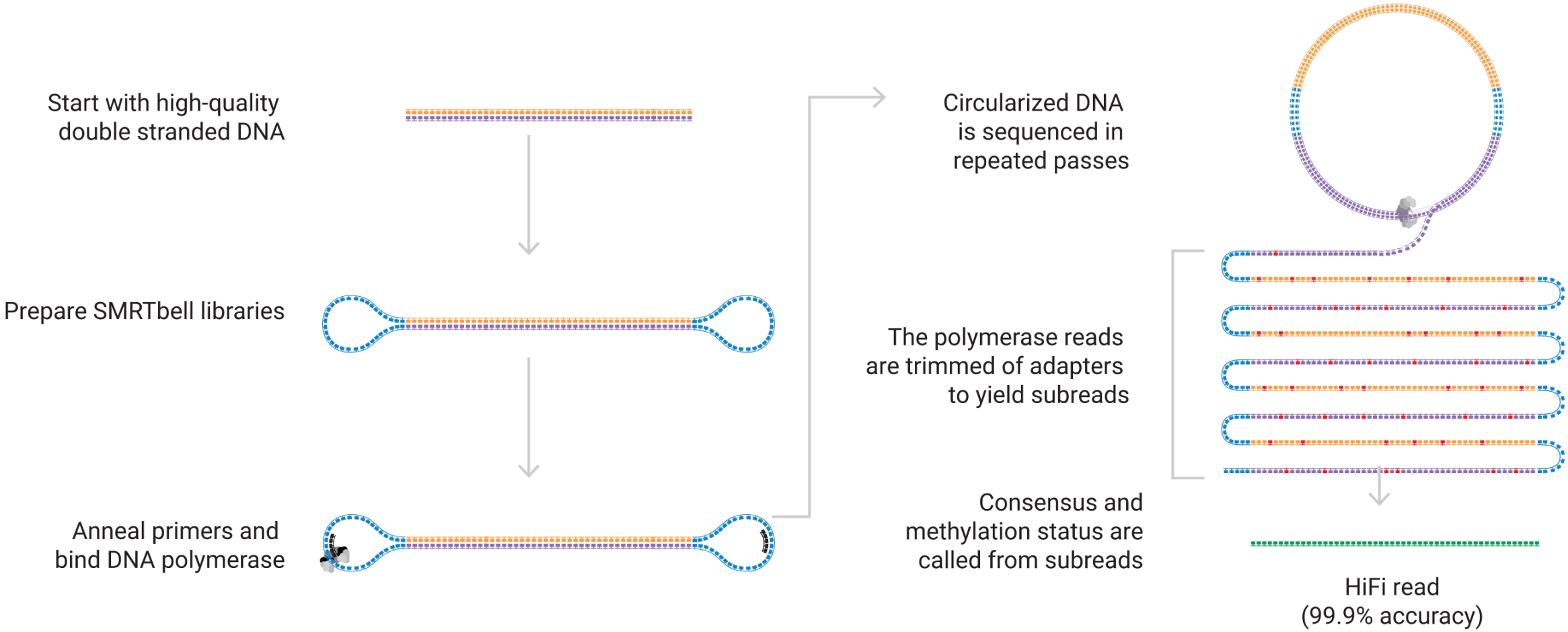
Caveats:

- higher error rate (1-2%)
- lower throughput : roughly 5 gigabases per run

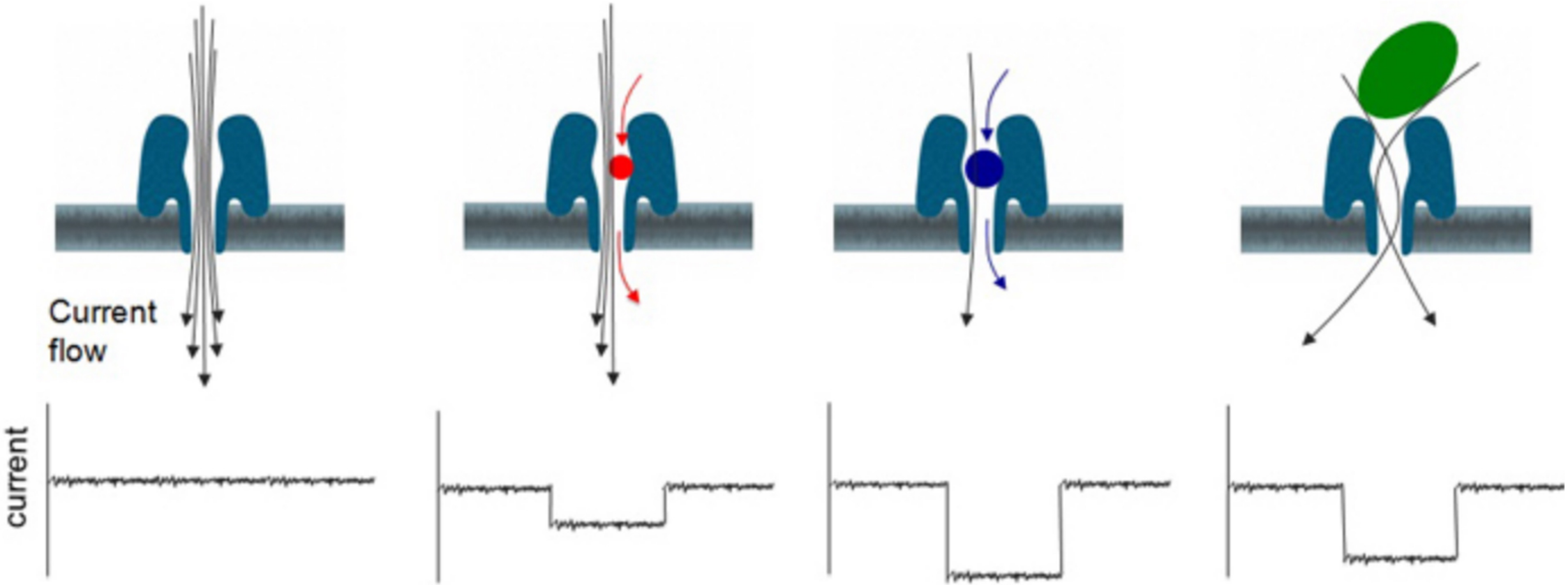
(this number is outdated!)

Slide from Ira Hall

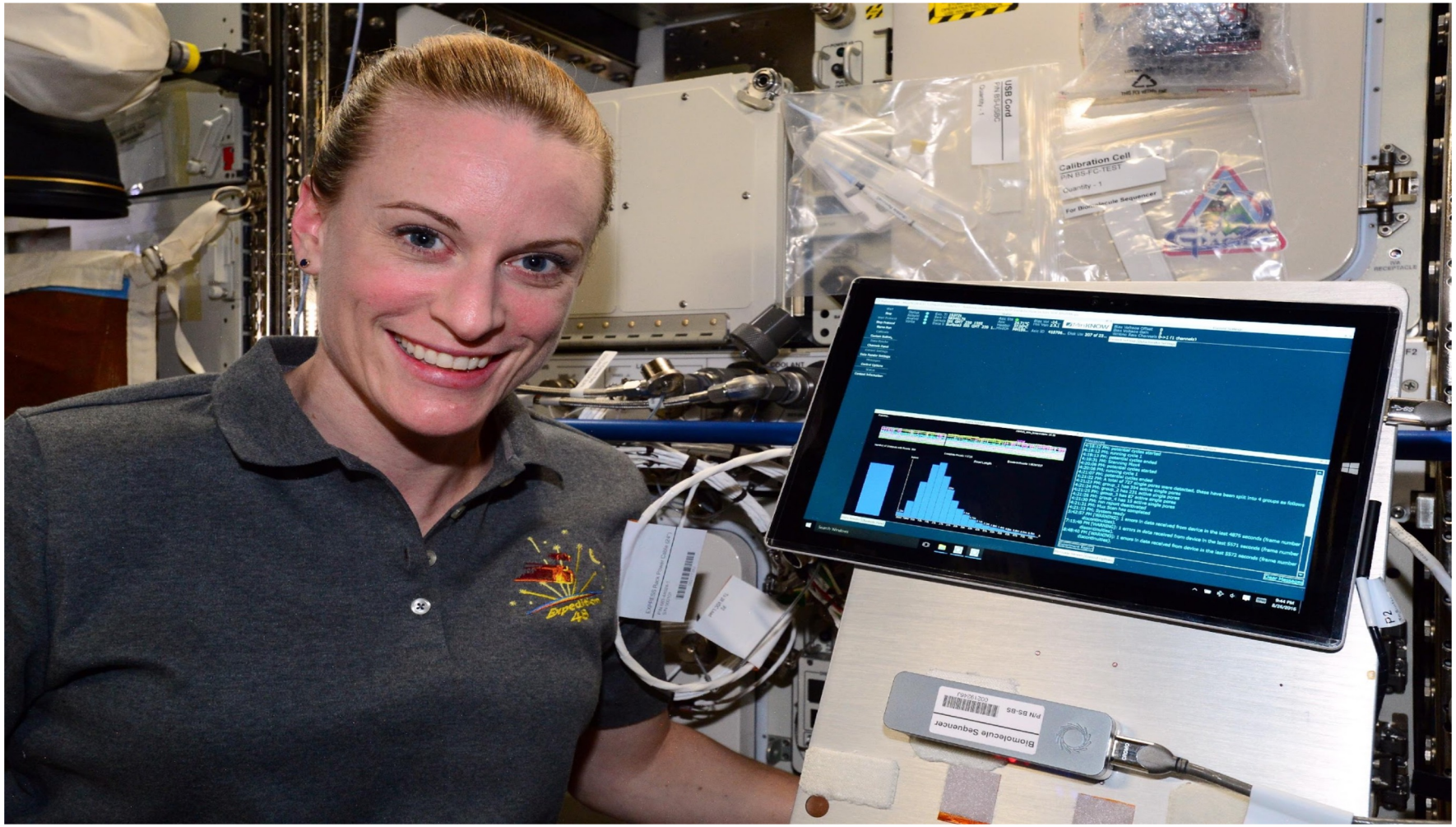
PacBio HiFi reads



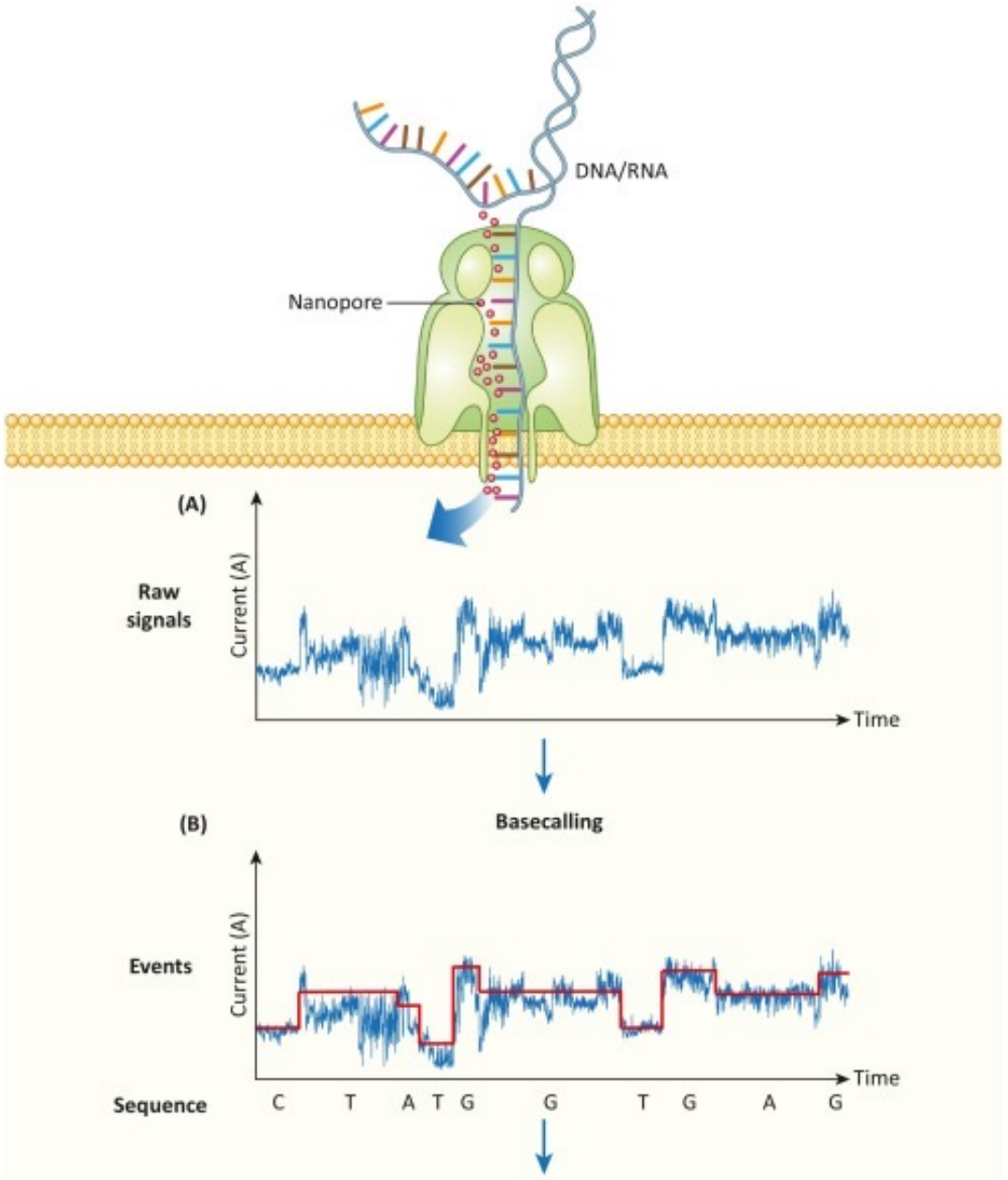
Oxford Nanopore Technologies

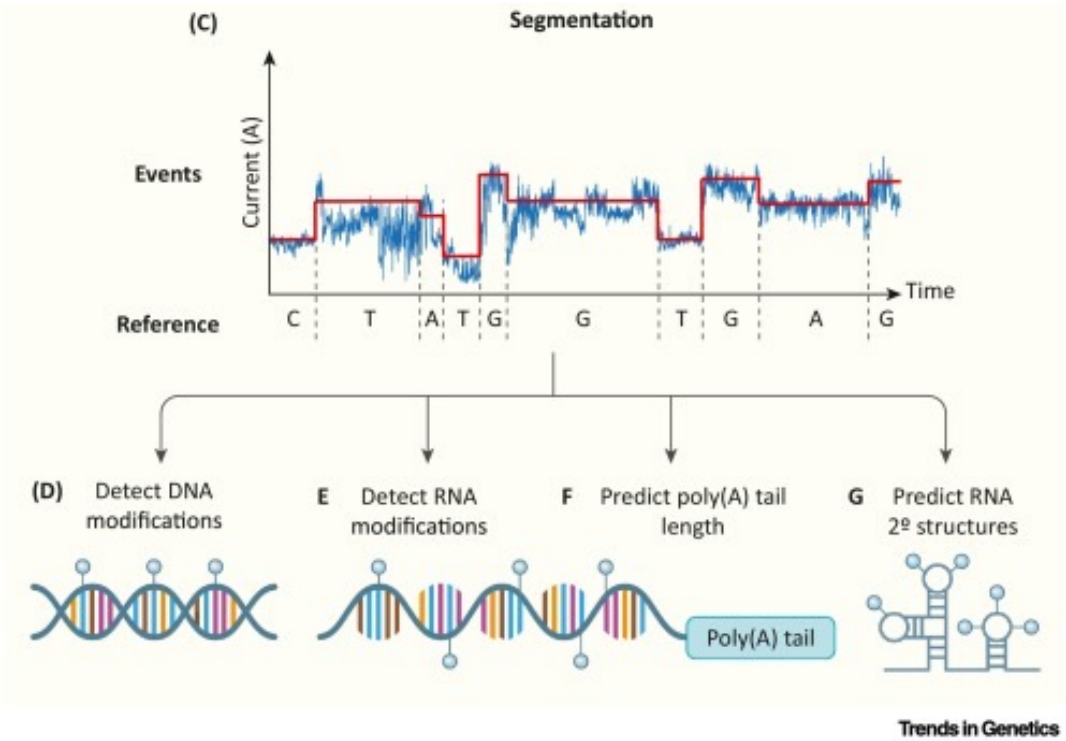


Nanopore sequencing is *extremely* portable

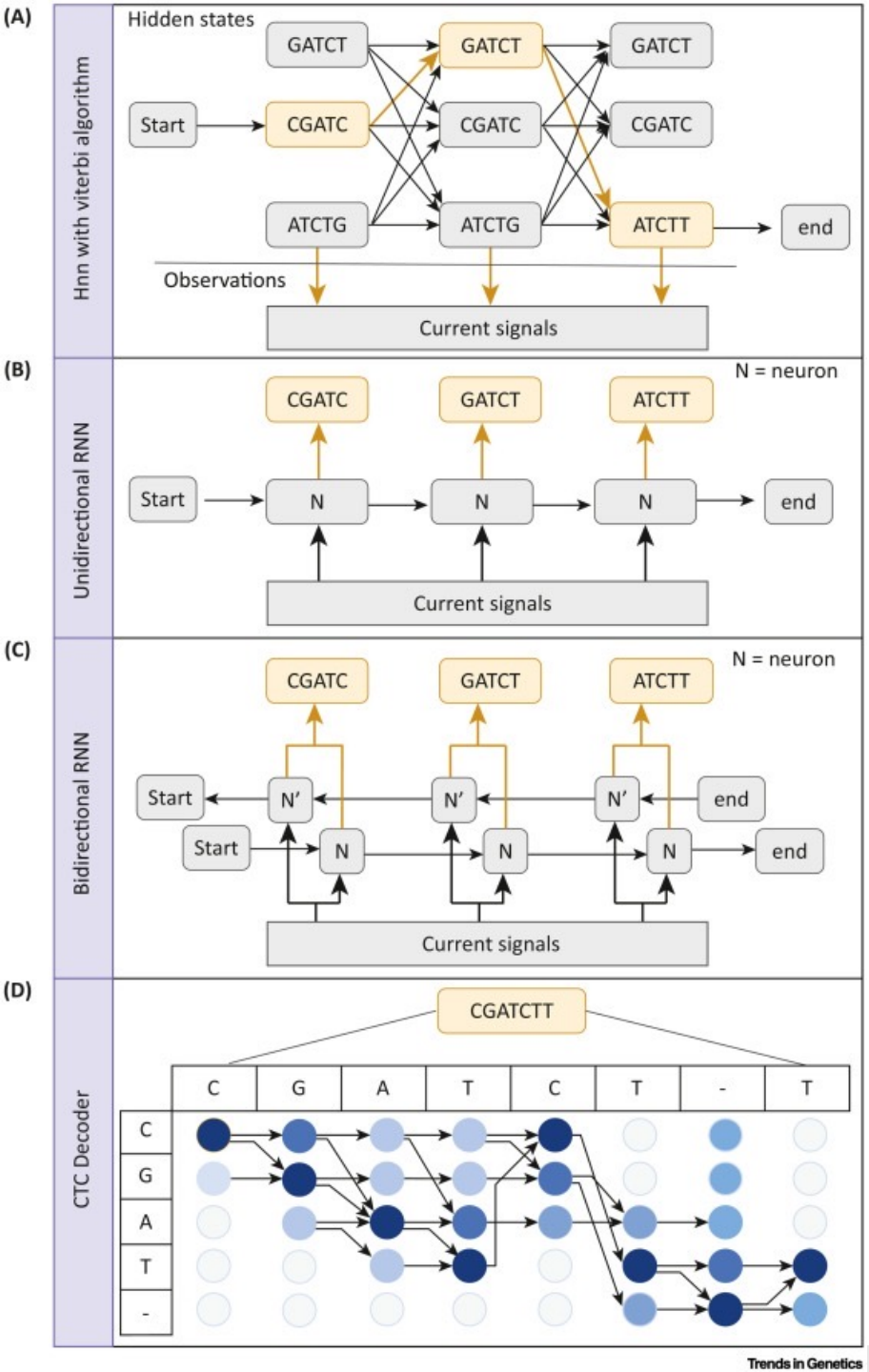


Kate Rubins sequencing DNA on the ISS





doi.org/10.1016/j.tig.2021.09.001



Trends in Genetics

Neural networks to translate signal into base calls

- Guppy (many versions)

- Dorado (v0.1, eventual guppy replacement)

- many others

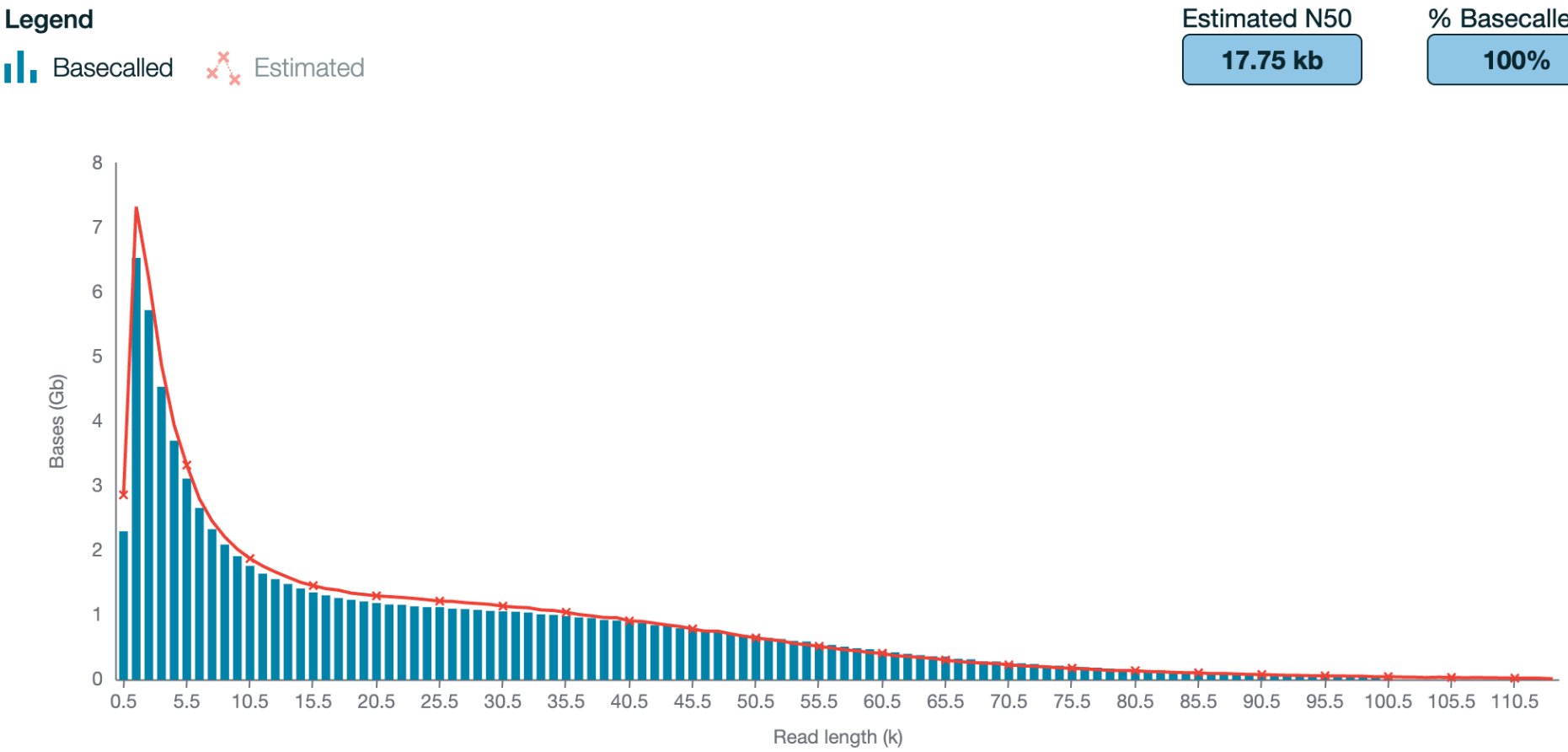
Practically, that means that we can't throw away our raw signal intensities. (1 Tb or more per run)

doi.org/10.1016/j.tig.2021.09.001

Alignment/pre-processing

- minimap2 algorithm is standard at this point
- has some options we're still exploring
  - input of known splice junctions for refinement of RNA/cDNA alignment

What does the data look like?

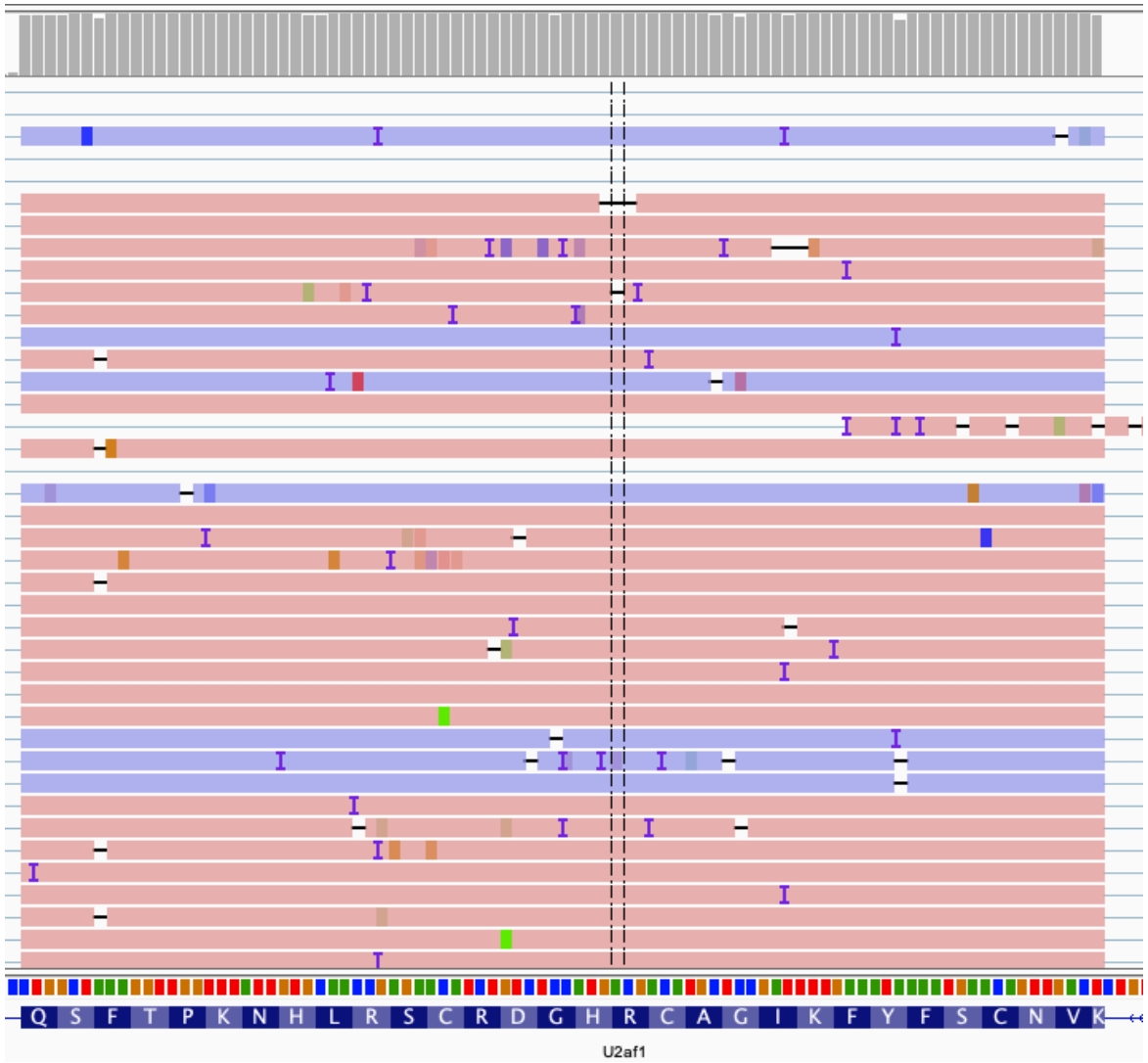


their relative amounts.

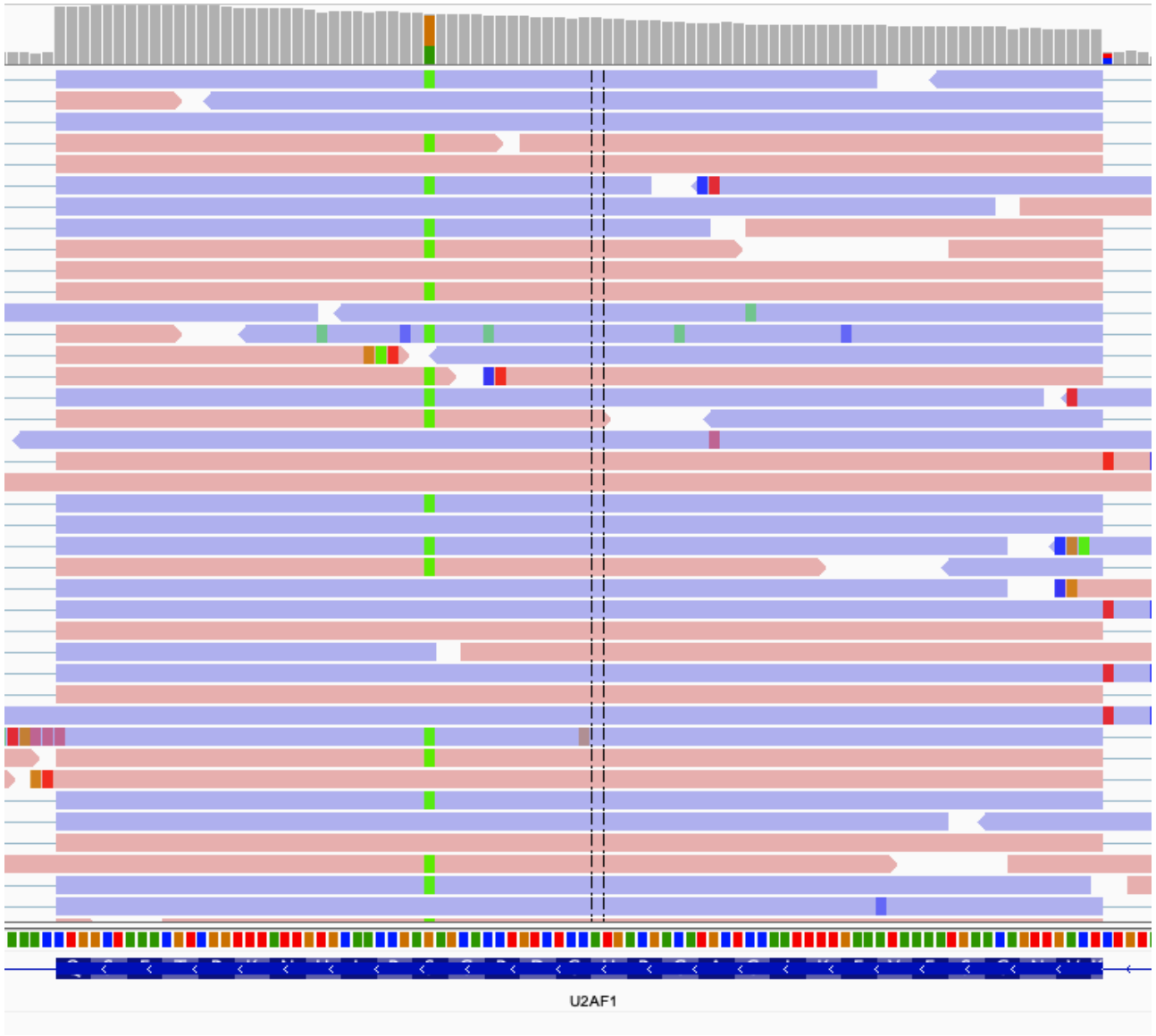
Read length (kb)	Aggregated reads (Mb)
100 - 164	886.98
164 - 228	36.06
228 - 292	4.02
292 - 344	0.35

Genomic DNA – standard prep

What does the data look like?



Long-read ONT ~5% base error rate



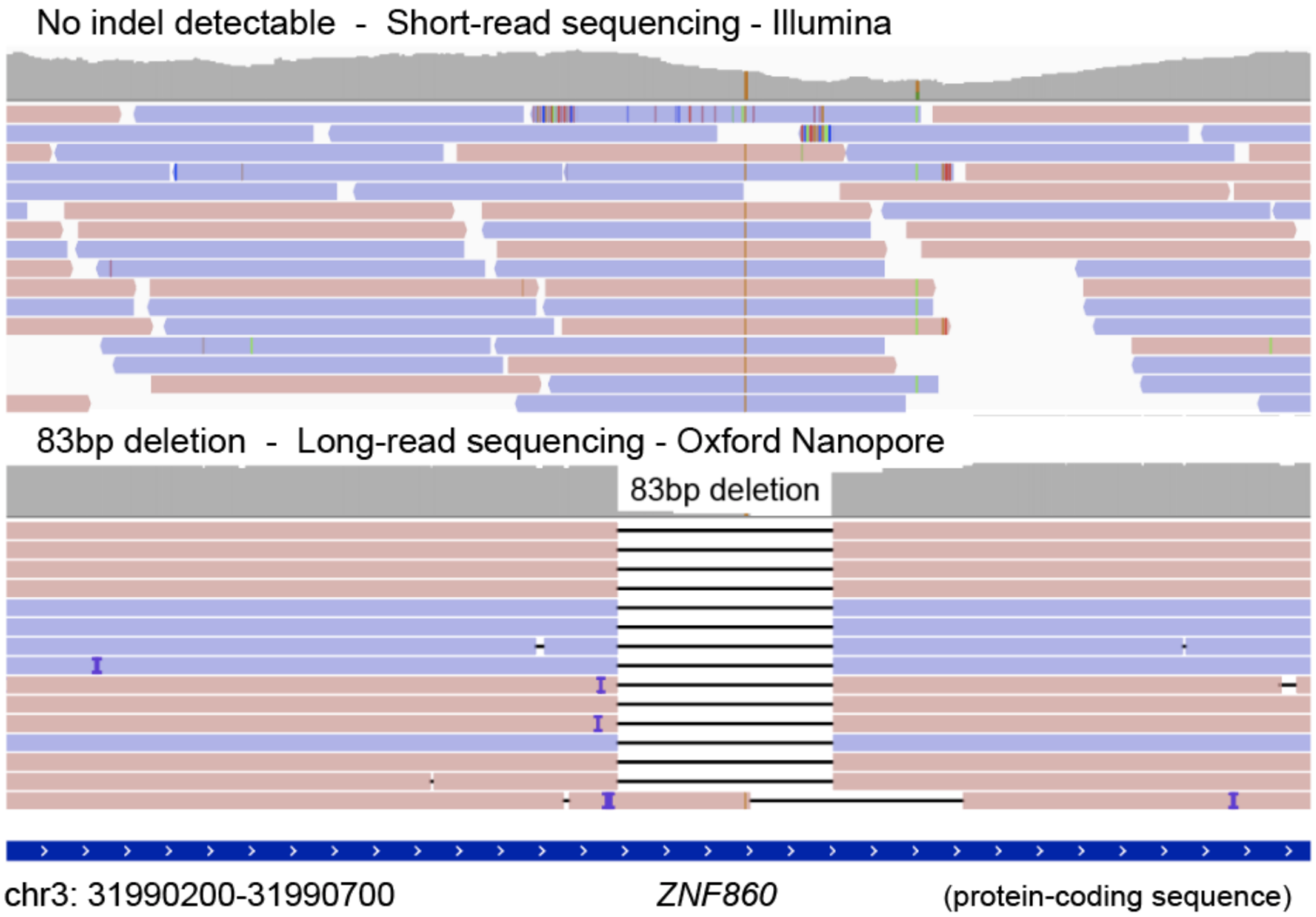
Short-read Illumina ~0.3% base error rate

Genomic DNA advantages

Figure 1: Blue-labeled genomic regions are accessible to long reads but not short, and have functional annotations (e.g. genes or enhancers)

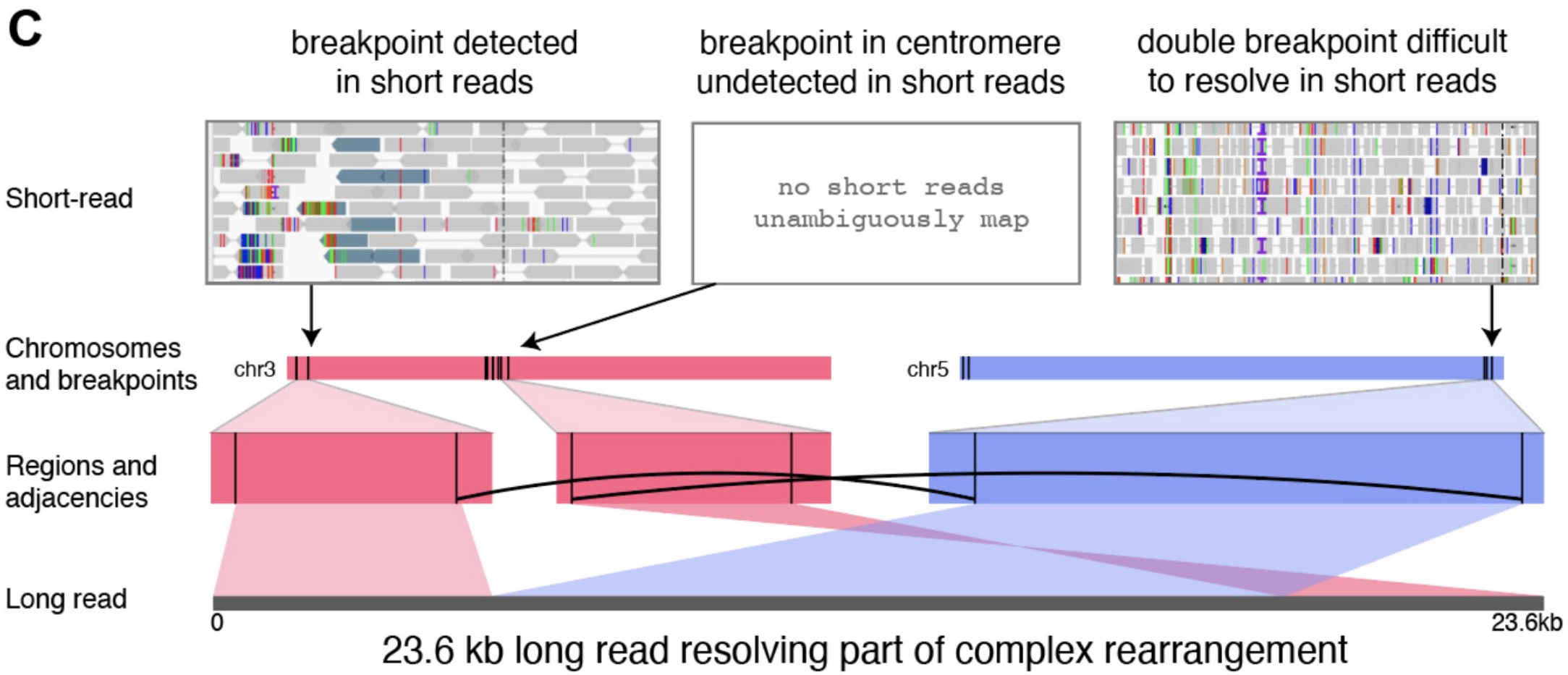
Dave Spencer

Genomic DNA advantages



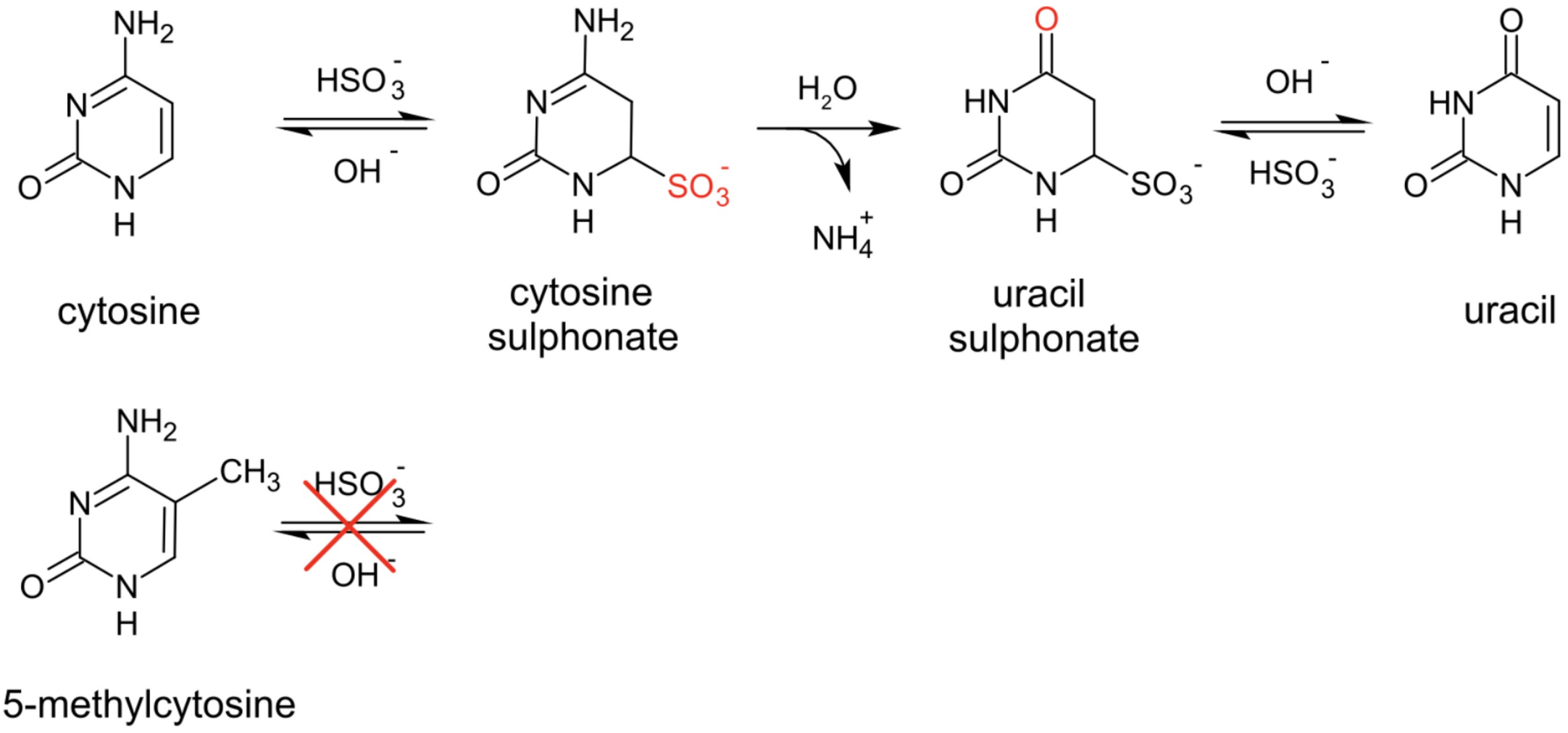
Haley Abel

Genomic DNA advantages



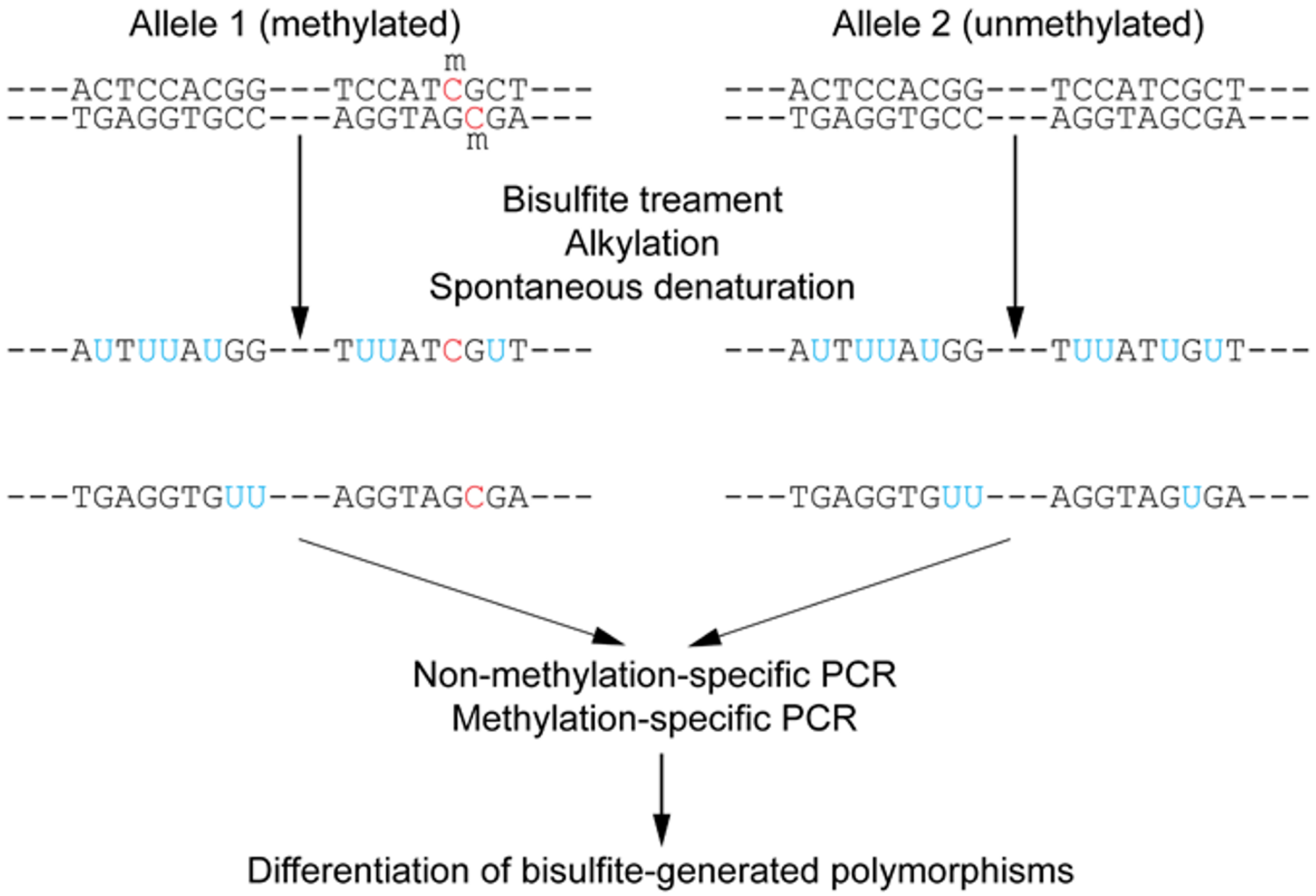
Haley Abel

Bisulfite sequencing

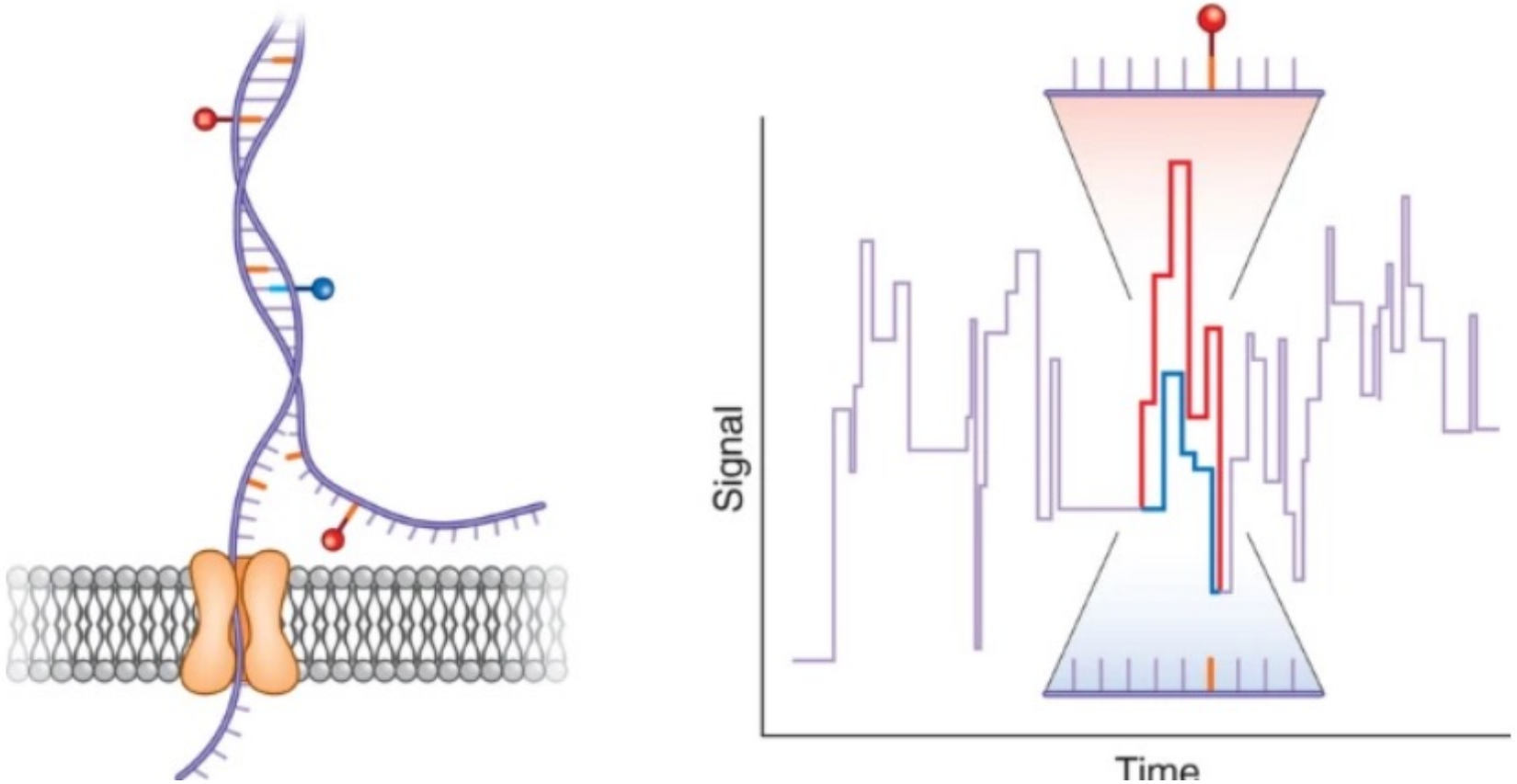


https://en.wikipedia.org/wiki/File:Bisulfite_conversion.svg

Bisulfite sequencing

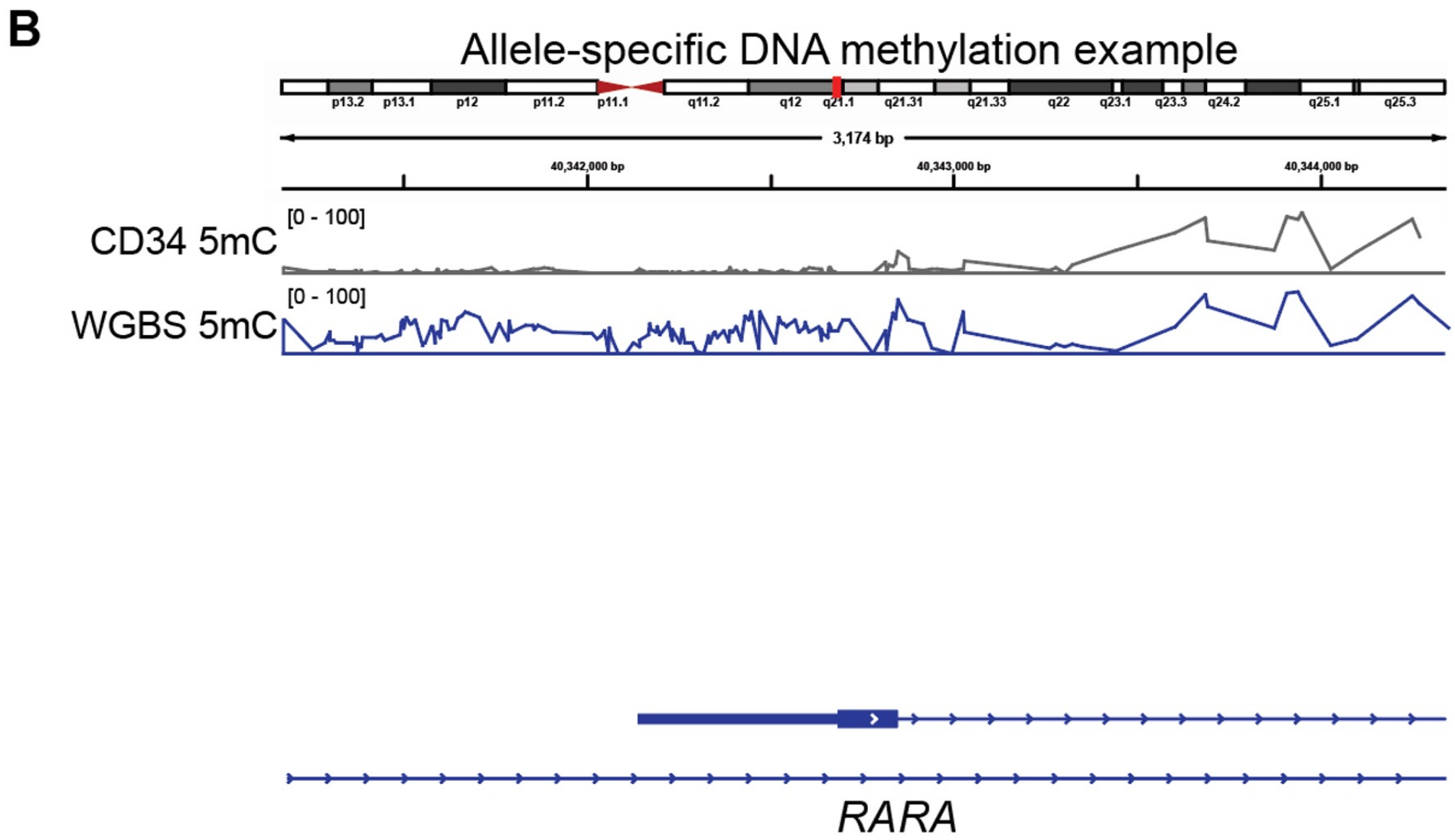


https://en.wikipedia.org/wiki/File:Wiki_Bisulfite_sequencing_Figure_1_small.png



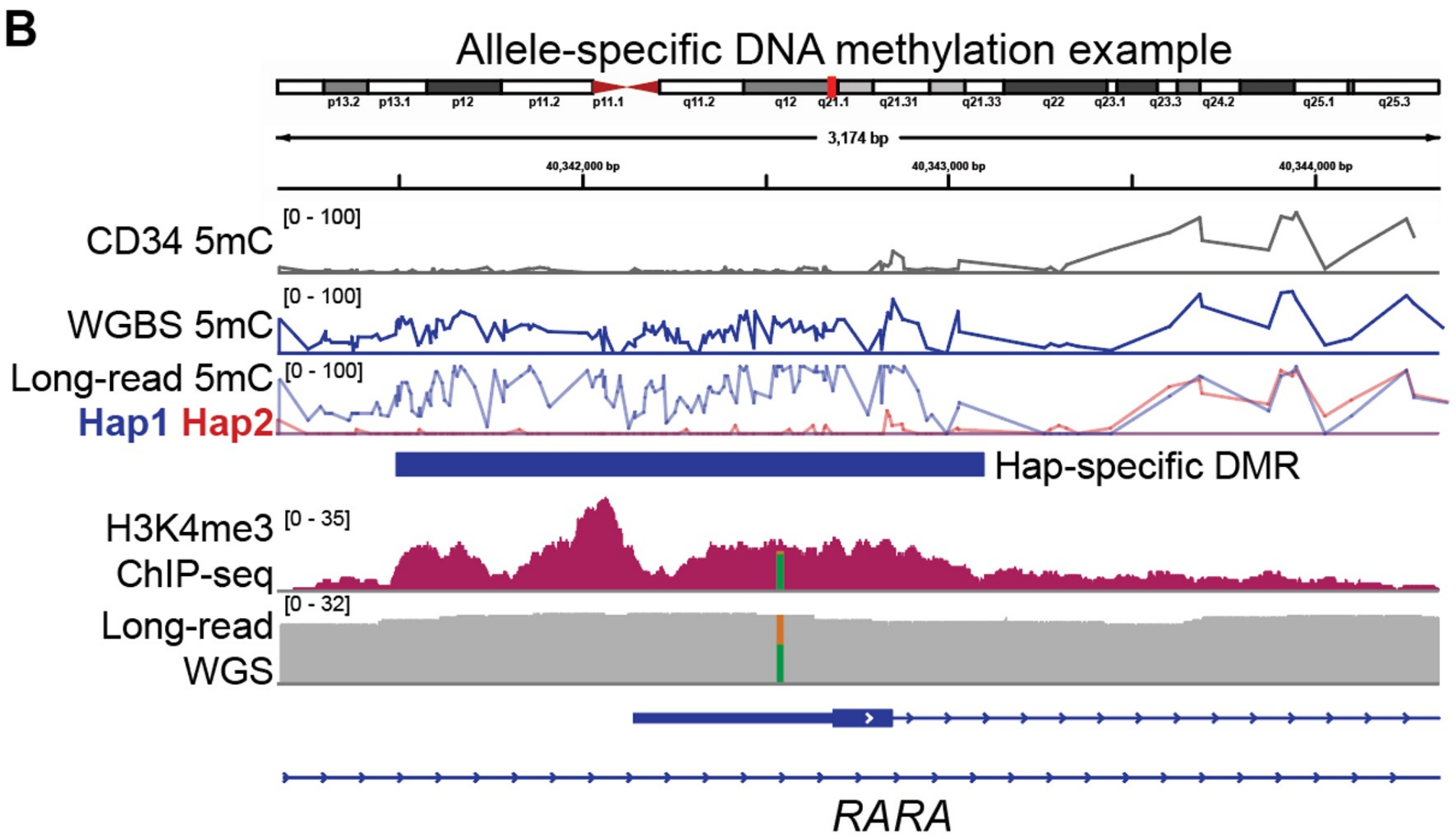
Can be used for 5mC as well as m6A in direct RNAseq

Genomic DNA advantages



Dave Spencer

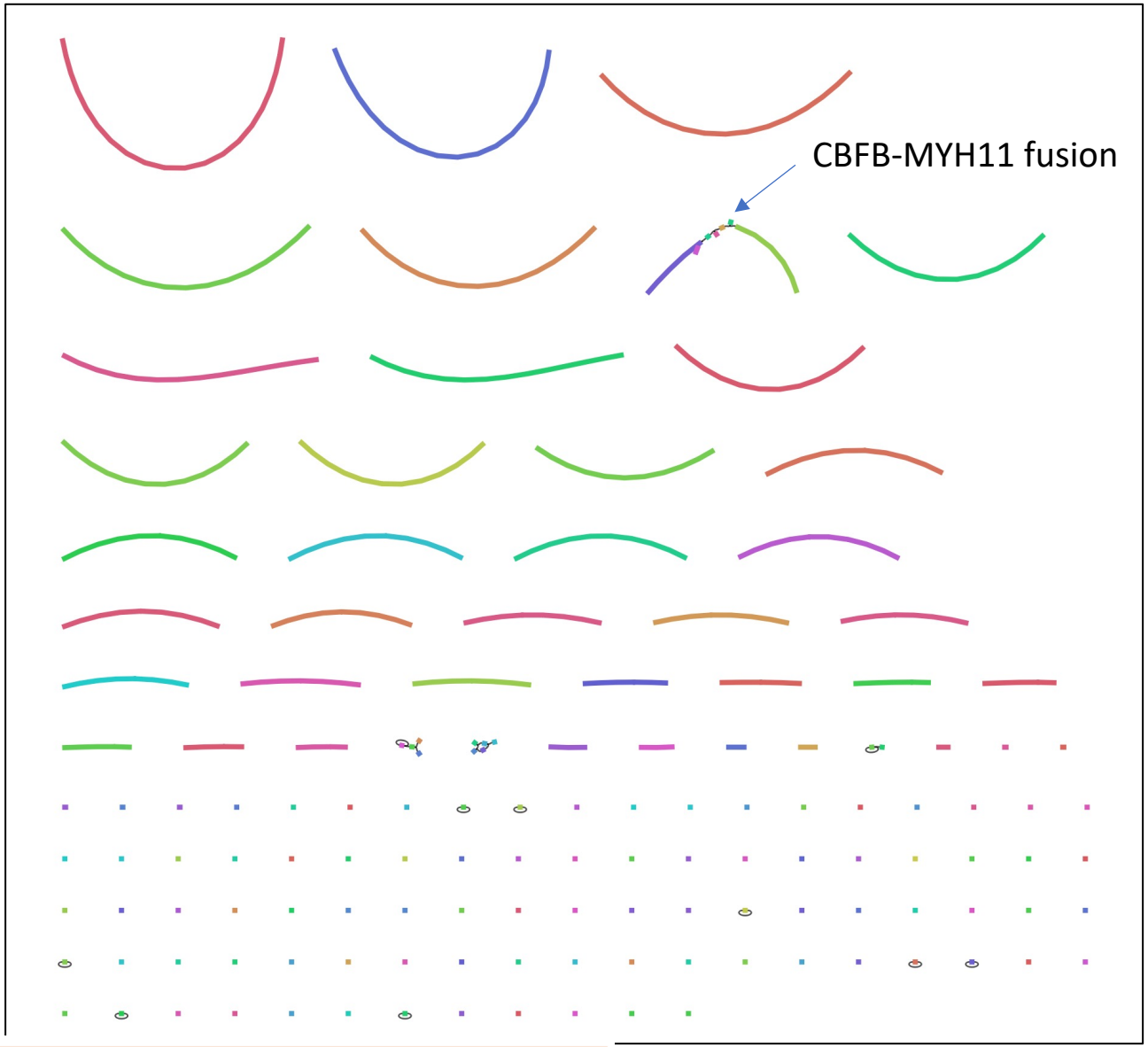
Genomic DNA advantages



Dave Spencer

Genomic DNA advantages

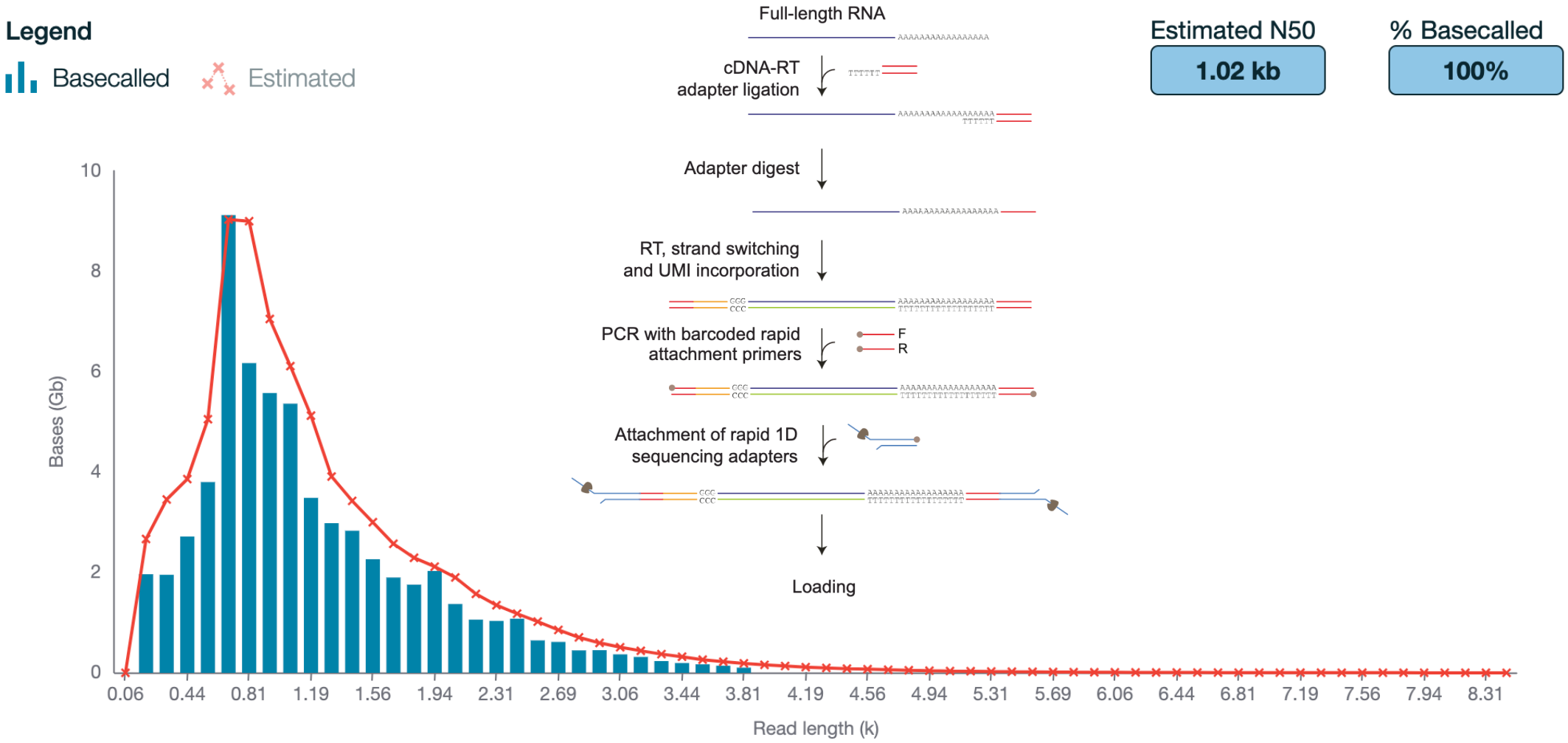
- Assembly of personal genomes



sample	hap	metric	value
Presentation	hap1	Number_of_contigs	402
Presentation	hap1	N50	88.3 Mbp

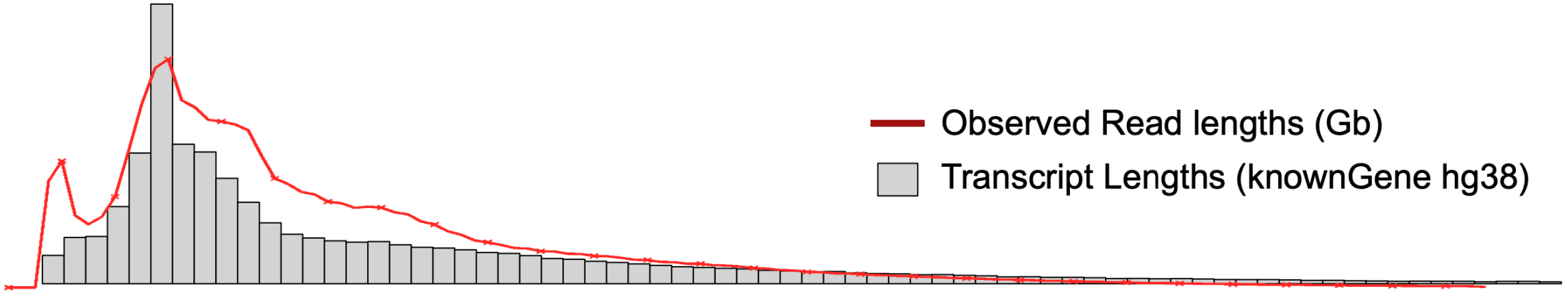
Dave Spencer, Haley Abel

What does the data look like?

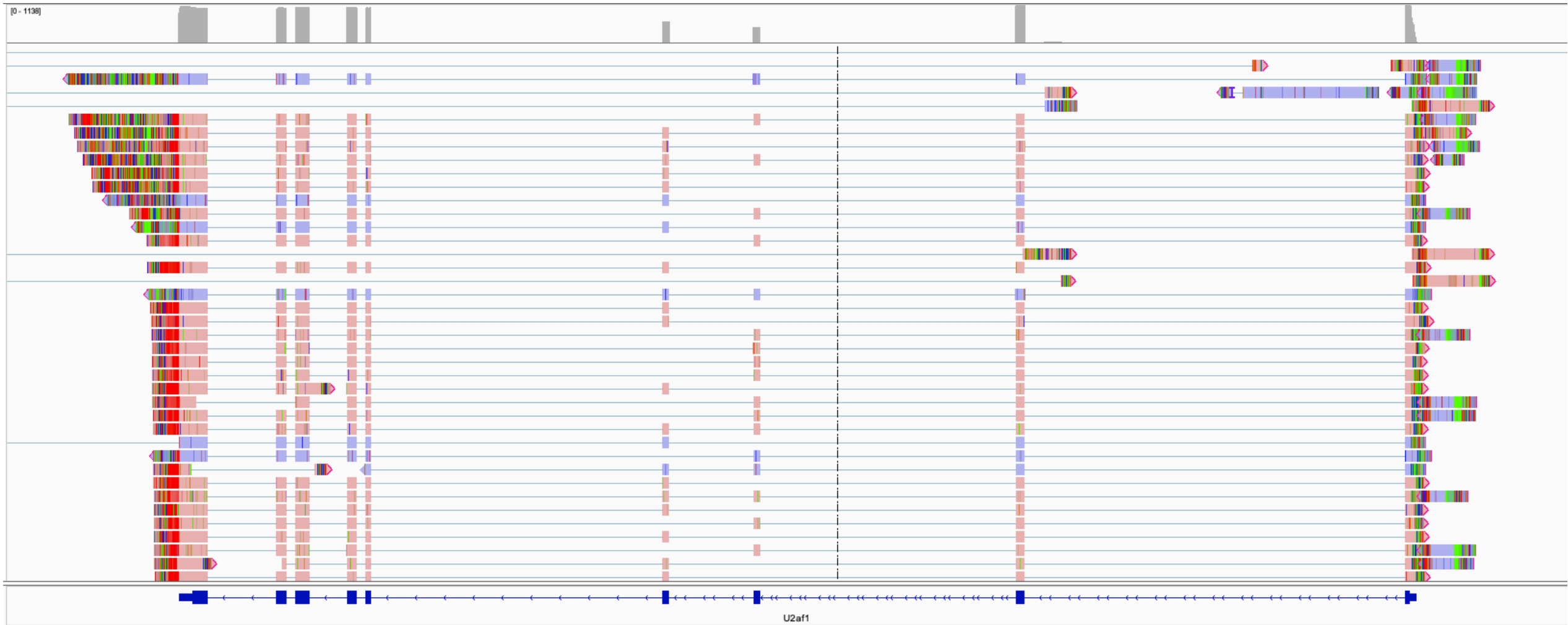


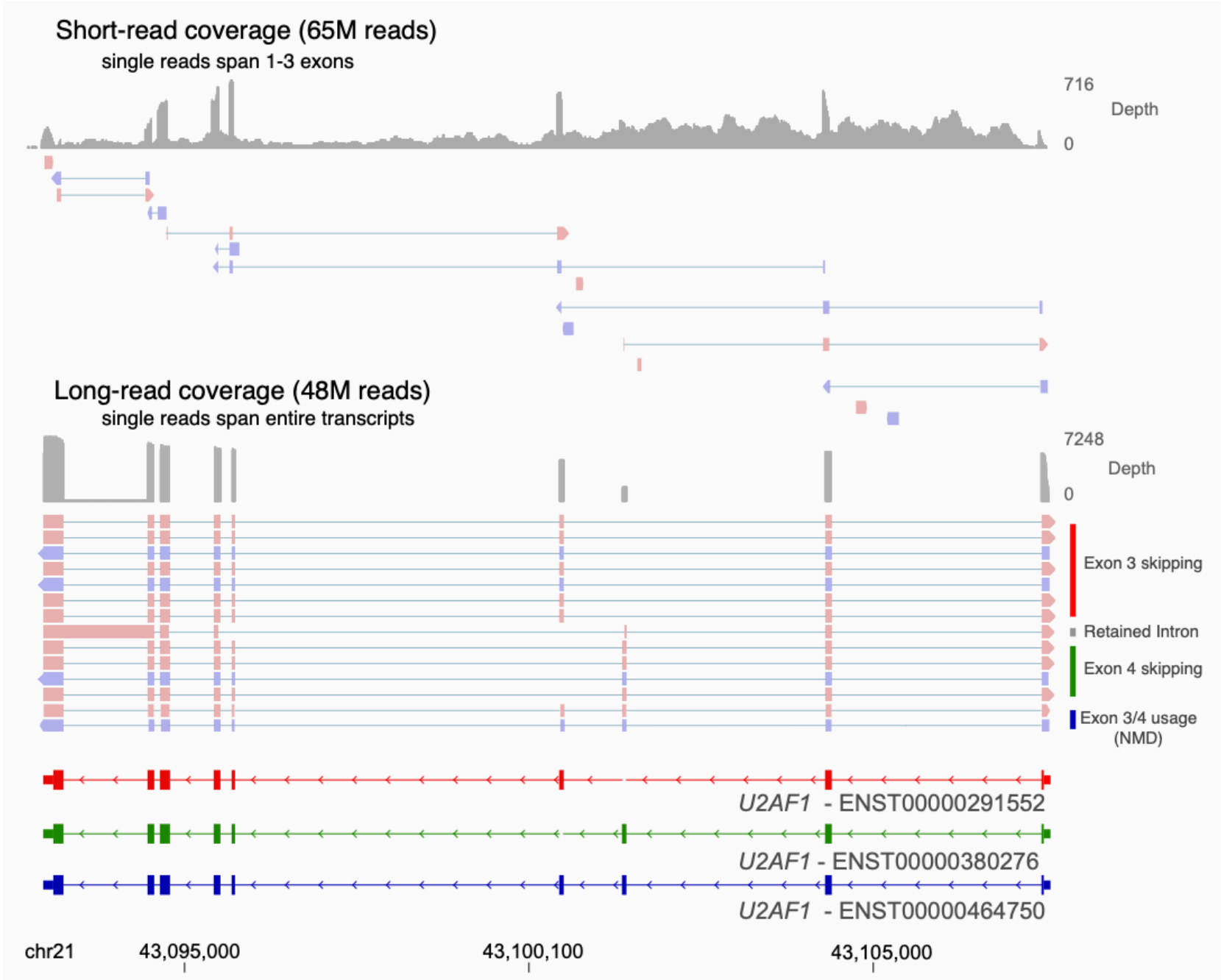
RNA/cDNA – standard prep

What does the data look like?



cDNA – standard prep


[0 - 1138]
U2af1

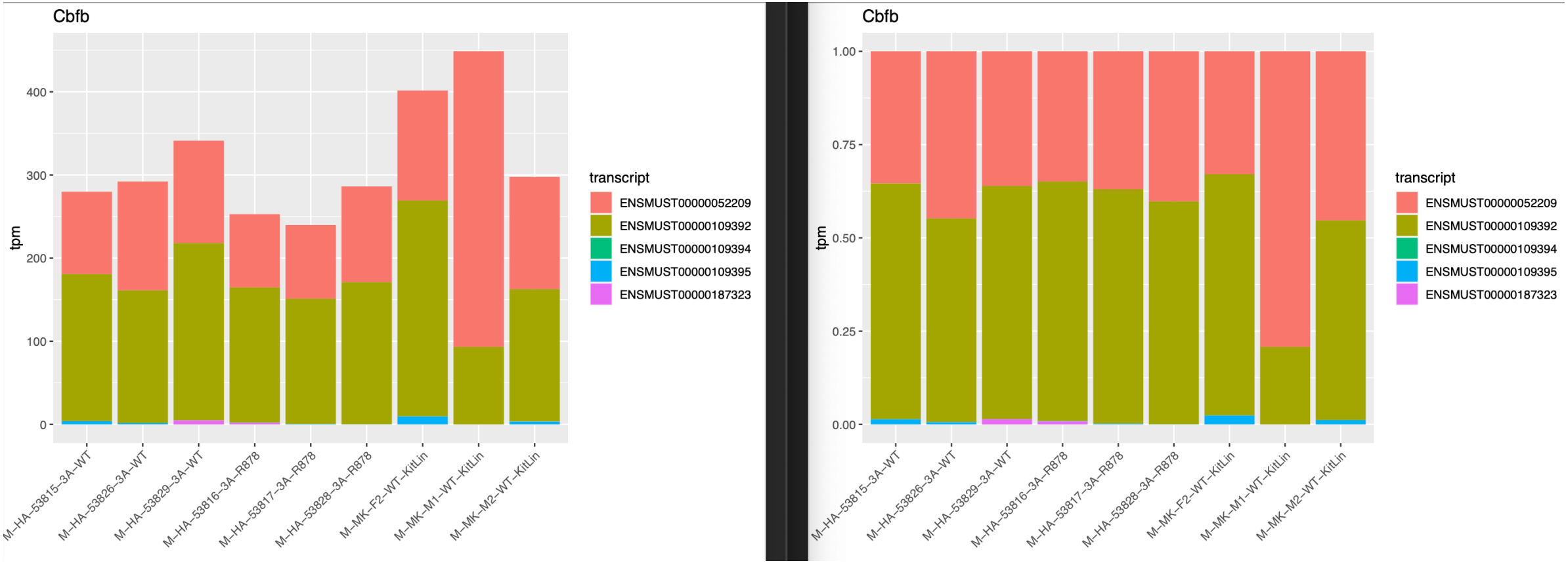

Short-read coverage (65M reads)
single reads span 1-3 exons
716
Depth
0
Long-read coverage (48M reads)
single reads span entire transcripts
7248
Depth
0
Exon 3 skipping
Retained Intron
Exon 4 skipping
Exon 3/4 usage (NMD)
U2AF1 - ENST00000291552
U2AF1 - ENST00000380276
U2AF1 - ENST00000464750
chr21
43,095,000
43,100,100
43,105,000

Estimating transcript abundance – long read

- We are producing much more data than most groups have to date
- FLAIR, Liqa, IsoQuant all either crash or run forever
- TALON, Bambu, and Stringtie all run in a reasonable time frame

50M long reads ~= 380M short reads

Transcript usage of mouse Cbfb



Takeaways

- ONT data looks really promising for transcript abundance estimation, alt-splicing, etc

- Still a fair number of technical hurdles to overcome
  - validating toolchains and workflows
  - batch effects, kits that are being continuously improved
  - thinking at a transcript level