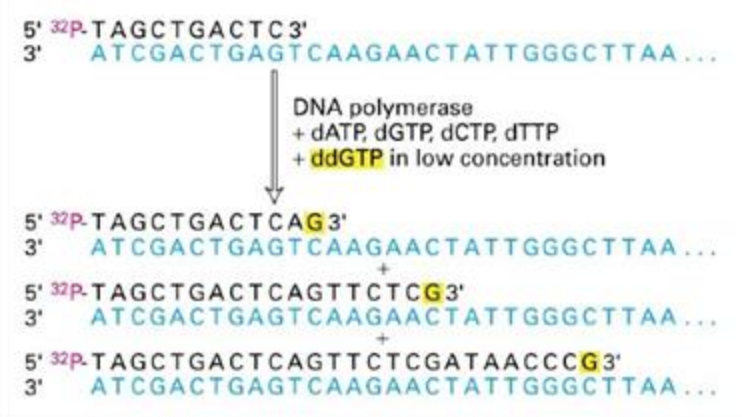
Long Read Sequencing

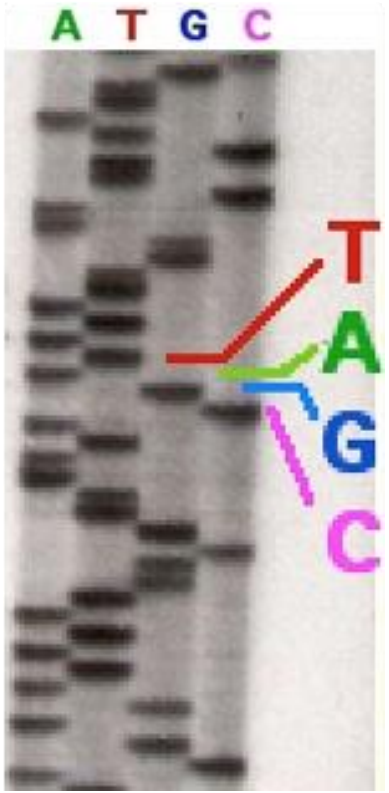
Chris Miller, Ph.D.
Washington University in St Louis

How to sequence a human genome: Sanger method

Key points:

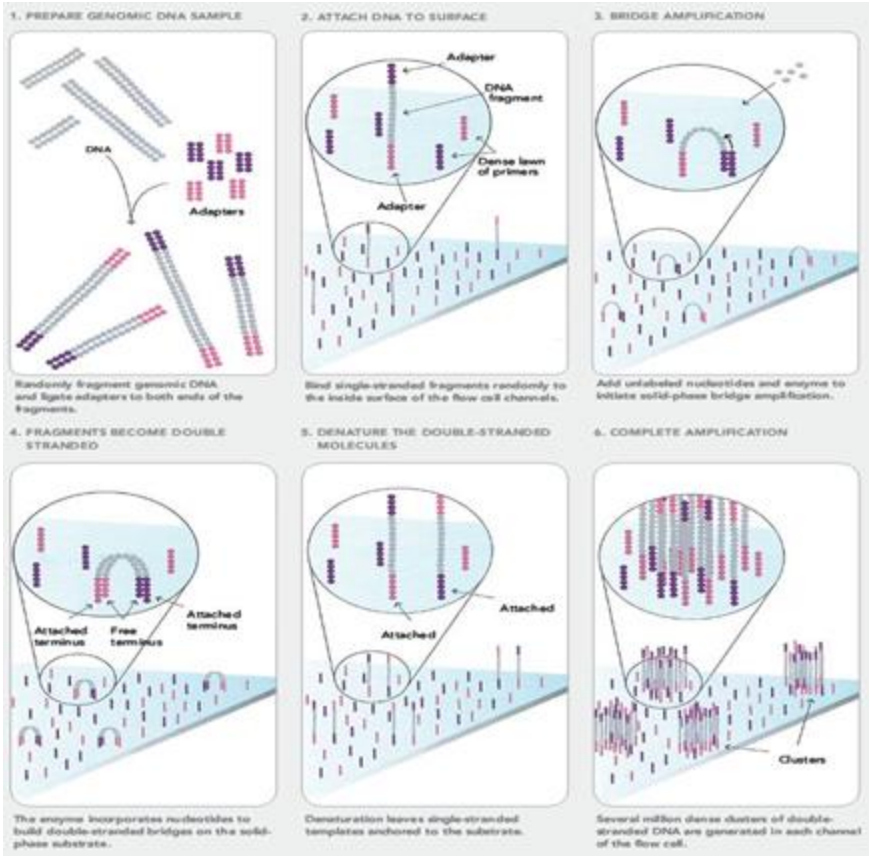
1) sequencing by synthesis (not degradation)
2) primers hybridize to DNA
3) polymerase + dNTPS + ddNTP terminators at low concentration
4) 1 lane per base, visually interpret ladder





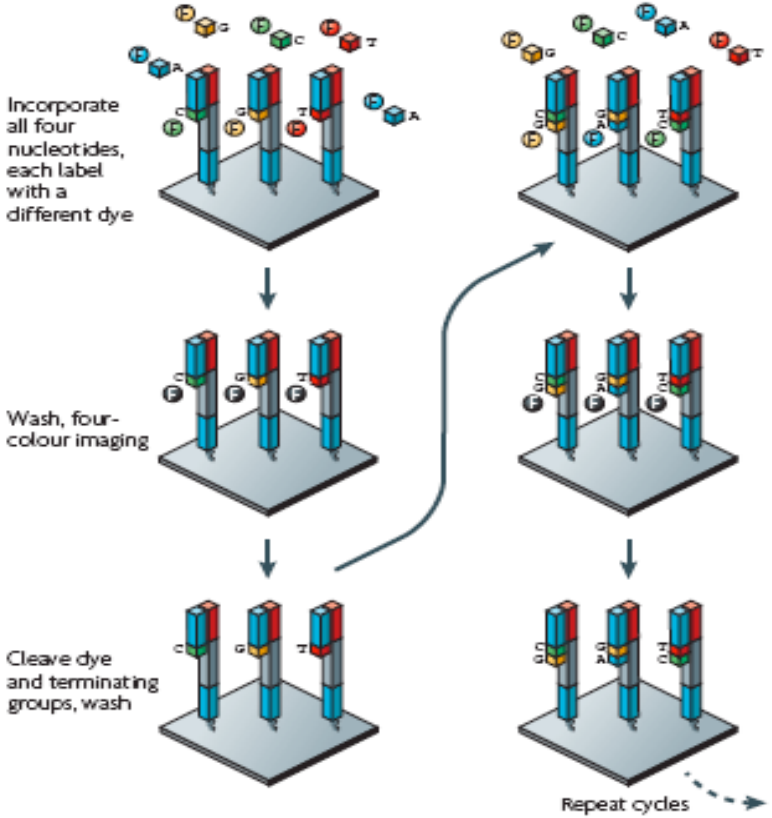
Solexa (Illumina) sequencing (2006)

- **PCR amplify sample (opt.)**
- **Immobilize and amplify single molecules on a solid surface**
- **<u>Reversible</u> terminator sequencing with 4 color dye-labelled nucleotides**

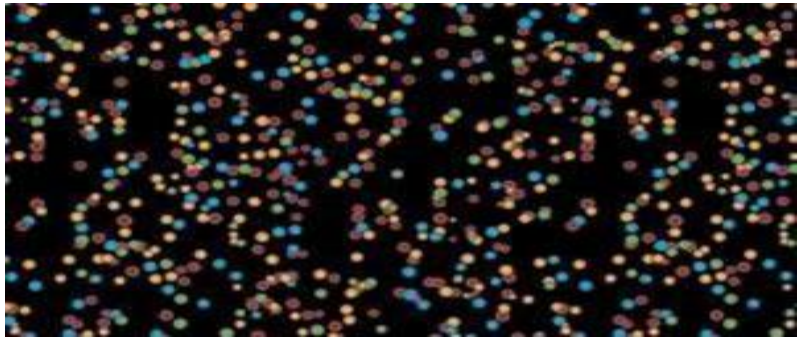


Slide from Ira Hall

Illumina sequencing (2005)



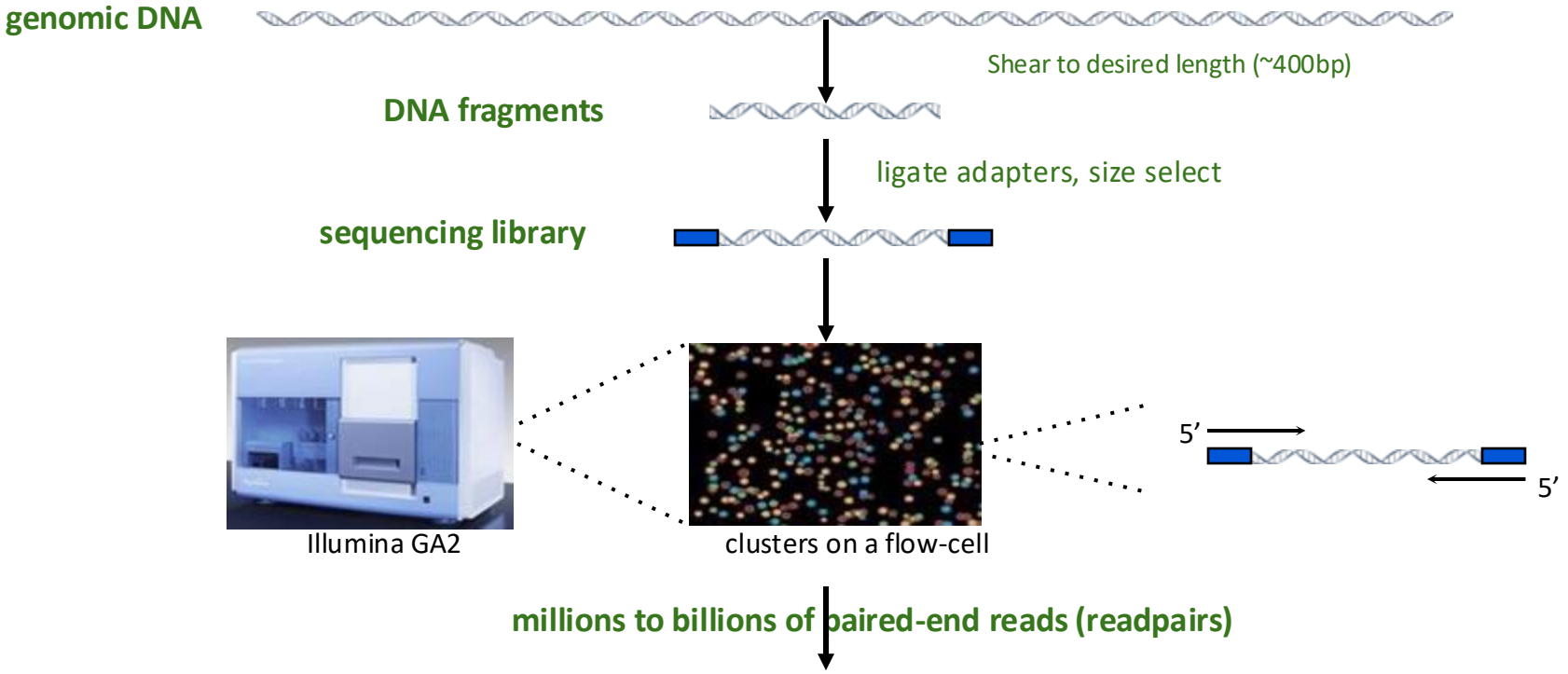
4 different images merged



6 cycles w/ base-calling

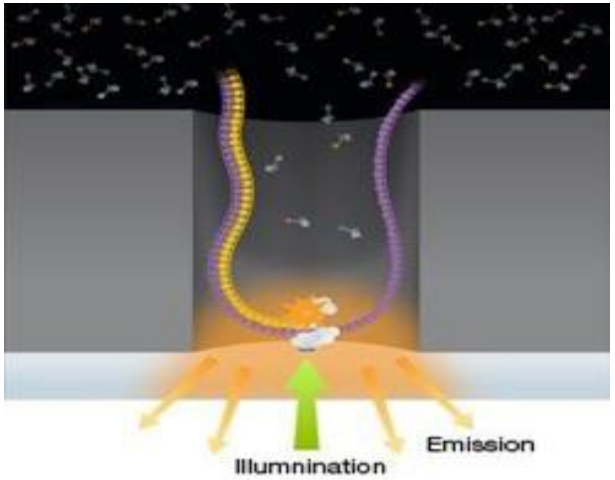
Slide from Ira Hall

Paired-end sequencing:
A molecular hack to sequence longer fragments



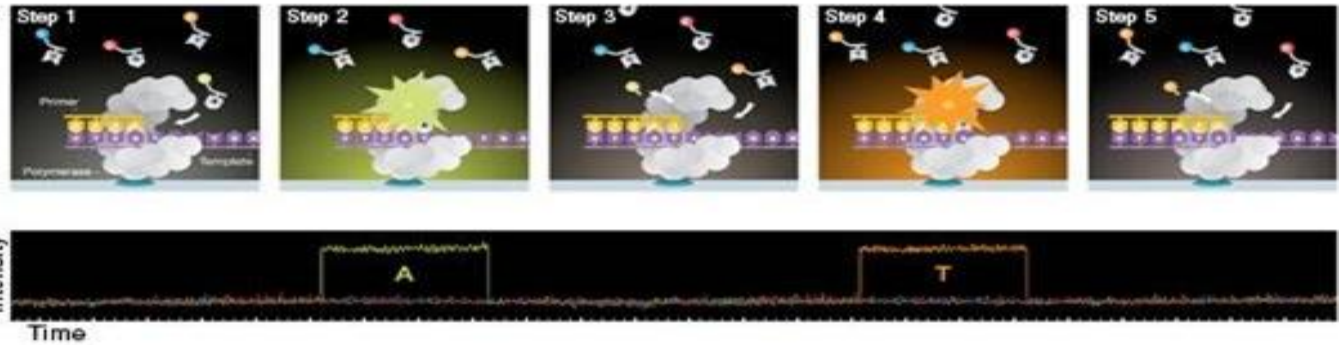
Slide from Ira Hall

Pacific Biosciences



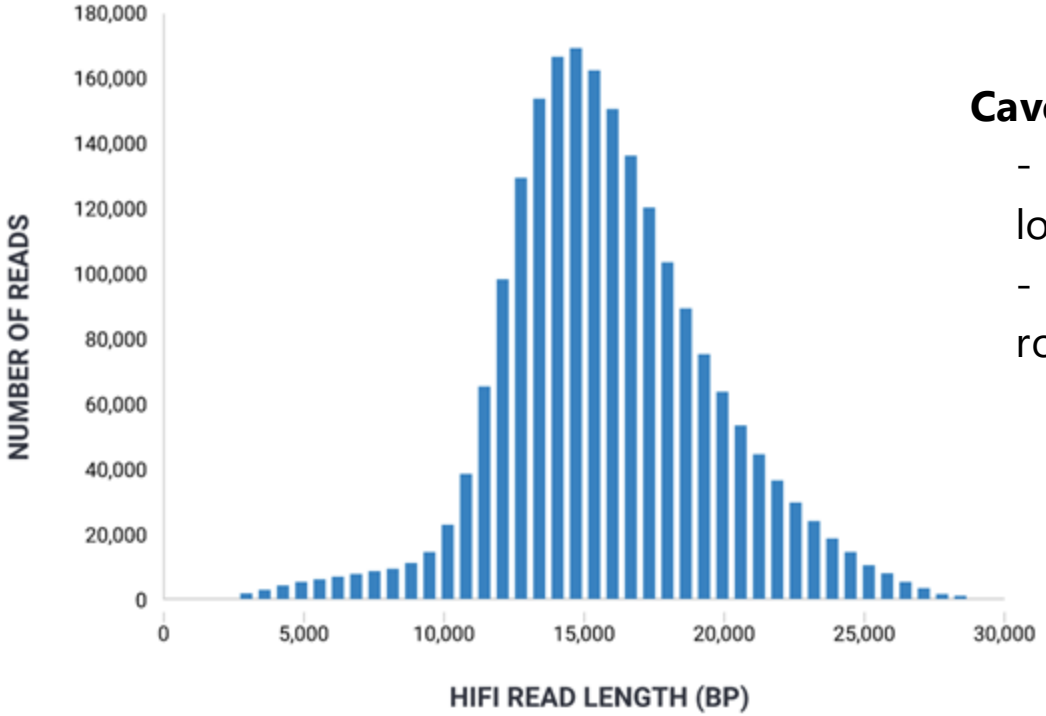
Key Points:

- 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



Slide from Ira Hall

Pacific Biosciences: long reads. Great for genome assembly

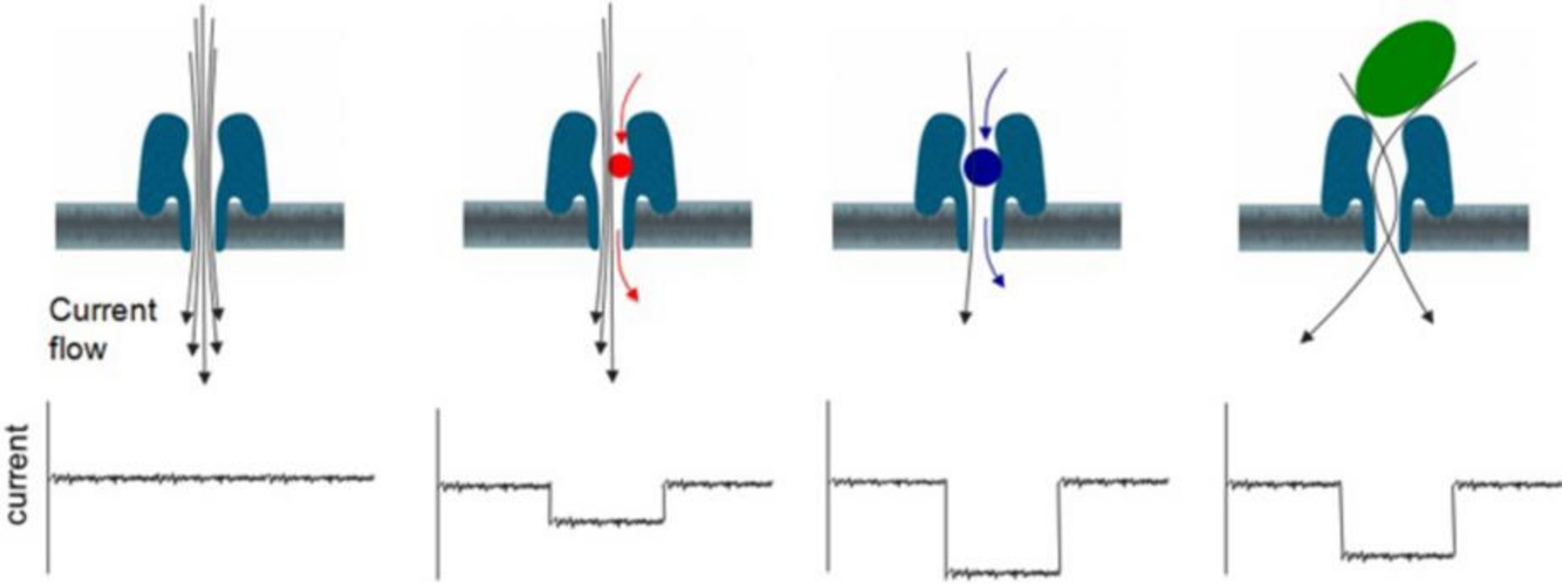


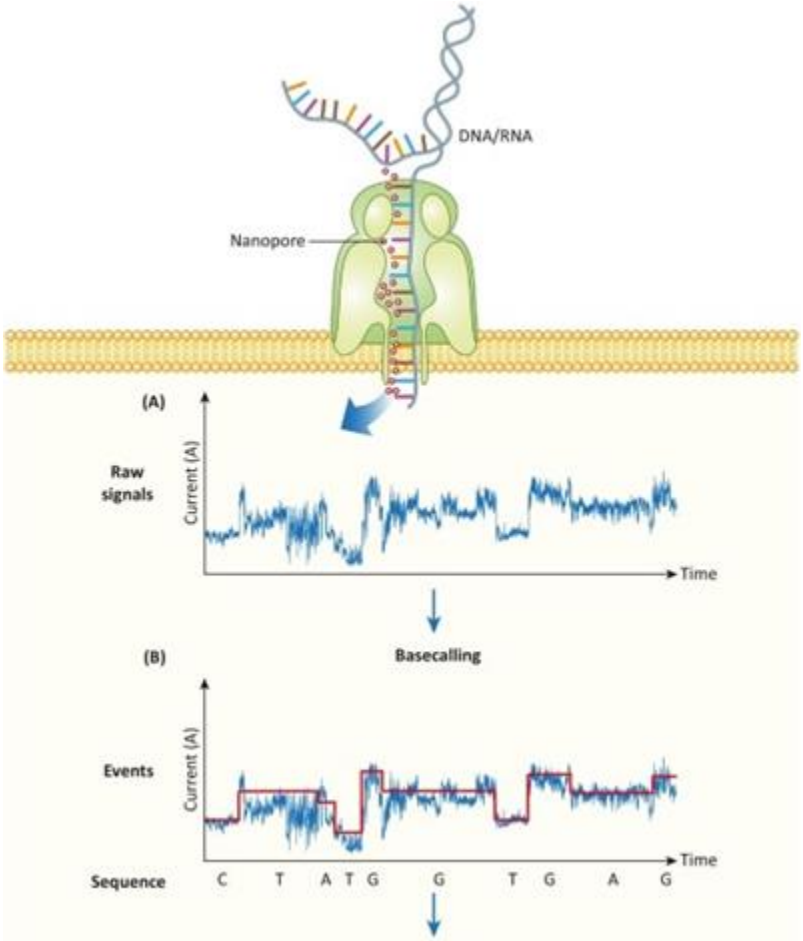
Caveats:

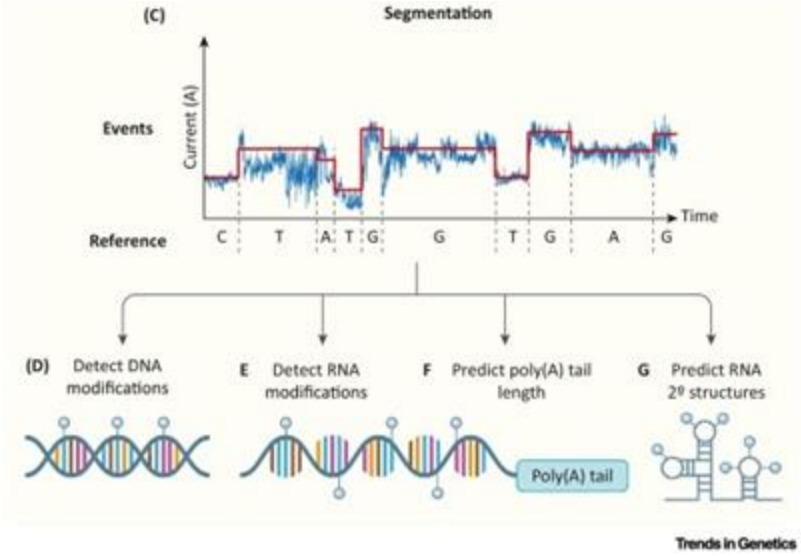
- higher error rate (1-2%), lower with Duplex runs
- lower throughput : roughly 90 gigabases per run

About $4,000 for a 30x human genome on the RevIO machine

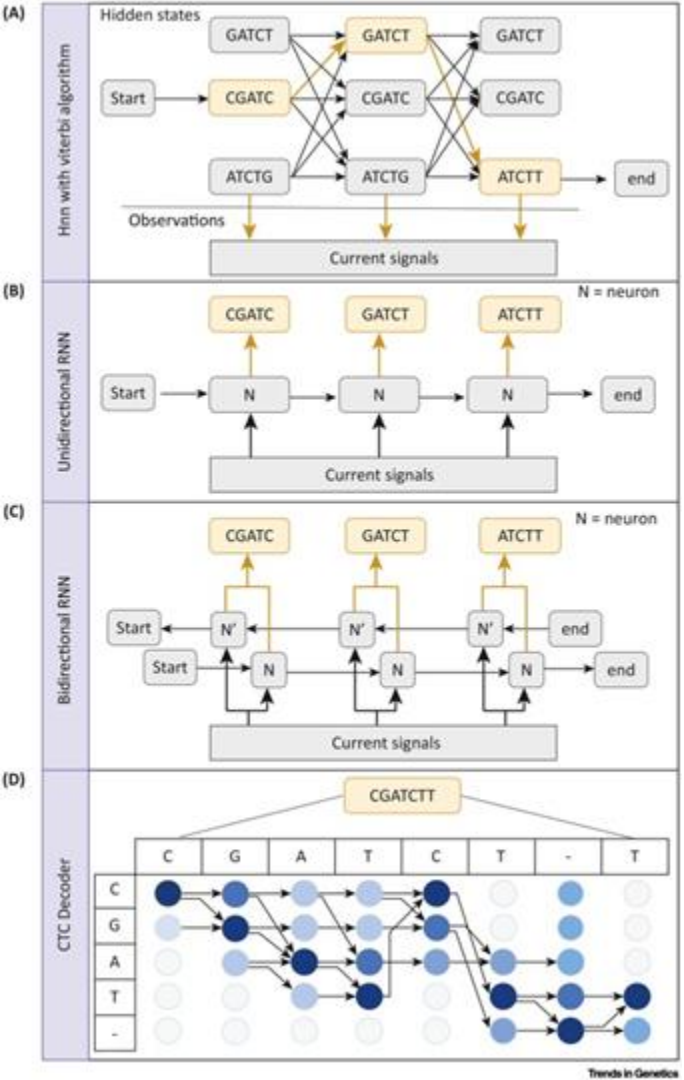
Oxford Nanopore Technologies







doi.org/10.1016/j.tig.2021.09.001



Neural networks to translate signal into base calls

- Guppy (many versions)

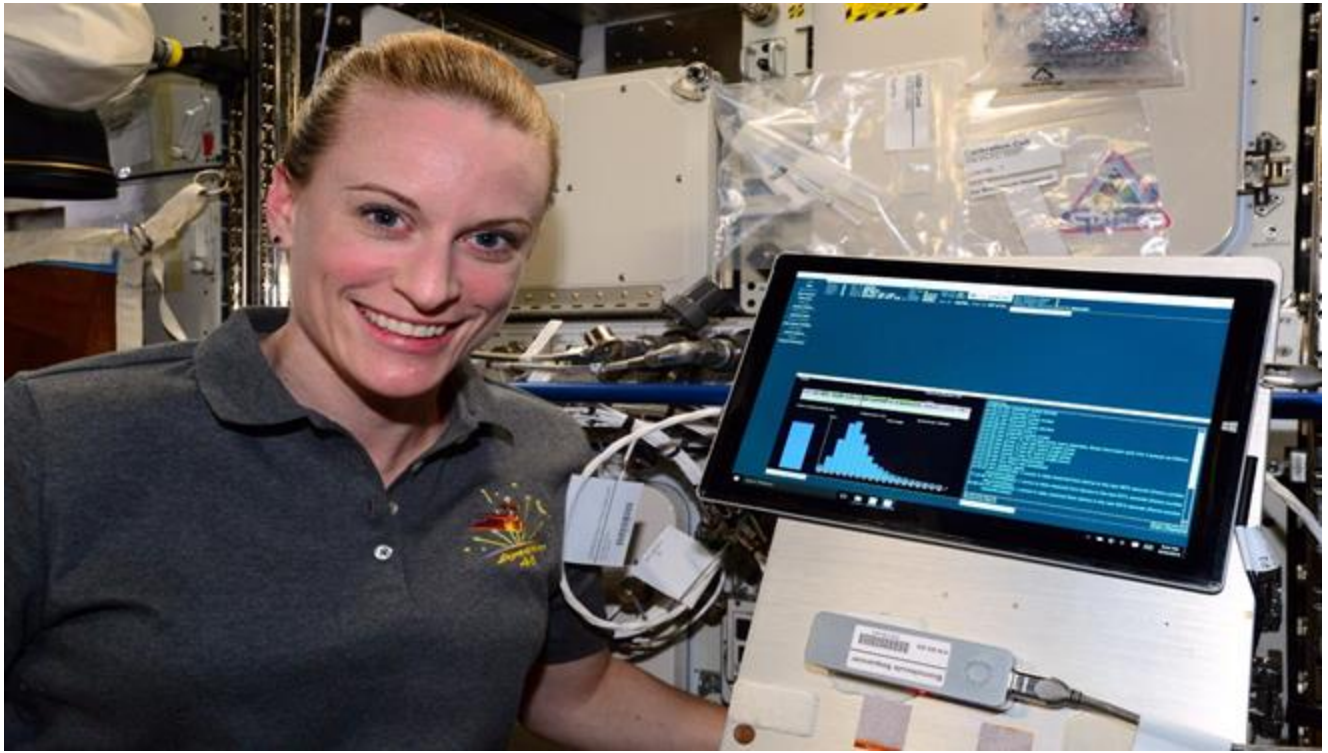
- Dorado (v0.4, eventual guppy replacement)

- many others

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

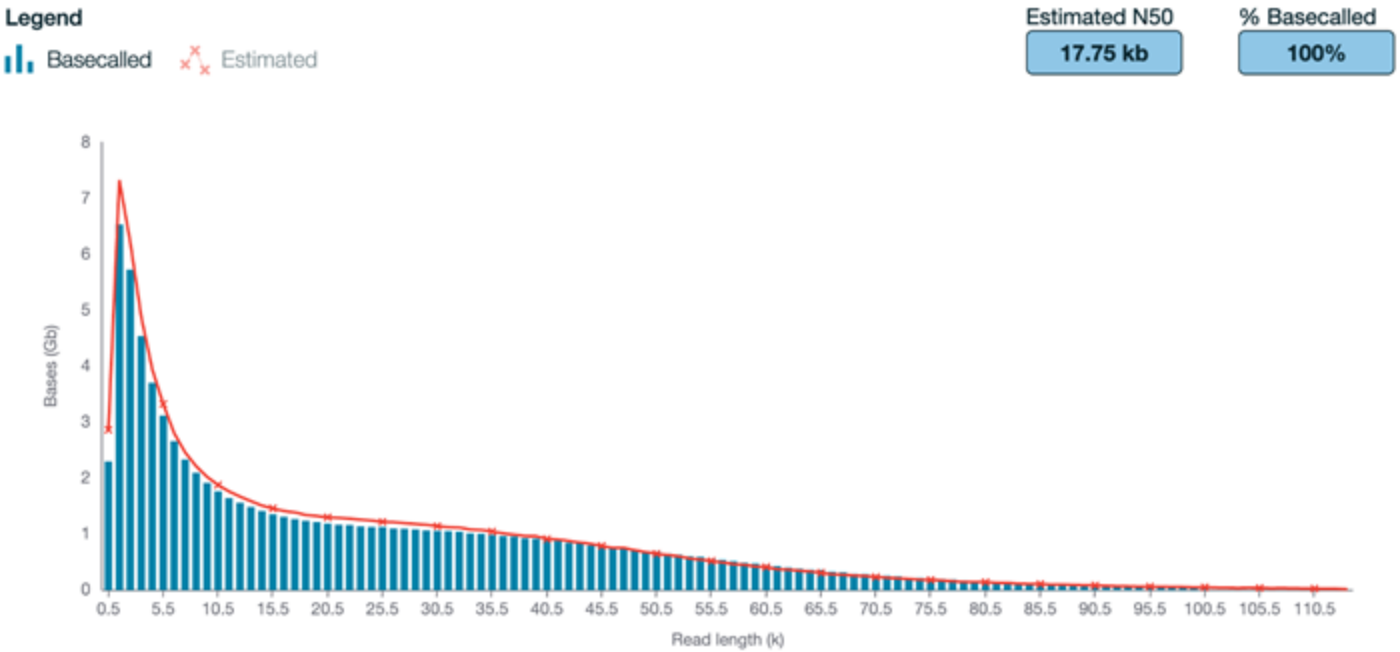
doi.org/10.1016/j.tig.2021.09.001

Nanopore sequencing is *extremely* portable



Kate Rubins sequencing DNA on the ISS

ONT sequence length distribution



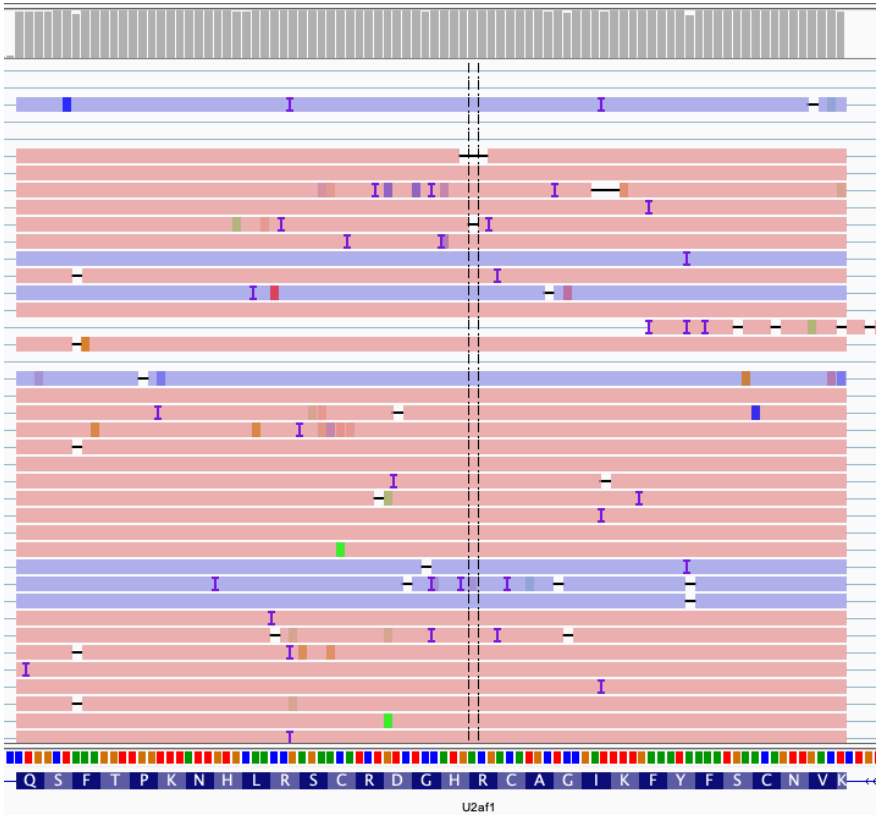
their relative amounts.

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

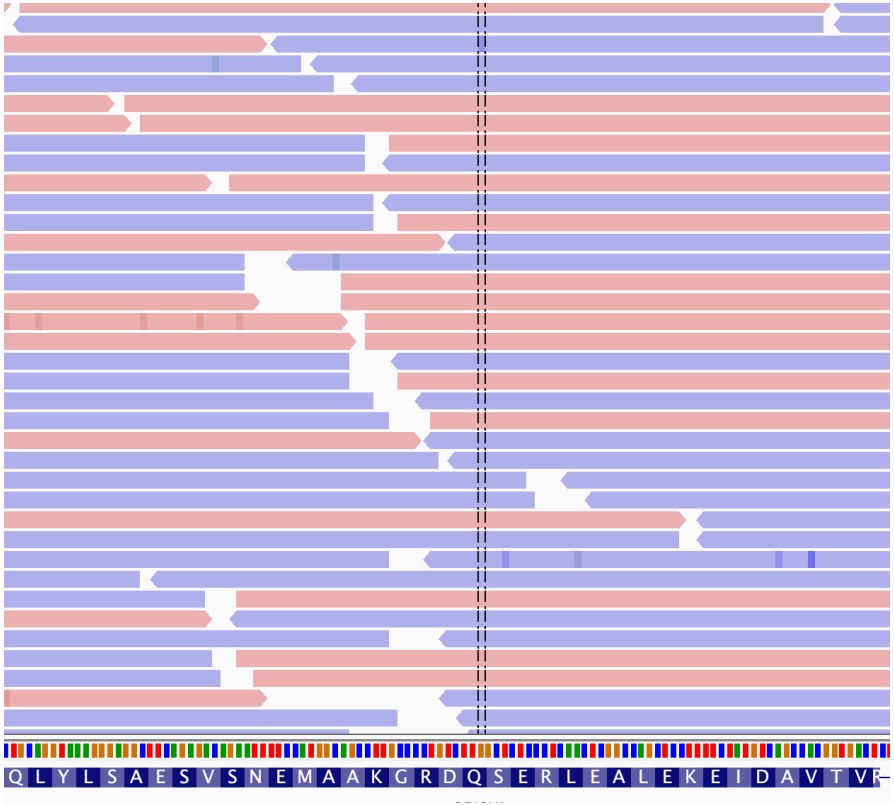
Recent run of a tumor sample

About $3,500 for a 30x human genome on a PromethION

What does the data look like?



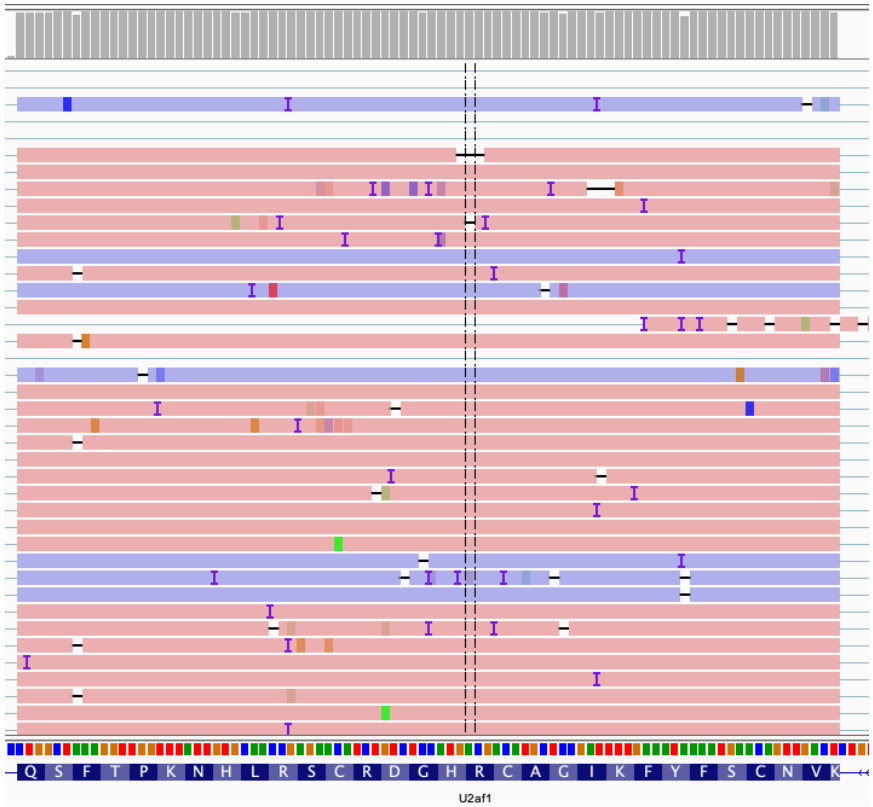
Long-read ONT ~5% base error rate



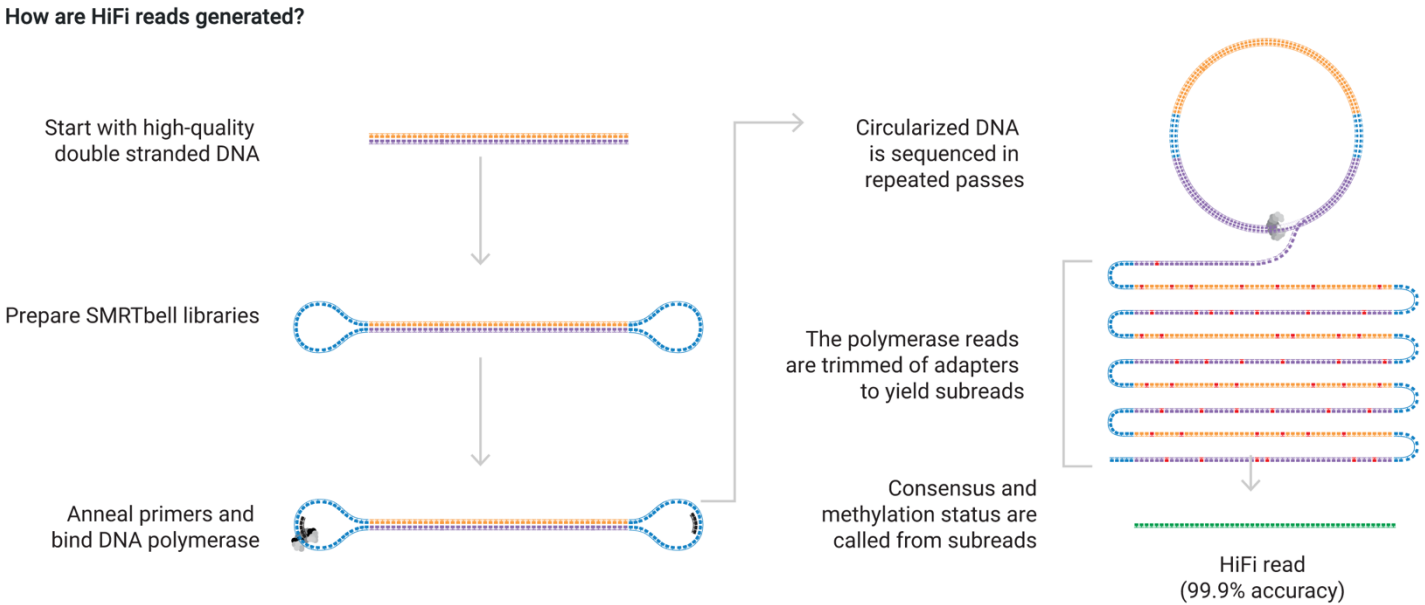
Short-read Illumina ~0.3% base error rate

Error rates are contentious and confusing

- How do you calculate error?
  Per base?
  Per read?
  Per variant call?
  (after collapsing all of the data?)



PacBio HiFi Sequencing

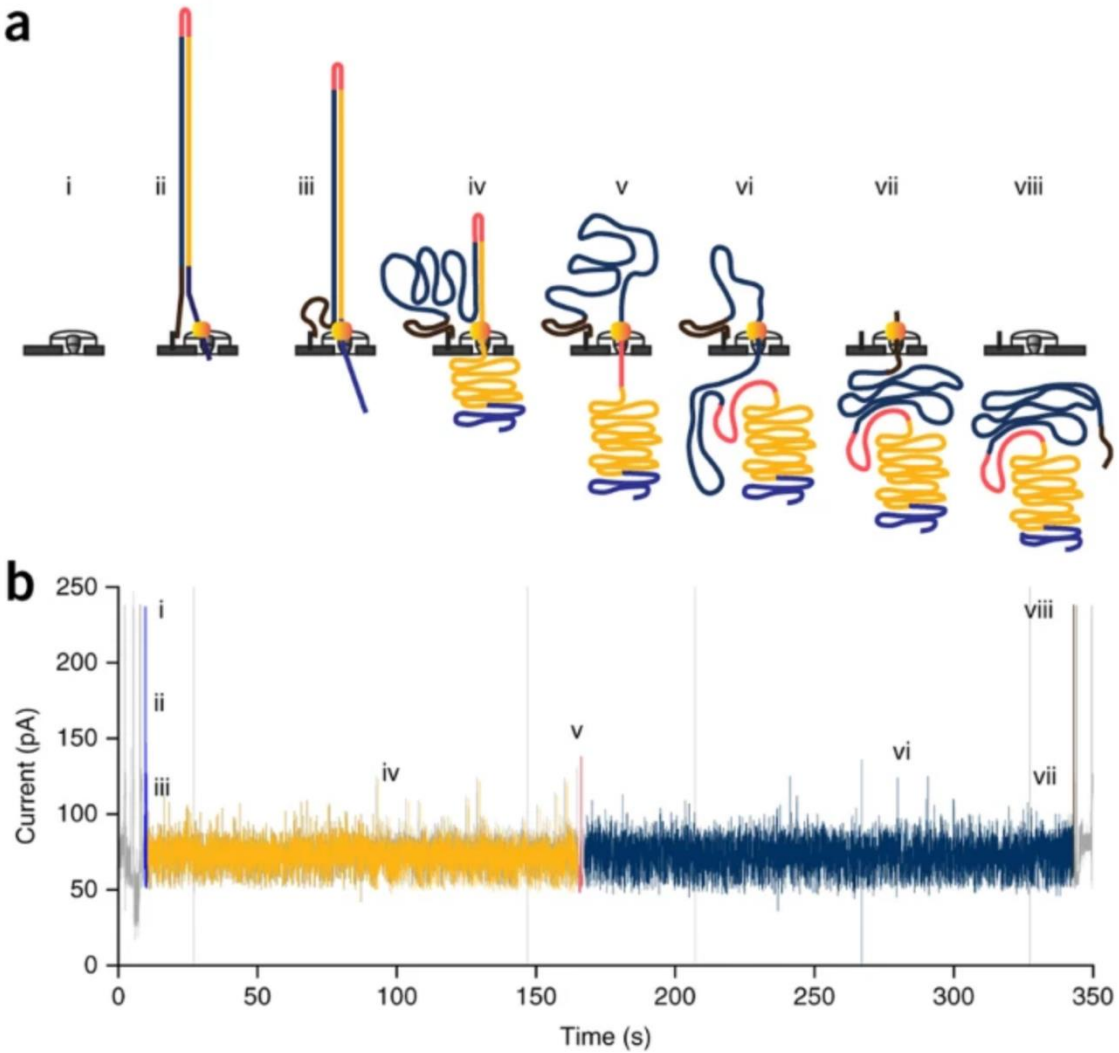


Improved error rates

higher cost/lower throughput

https://www.pacb.com/technology/hifi-sequencing/

ONT Duplex sequencing



Improved error rates

higher cost/lower throughput

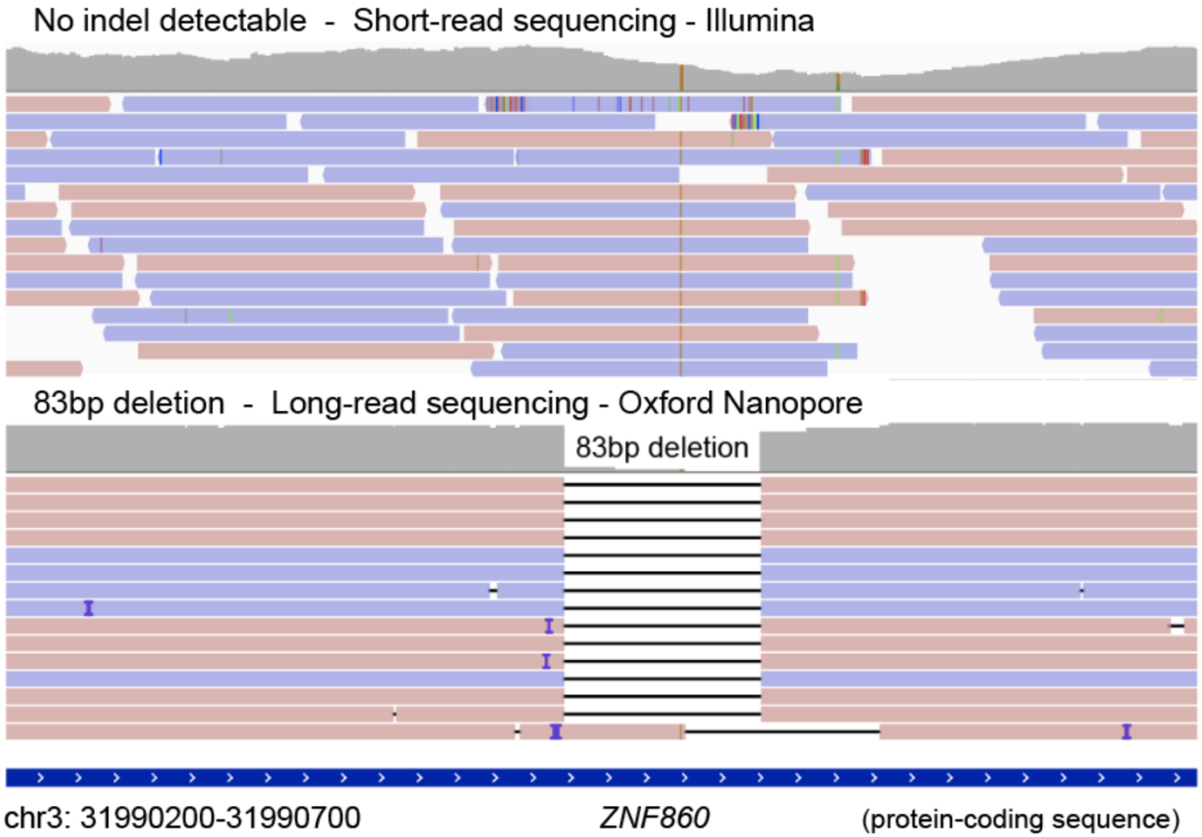
https://www.nature.com/articles/nmeth.3290

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)

David Spencer

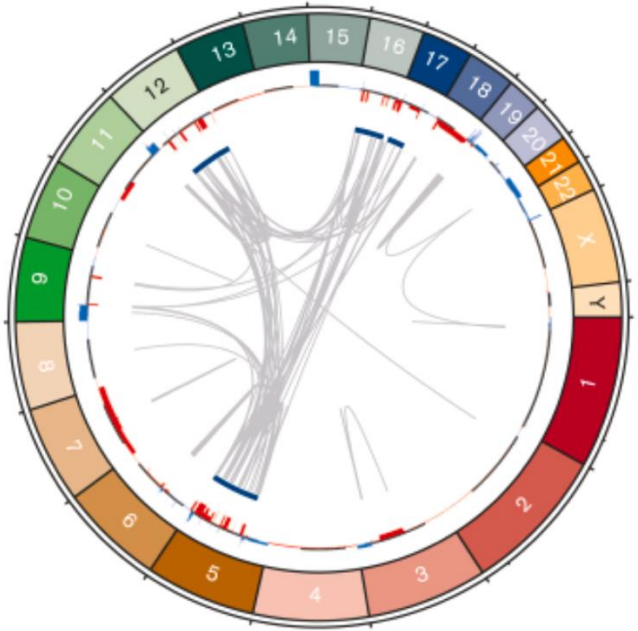
Large Indel detection



Haley Abel

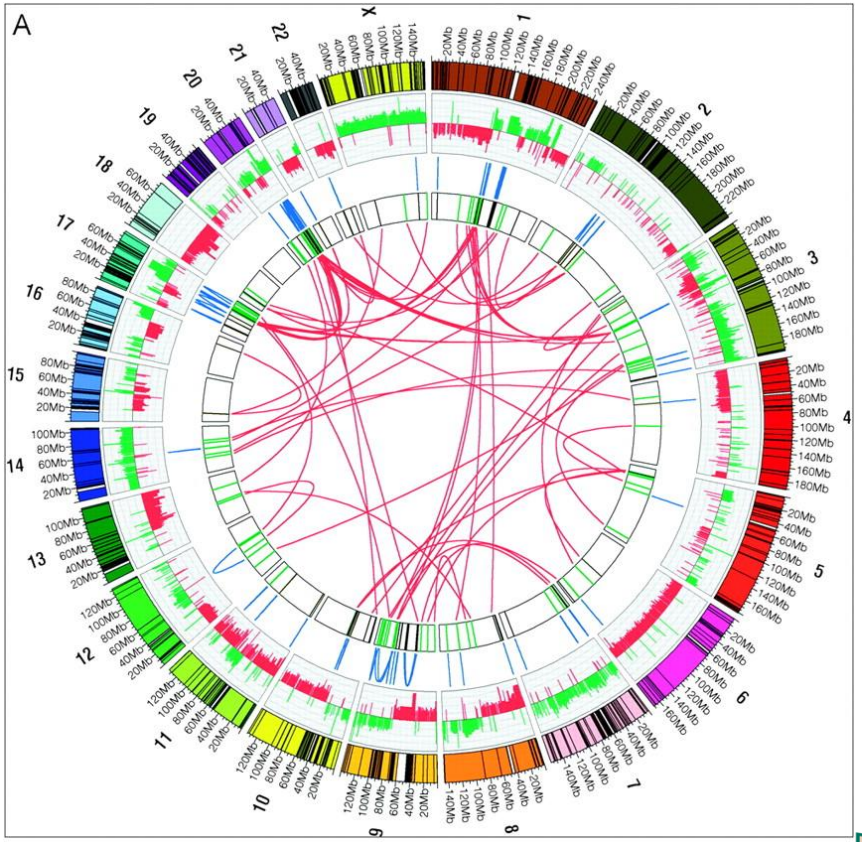
Structural variant resolution

TP53-mutated AML



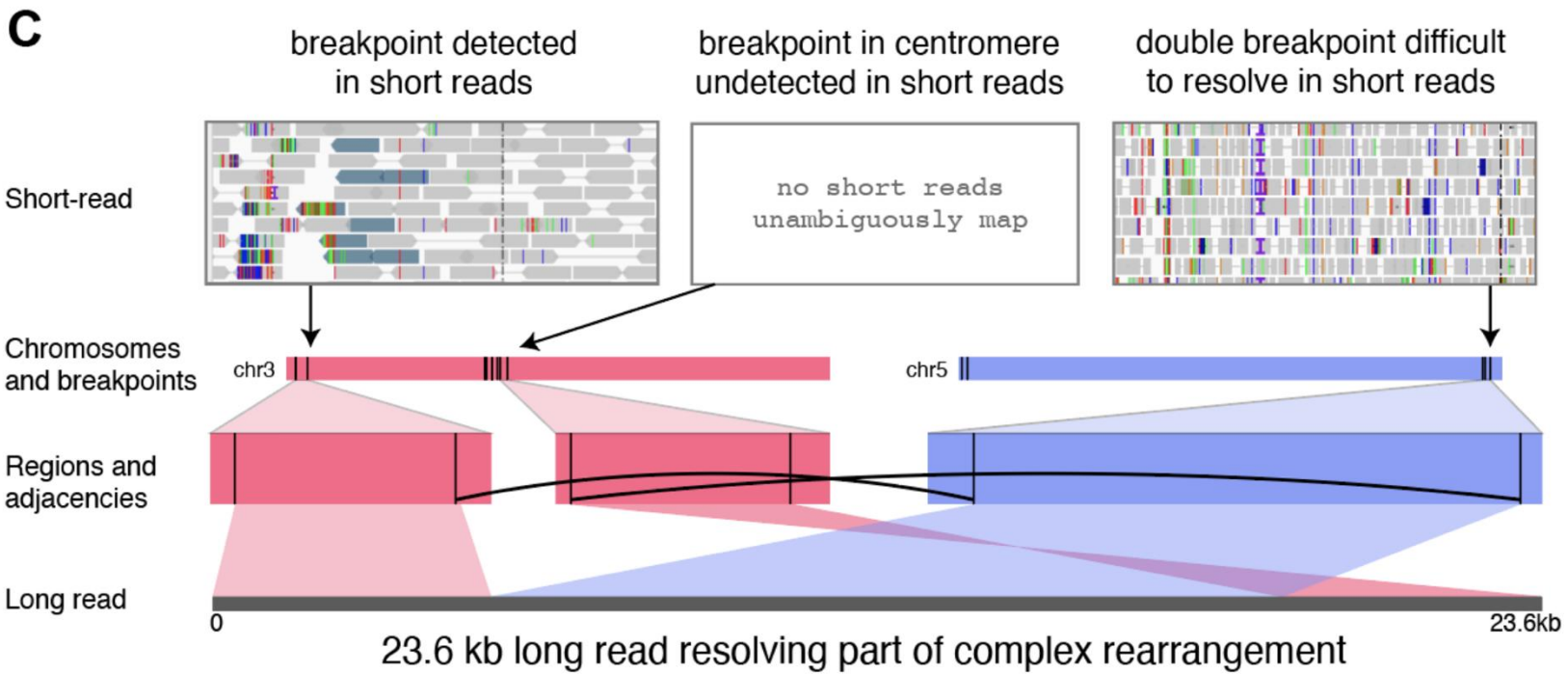
doi: 10.1182/bloodadvances.2023010156

MCF7 Breast Cancer Cell line



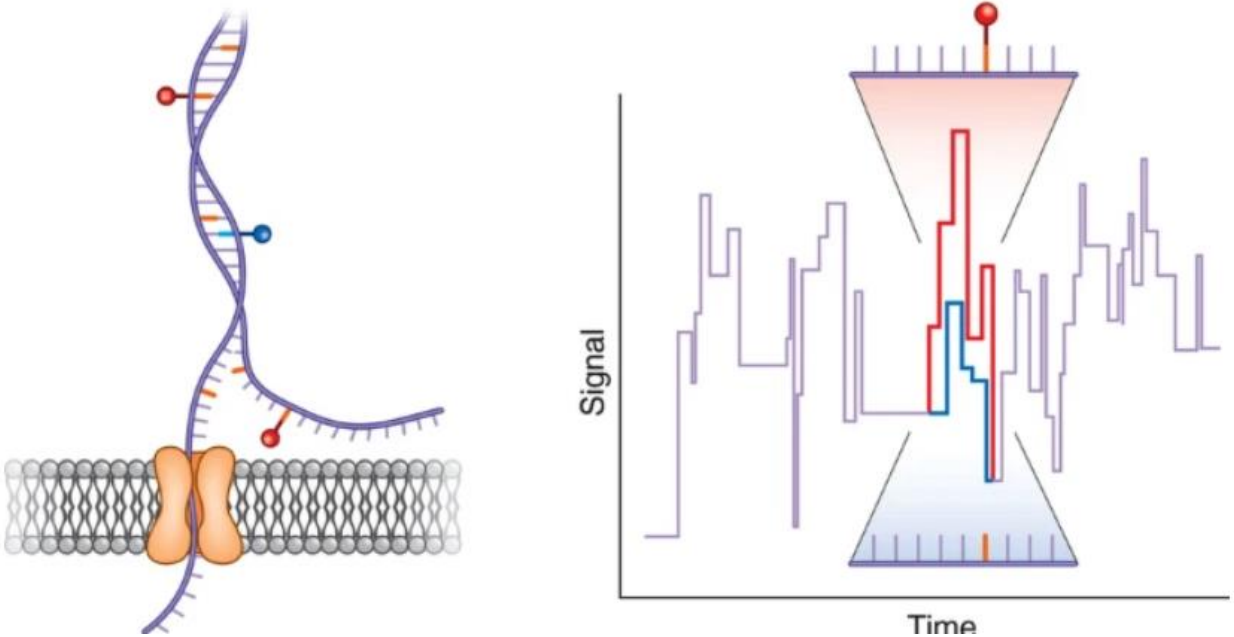
Hampton, et al. doi: 10.1101/gr.080259.108

Structural variant resolution



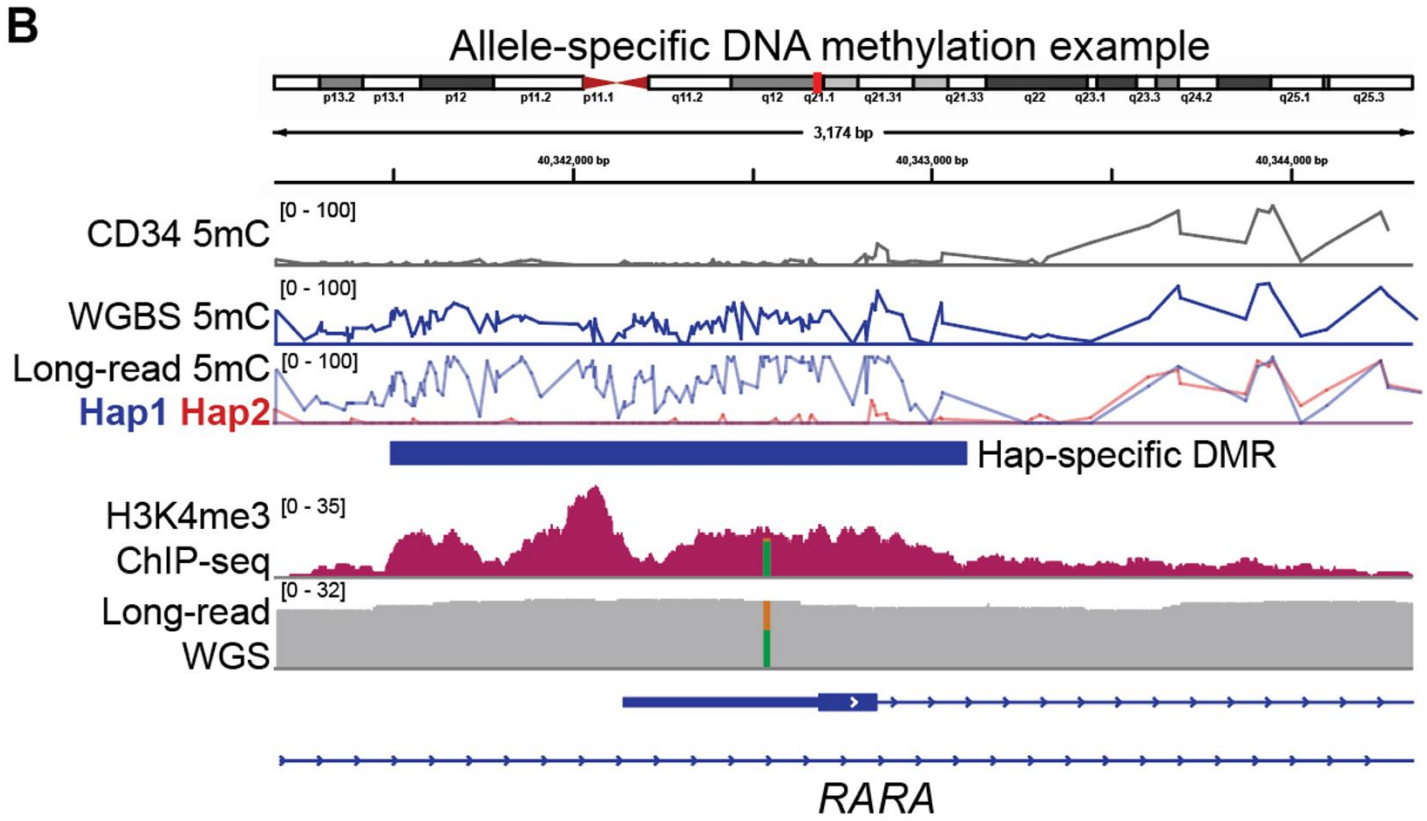
Haley Abel

Base modification detection



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

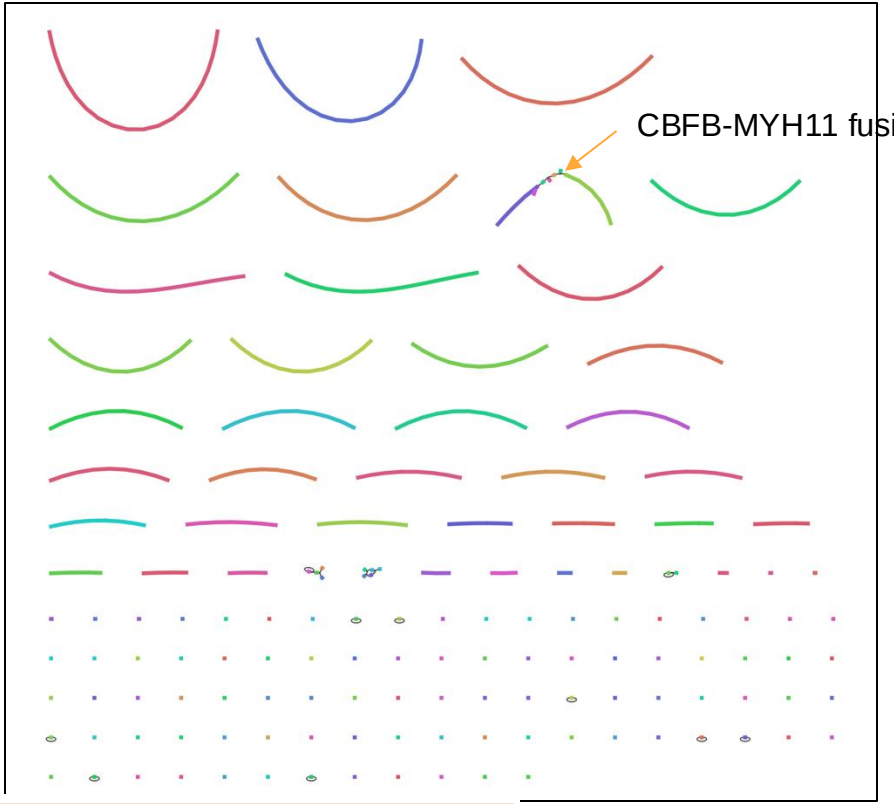
Phasing of reads/modifications



Dave Spencer

Genome assembly

- Assembly of personal genomes

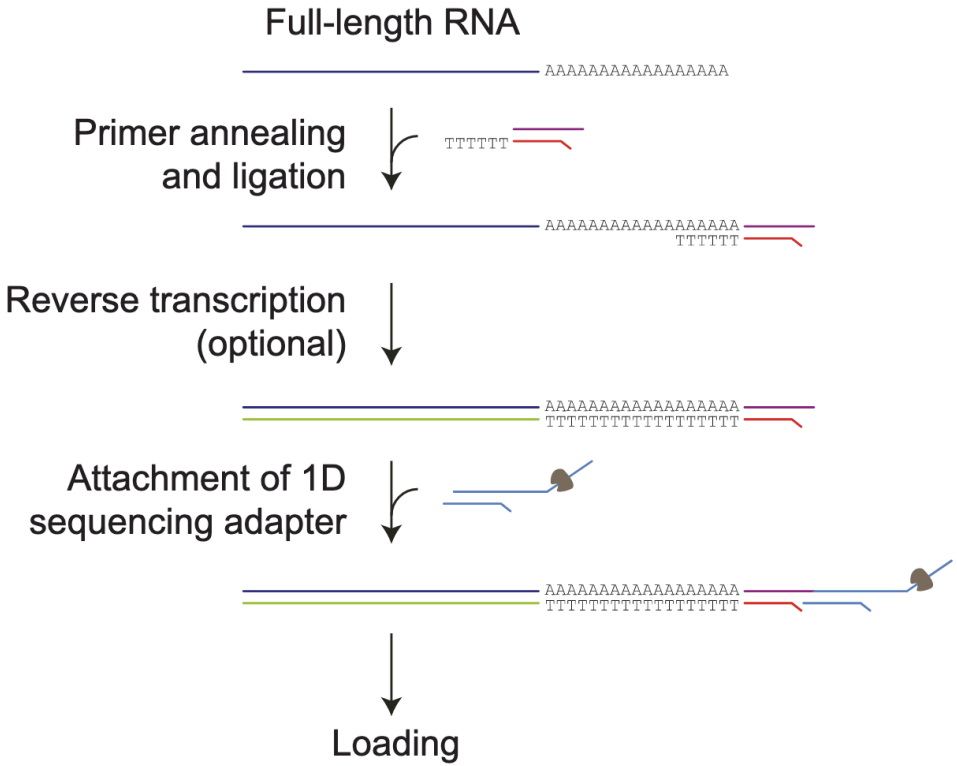


sample	hap	metric	value
Presentation	hap1	Number_of_contigs	402
Presentation	hap1	N50	88.3 Mbp

Dave Spencer, Haley Abel

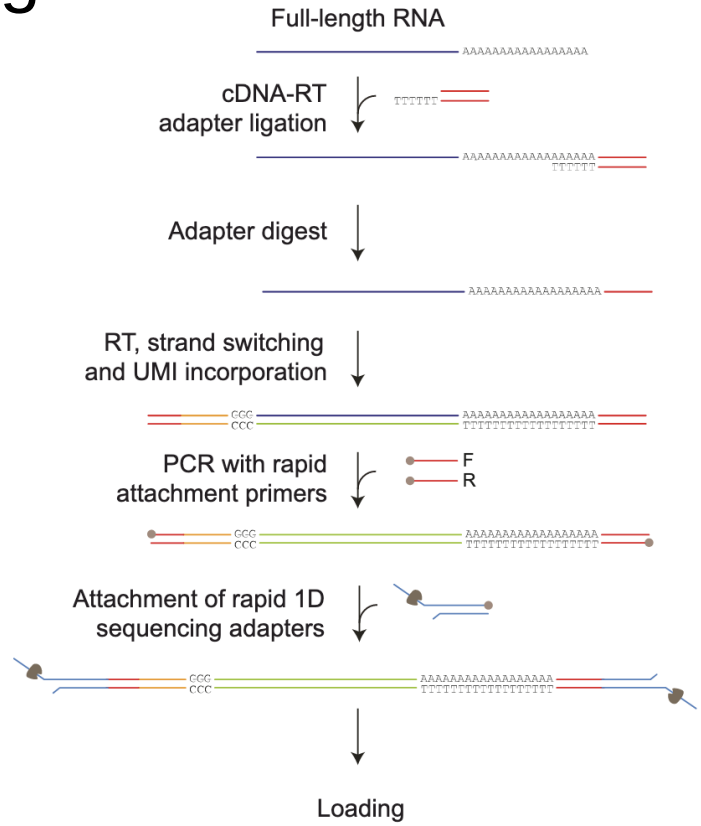
Long-read RNA sequencing

- Direct RNA
- No amplification, less bias
- Preserves base modifications (m6a, etc)



Long-read RNA sequencing

- cDNA sequencing
- much higher yields



Pychopper

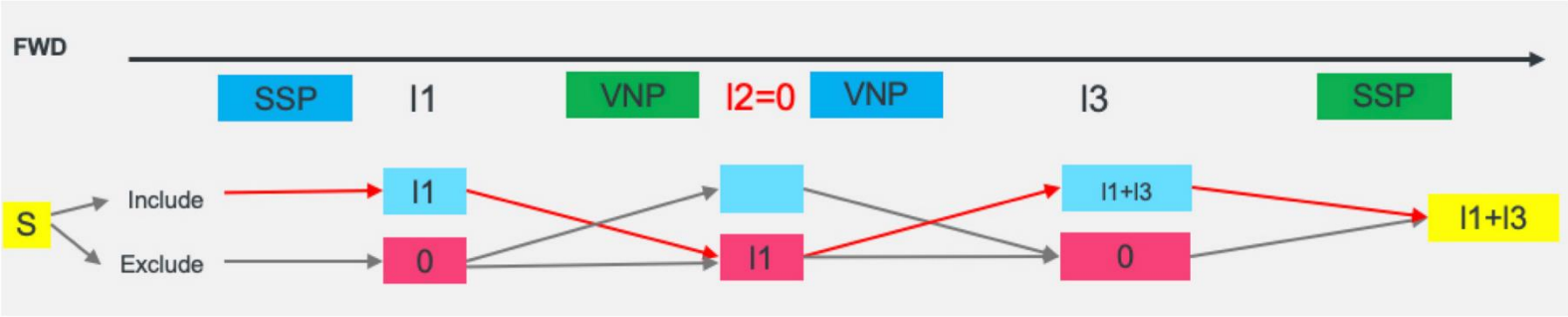
Primer Sequence Primer

Pychopper

Primer Sequence Primer

Pychopper

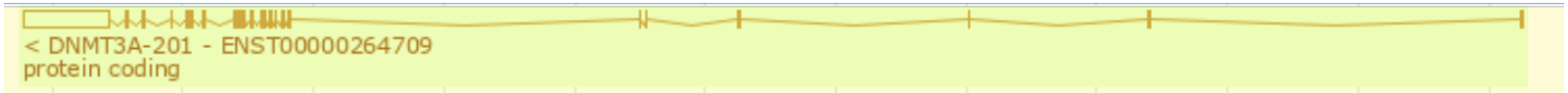
Primer Sequence Primer



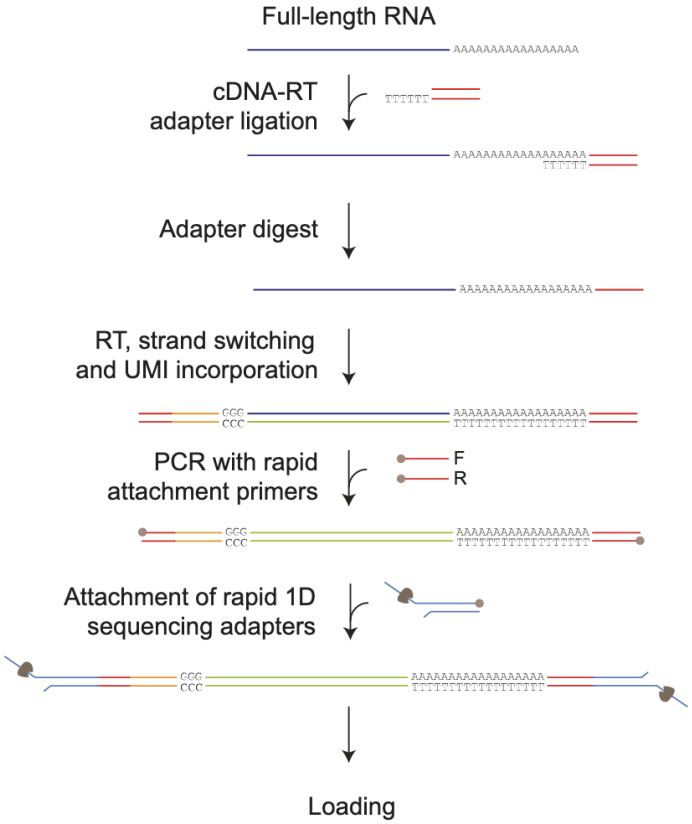
How to estimate duplication rates

- In short read data, reads at the same position are assumed to be duplicates

- How do we know if we're saturating our libraries?



UMIs


Full-length RNA
AAAAAAAAAAAAAAAAA
cDNA-RT
adapter ligation
TTTTTT
AAAAAAAAAAAAAAAAAA
TTTTTT
Adapter digest
AAAAAAAAAAAAAAAAAA
RT, strand switching
and UMI incorporation
GGG
CCC
AAAAAAAAAAAAAAAAAA
TTTTTTTTTTTTTTTTTT
PCR with rapid
attachment primers
F
R
GGG
CCC
AAAAAAAAAAAAAAAAAA
TTTTTTTTTTTTTTTTTT
Attachment of rapid 1D
sequencing adapters
GGG
CCC
AAAAAAAAAAAAAAAAAA
TTTTTTTTTTTTTTTTTT
Loading

UMIs

- UMI at the 3' end of the read

UMIs

- UMI at the 3' end of the read

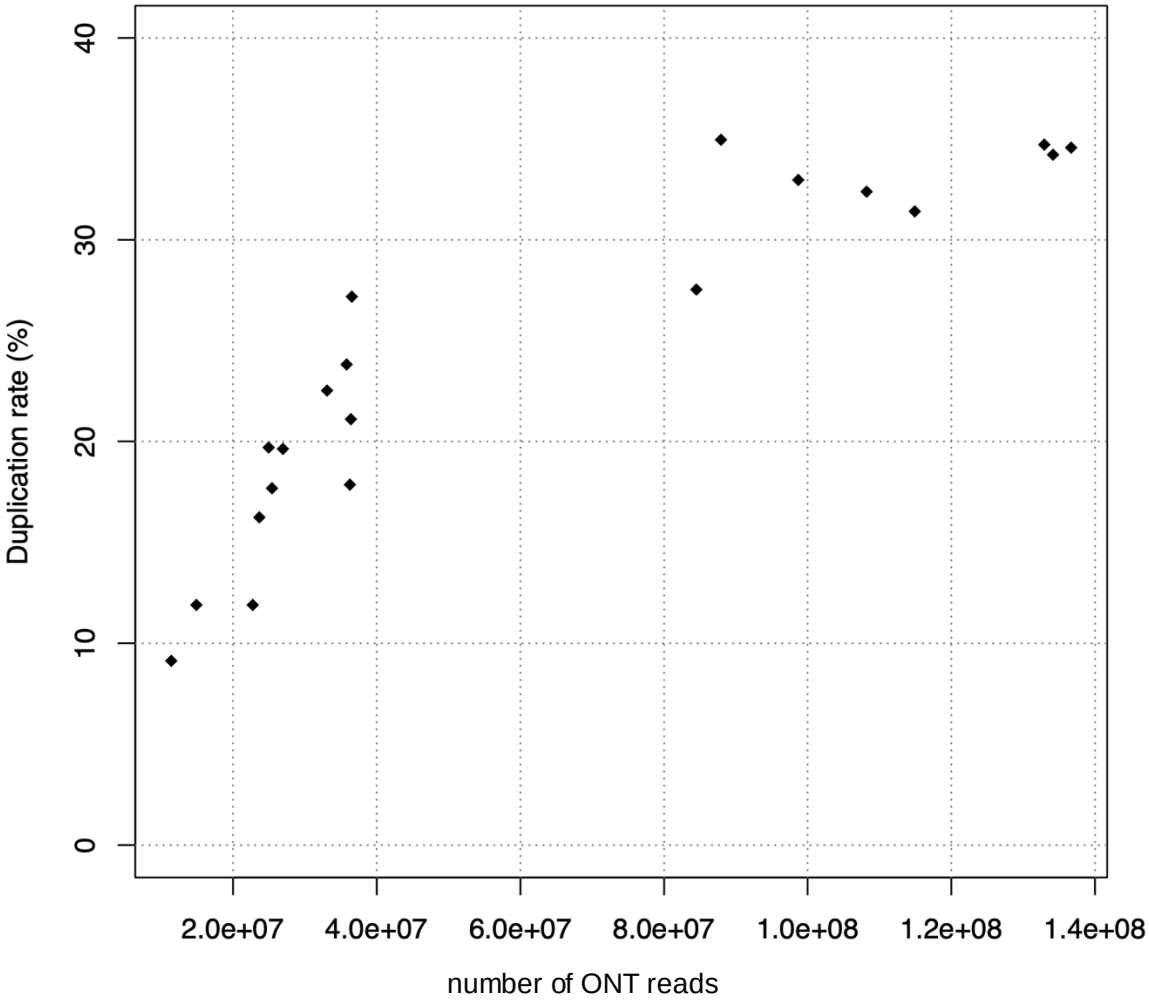
```
30 TTTCACCCTCCACTTCCCGTCTCAGAATT
29 TTTGAAACAGCTTCACCTTGAACTTT
29 TTTCCAATAAAAAAAATTACAATTT
29 TTTCAGCAAAATAAAATTCCGGTTT
27 TTTGGAGTTGGGGTTGCGCTTGGGGTTT
27 TTTGAGGTTGGAGTTGGGGTTGGCGTTT
24 TTTGGAGTTGGCGTTGCGGTTGGGGTTT
23 TTTGGGGTTGGAATTGGCGTTGGCATTT
23 TTTGGGATTAAGATTGGCATTGCGGTTT
23 TTTAGGGTTCGCGTTGGGGTTGCAGTTT
23 TTTAGGGTTAGCGTTGGAGTTGGGGTTT
22 TTTGGCGTTGGGGTTGGCGTTGGCGTTT
22 TTTGGCGTTGGAGTTCAGCTTACGGTTT
22 TTTGCGGTTGGAGTTGGGCTTGGCGTTT
22 TTTACACTTGTGCTCTCCTTAGCCTTT
21 TTTGGGGTTGGAGTTGGCGTTGGCATTT
21 TTTGGCGTTGGCATTGGCGTTGGGGTTT
21 TTTGGCGTTCGGGTTGGAATTCGCGTTT
```

UMIs

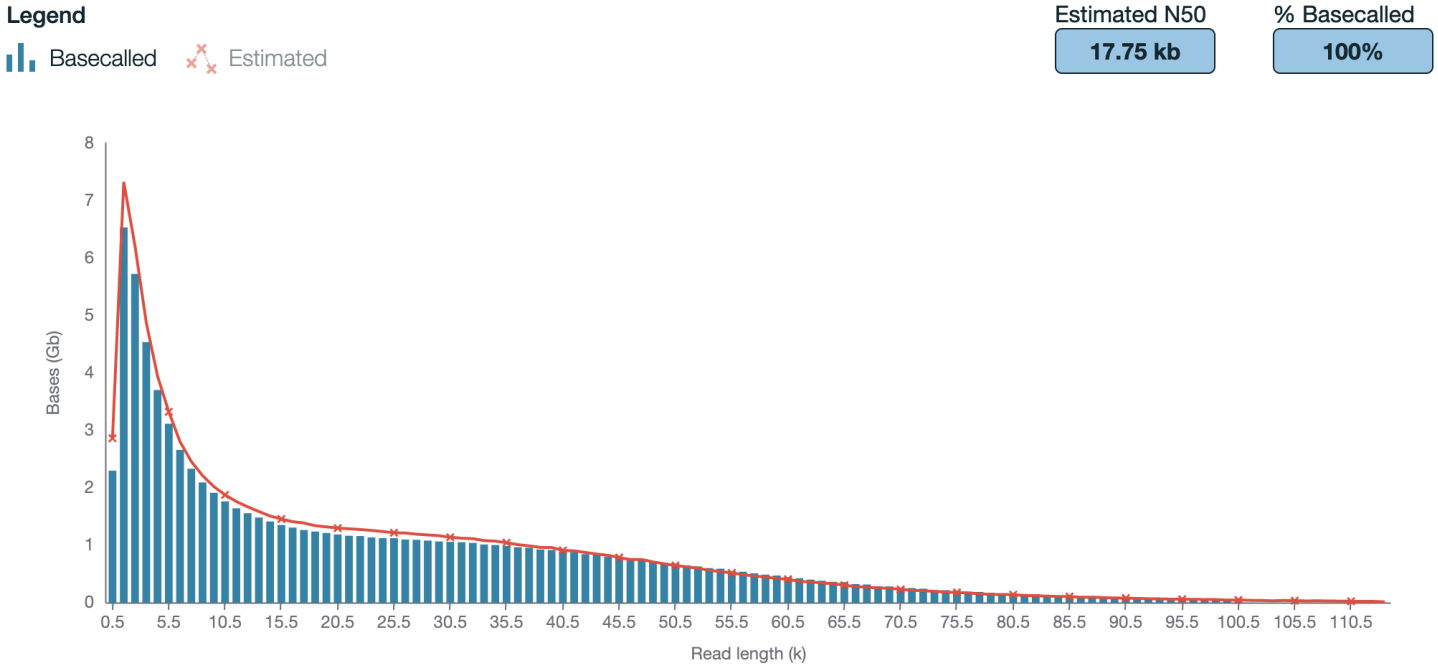
- UMI at the 3' end of the read
- Different lengths indicative of high error rate
- only 47% of reads have fully intact UMI
- 7% have no UMI at all
- Even using some error correction with Levenshtein distance, it's ugly

```
30 TTTCACCCTCCACTTCCCGTCTCAGAATT
29 TTTGAAACAGCTTCACCTTGAACTTT
29 TTTCCAATAAAAAAAATTACAATTT
29 TTTCAGCAAAATAAAATTCCGGTTT
27 TTTGGAGTTGGGGTTGCGCTTGGGGTTT
27 TTTGAGGTTGGAGTTGGGGTTGGCGTTT
24 TTTGGAGTTGGCGTTGCGGTTGGGGTTT
23 TTTGGGGTTGGAATTGGCGTTGGCATTT
23 TTTGGGATTAAGATTGGCATTGCGGTTT
23 TTTAGGGTTCGCGTTGGGGTTGCAGTTT
23 TTTAGGGTTAGCGTTGGAGTTGGGGTTT
22 TTTGGCGTTGGGGTTGGCGTTGGCGTTT
22 TTTGGCGTTGGAGTTCAGCTTACGGTTT
22 TTTGCGGTTGGAGTTGGGCTTGGCGTTT
22 TTTACACTTGTGCTCTCCTTAGCCTTT
21 TTTGGGGTTGGAGTTGGCGTTGGCATTT
21 TTTGGCGTTGGCATTGGCGTTGGGGTTT
21 TTTGGCGTTCGGGTTGGAATTCGCGTTT
```

0.95^28 = 0.237


Duplication rate (%)
0
10
20
30
40
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
1.4e+08
number of ONT reads

What does the data look like?

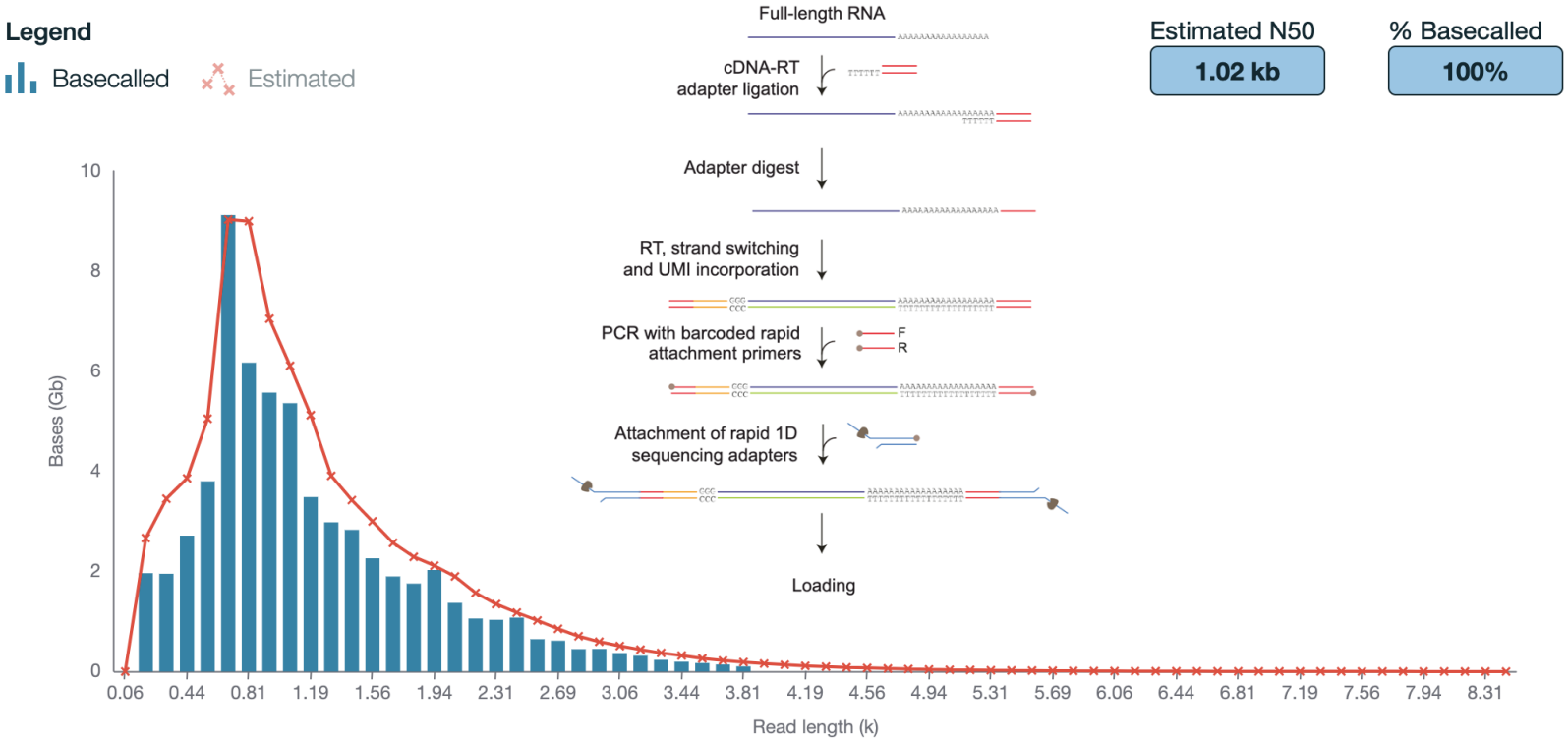


their relative amounts.

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

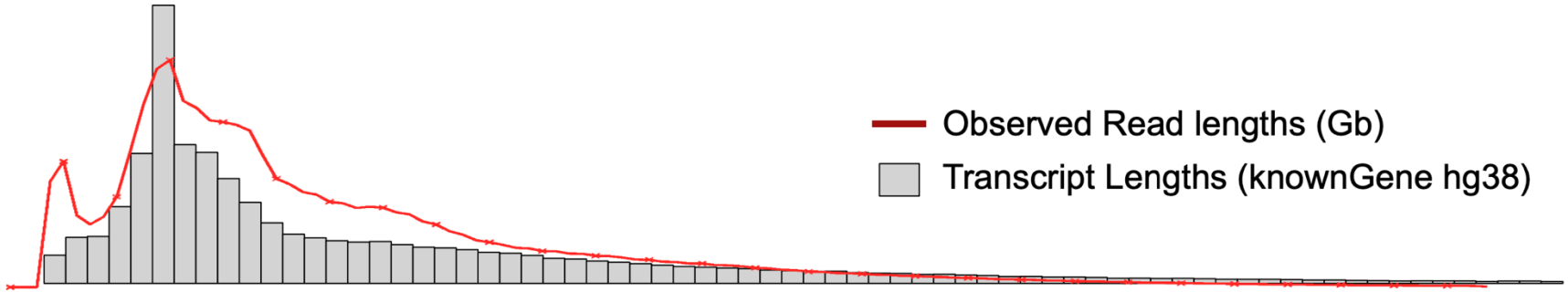
Genomic DNA – standard prep

What does the data look like?

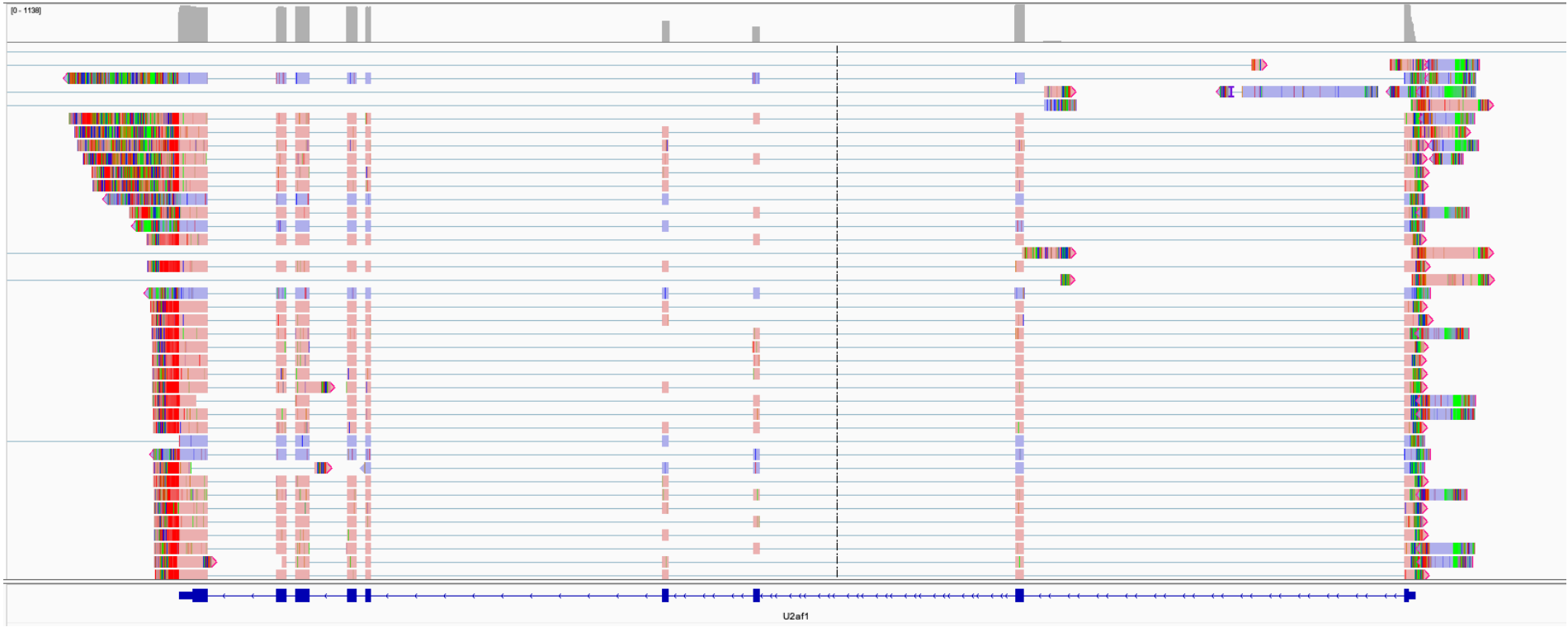


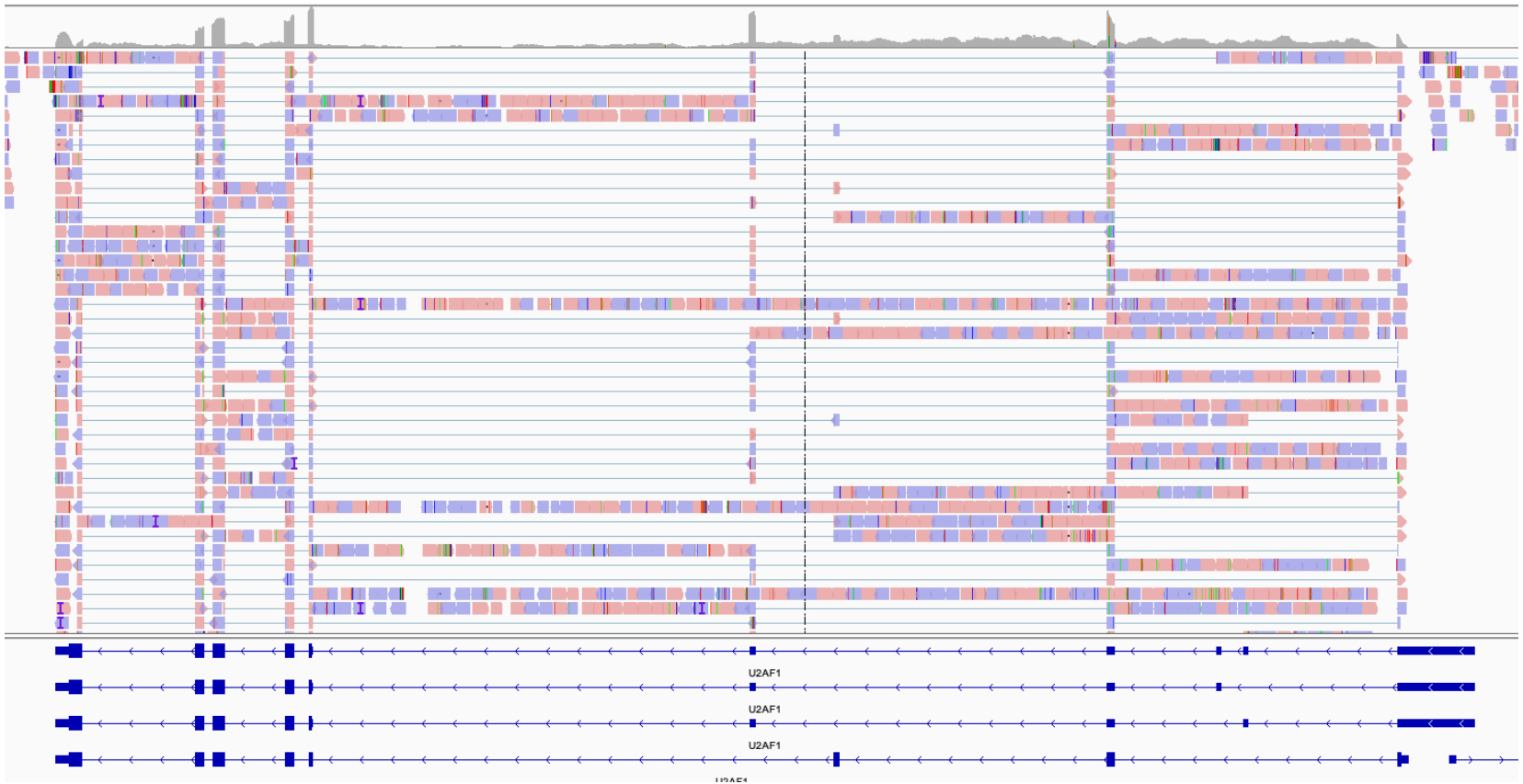
RNA/cDNA – standard prep

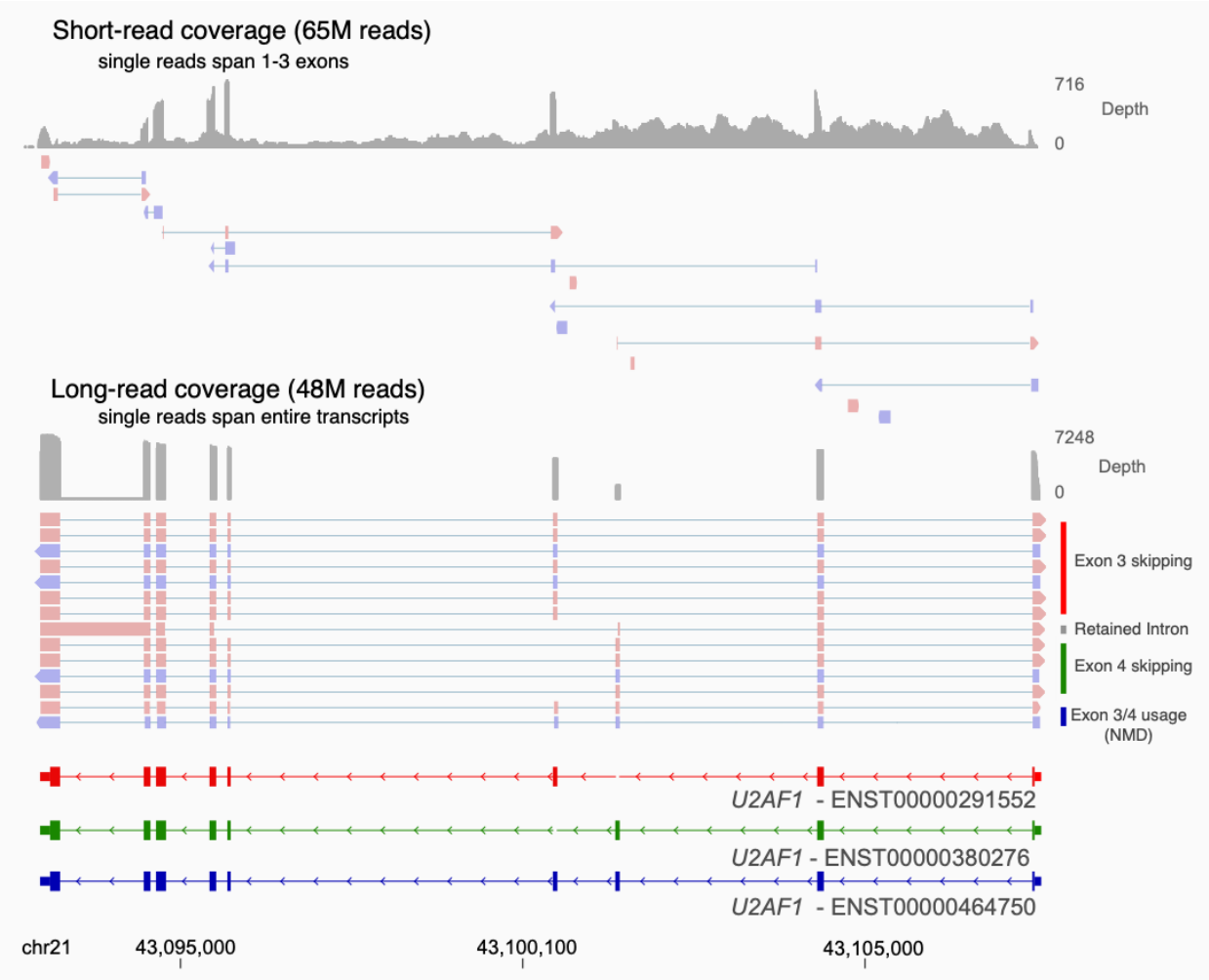
What does the data look like?



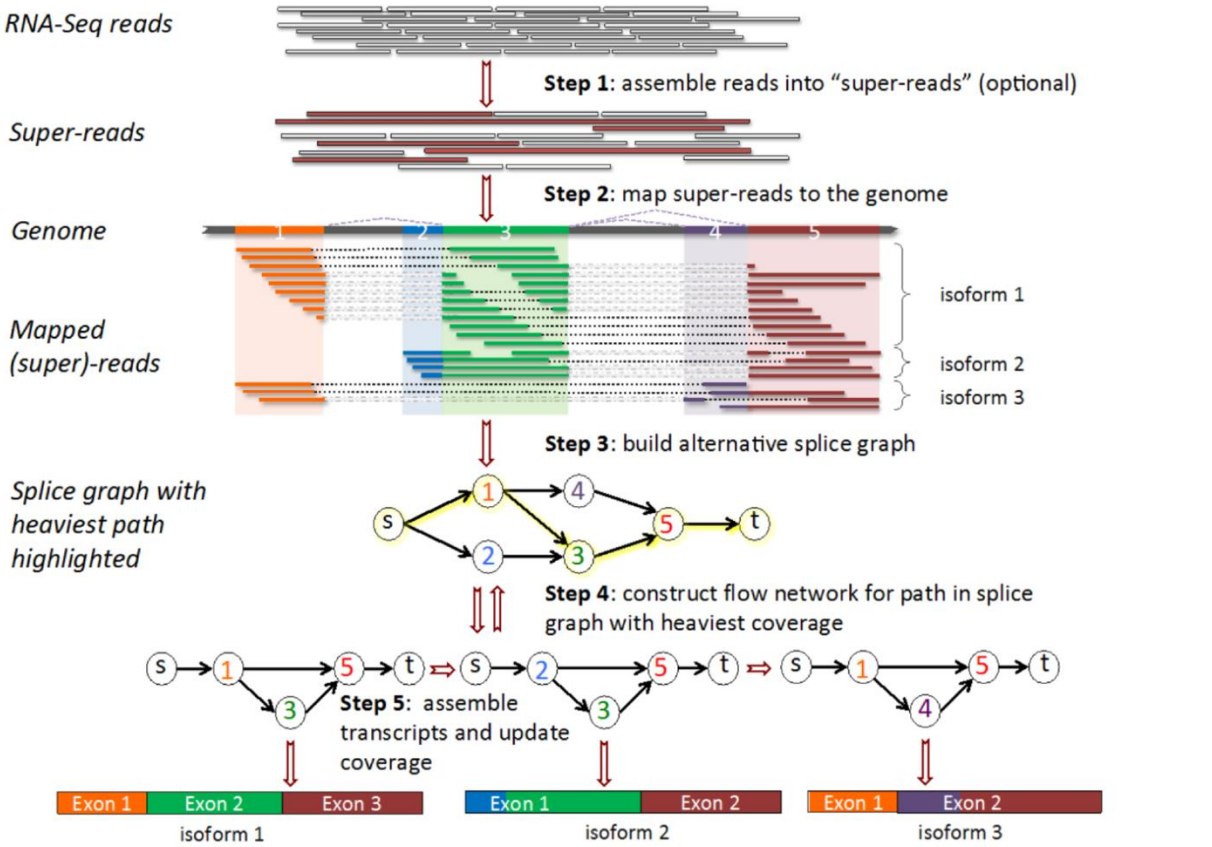
cDNA – standard prep


[0 - 1138]
U2af1


U2AF1
U2AF1
U2AF1

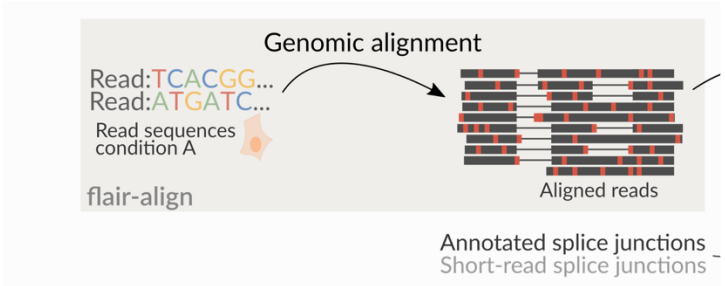

Short-read coverage (65M reads)
single reads span 1-3 exons
716
Depth
0
Long-read coverage (48M reads)
single reads span entire transcripts
7248
Depth
0
Exon 3 skipping
Retained Intron
Exon 4 skipping
Exon 3/4 usage (NMD)
U2AF1 - ENST00000291552
U2AF1 - ENST00000380276
U2AF1 - ENST00000464750
chr21
43,095,000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

Estimating transcript abundance – short-read, Stringtie



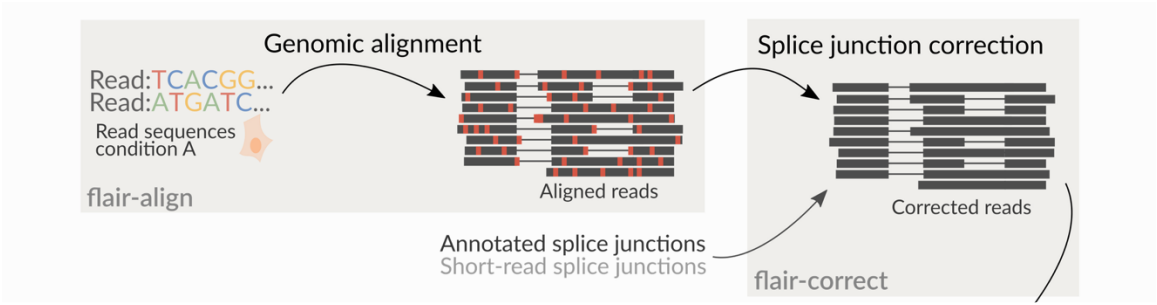
Pertea et al. Nature Biotechnology, 2015

Estimating transcript abundance – long read



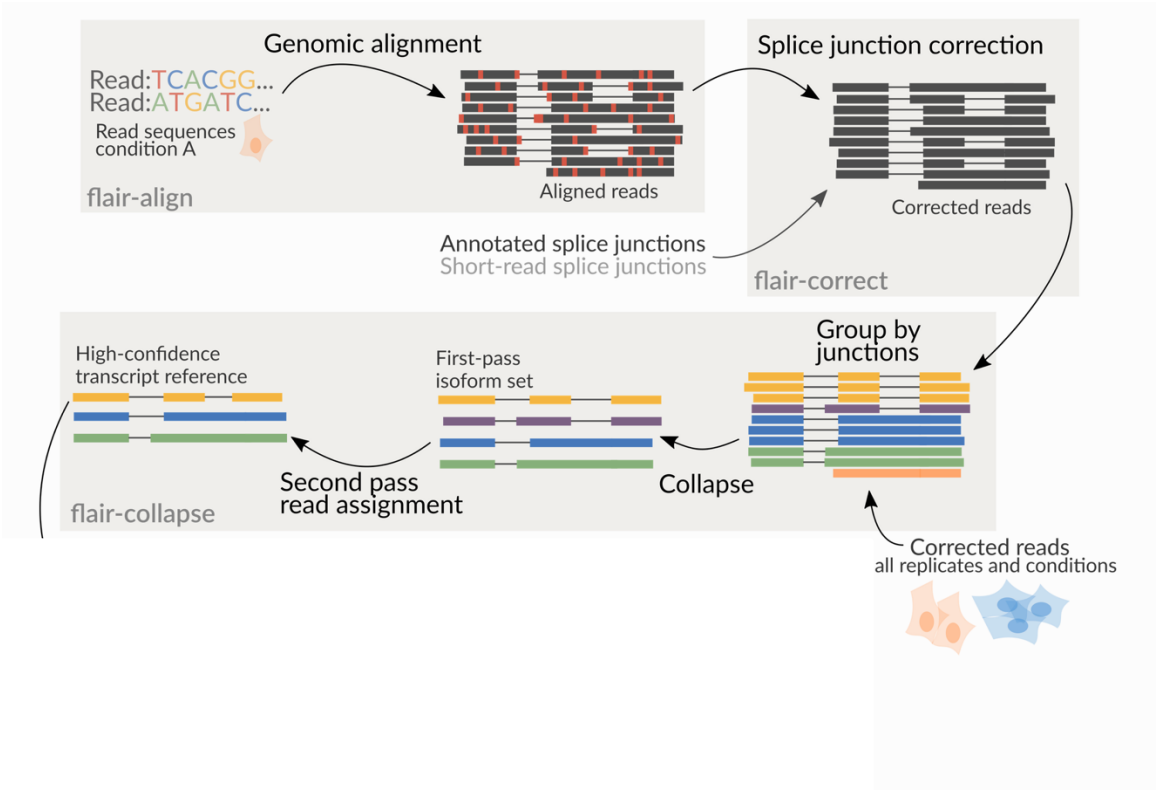
https://github.com/BrooksLabUCSC/flair

Estimating transcript abundance – long read



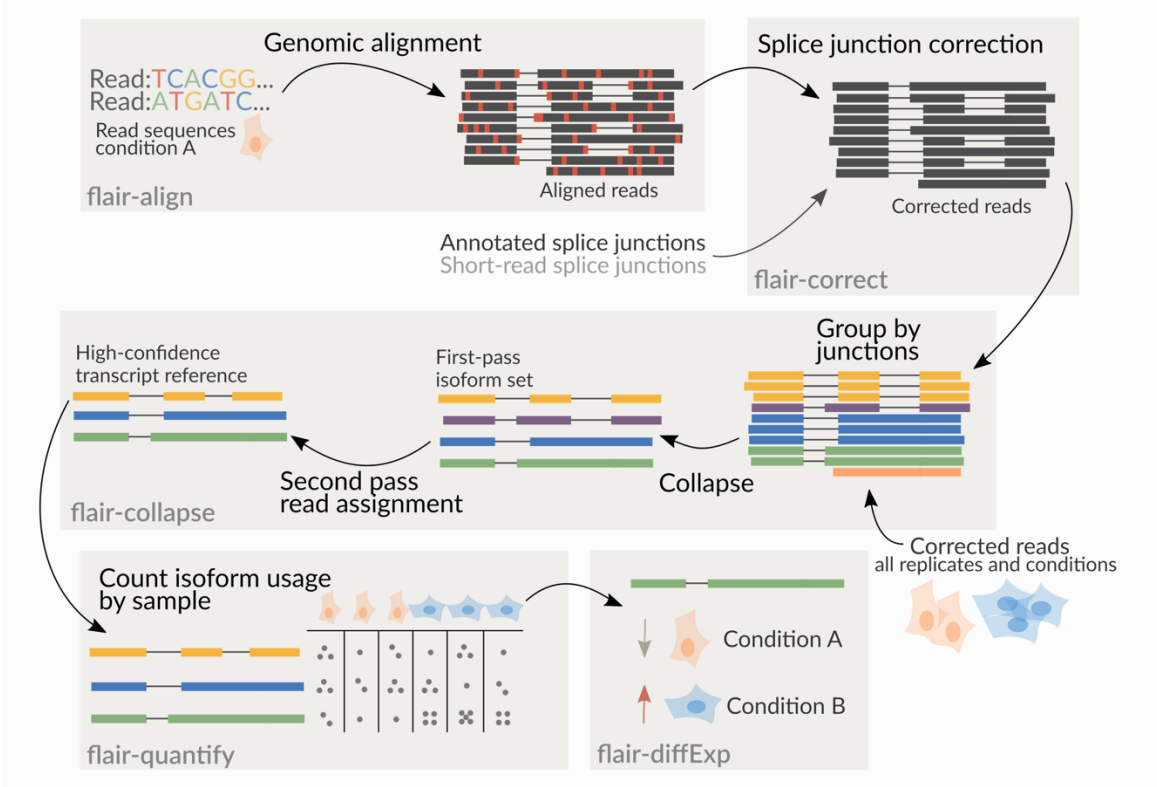
https://github.com/BrooksLabUCSC/flair

Estimating transcript abundance – long read



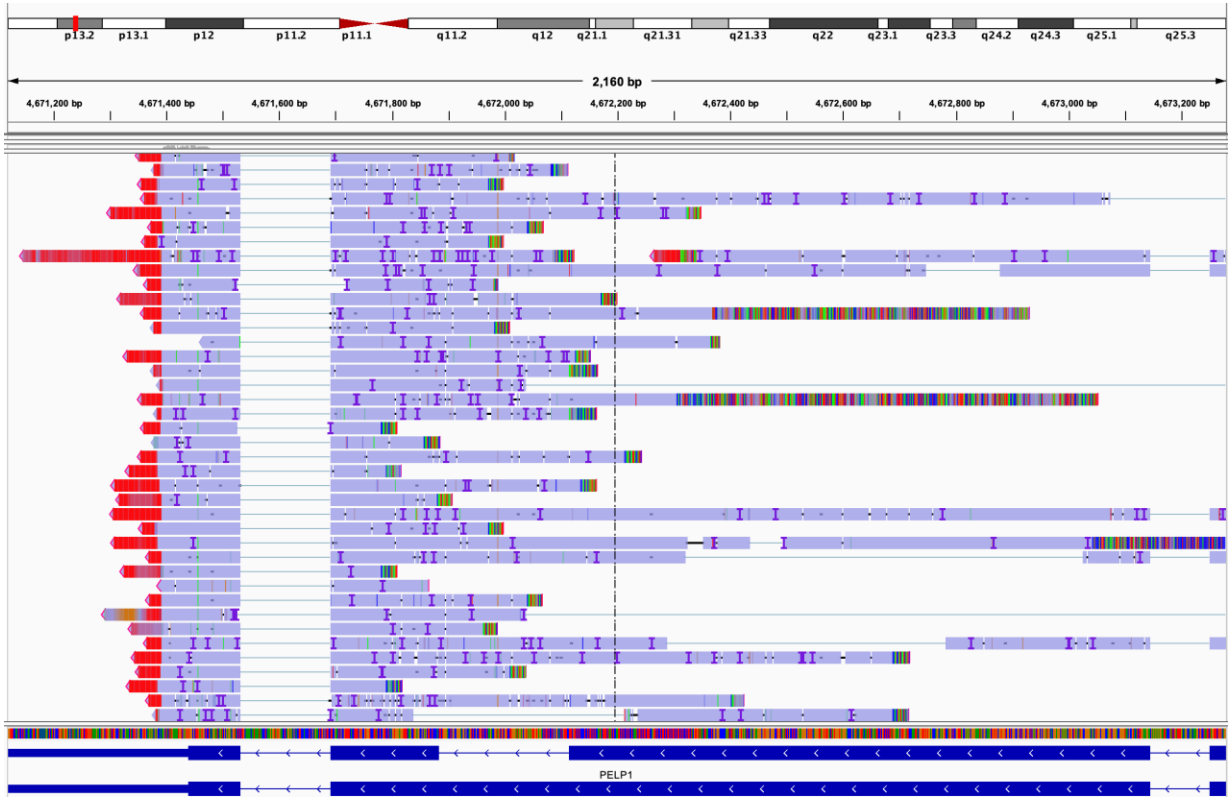
https://github.com/BrooksLabUCSC/flair

Estimating transcript abundance – long read



https://github.com/BrooksLabUCSC/flair

Technical artifacts



Looks like fragmentation of this RNA throughout the long exon. (This is an egregious case – most are more mild)

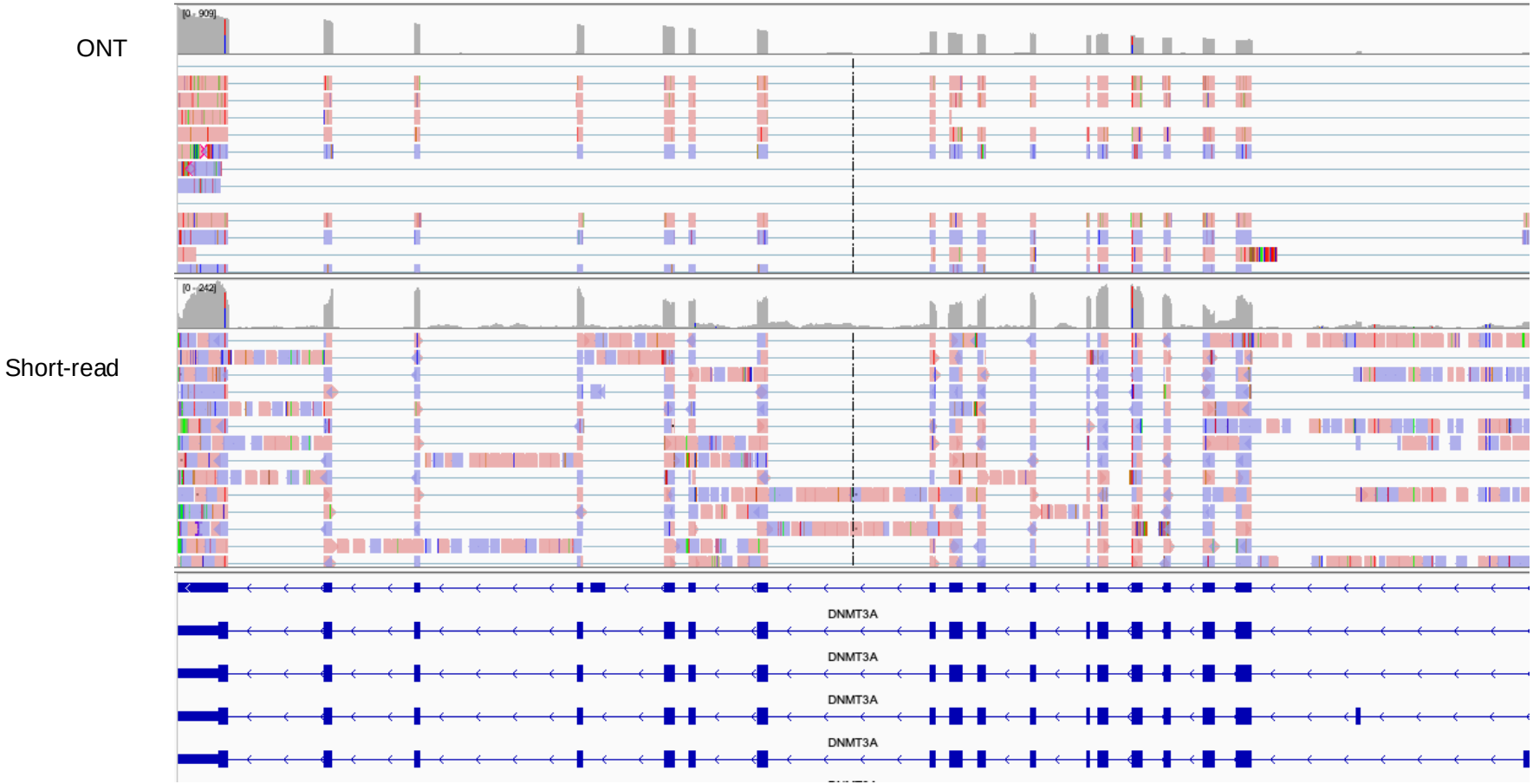
"Ful

Short-read

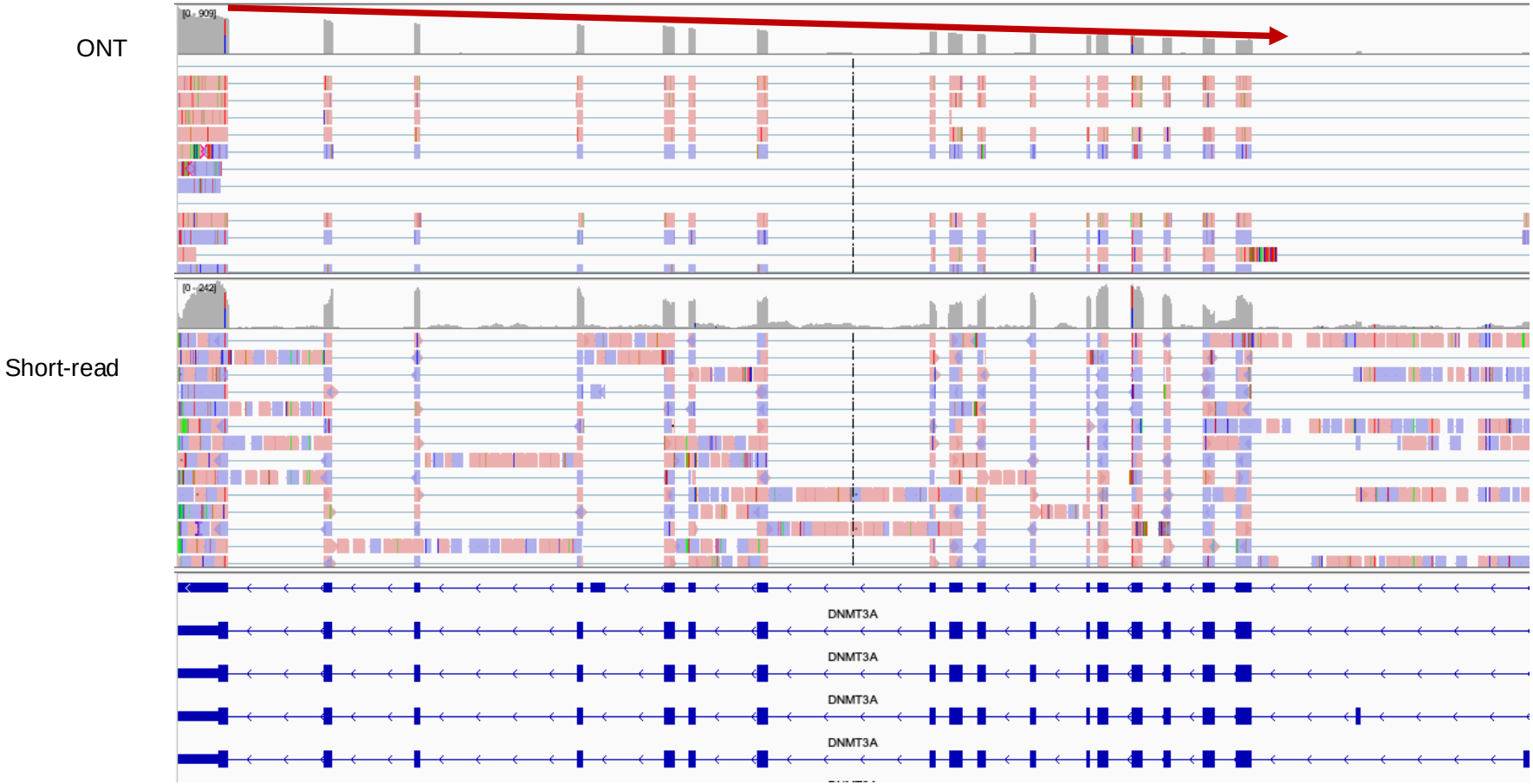
Long-read

ESPRESSO transcript assembly

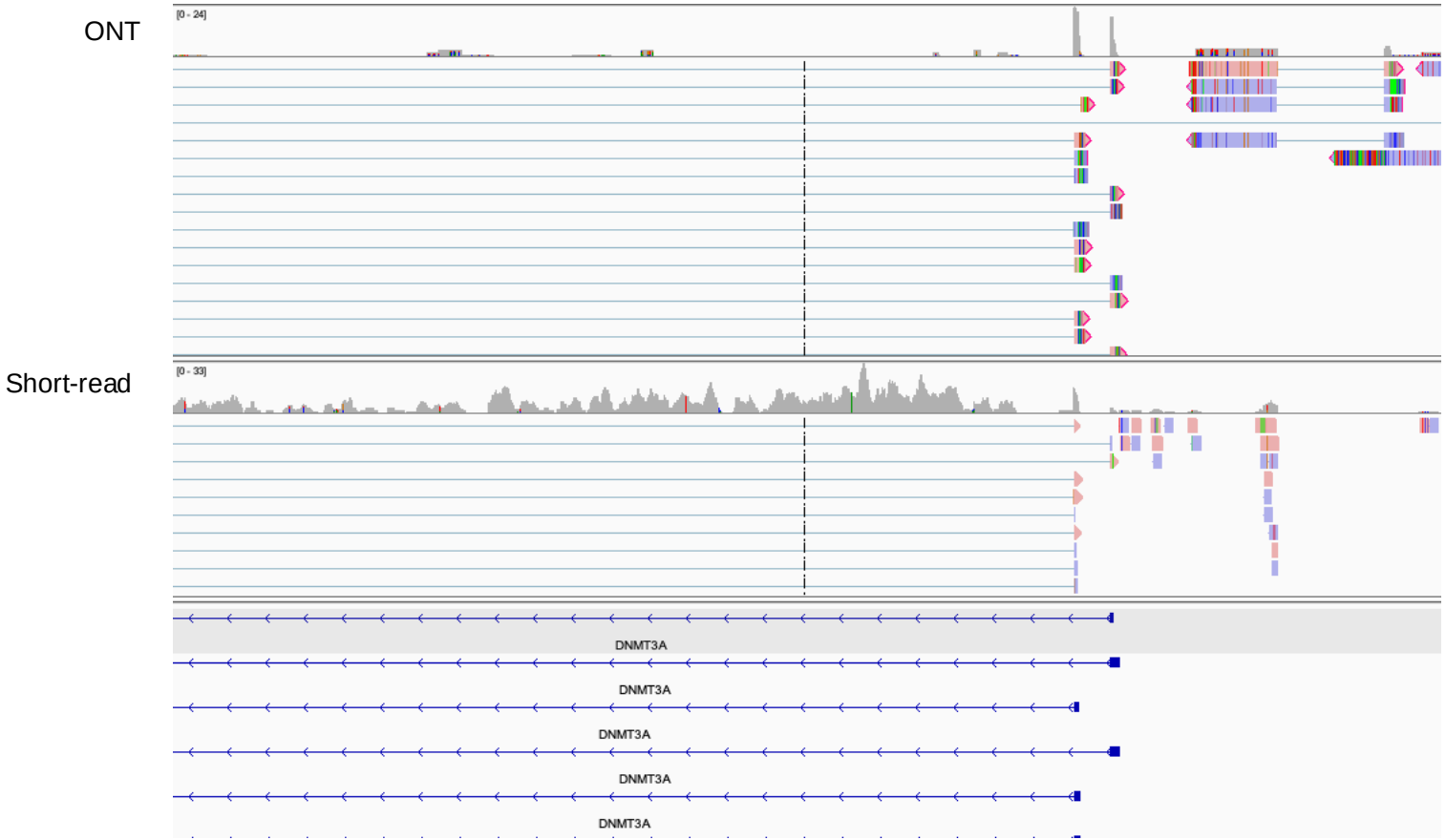
human DNMT3A



human DNMT3A



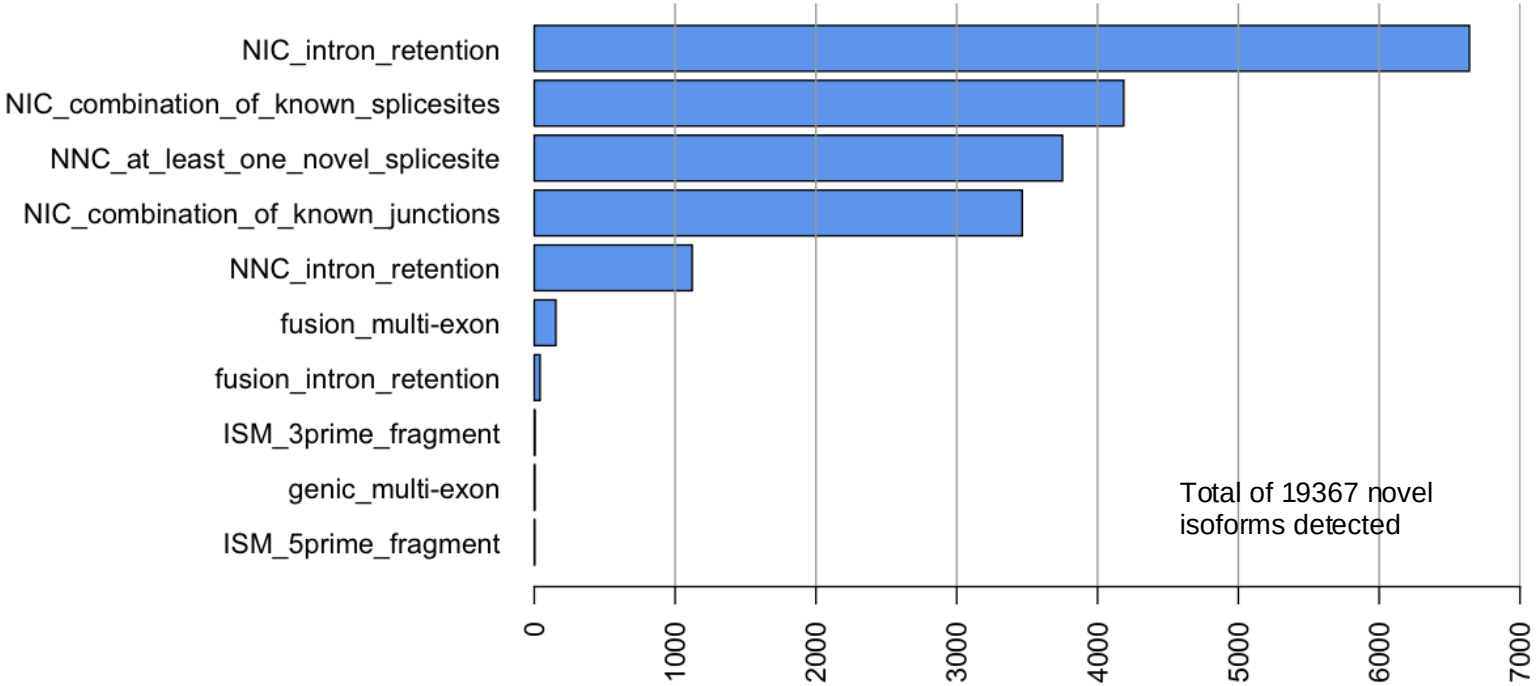
human DNMT3A



Truncated reads

- Appear to be caused by RNA fragmentation
- assessing RIN values of your samples can help – choose clean ones when possible
- When not possible, iteratively assemble transcripts and remove non-full-length reads

Characterization of novel isoforms

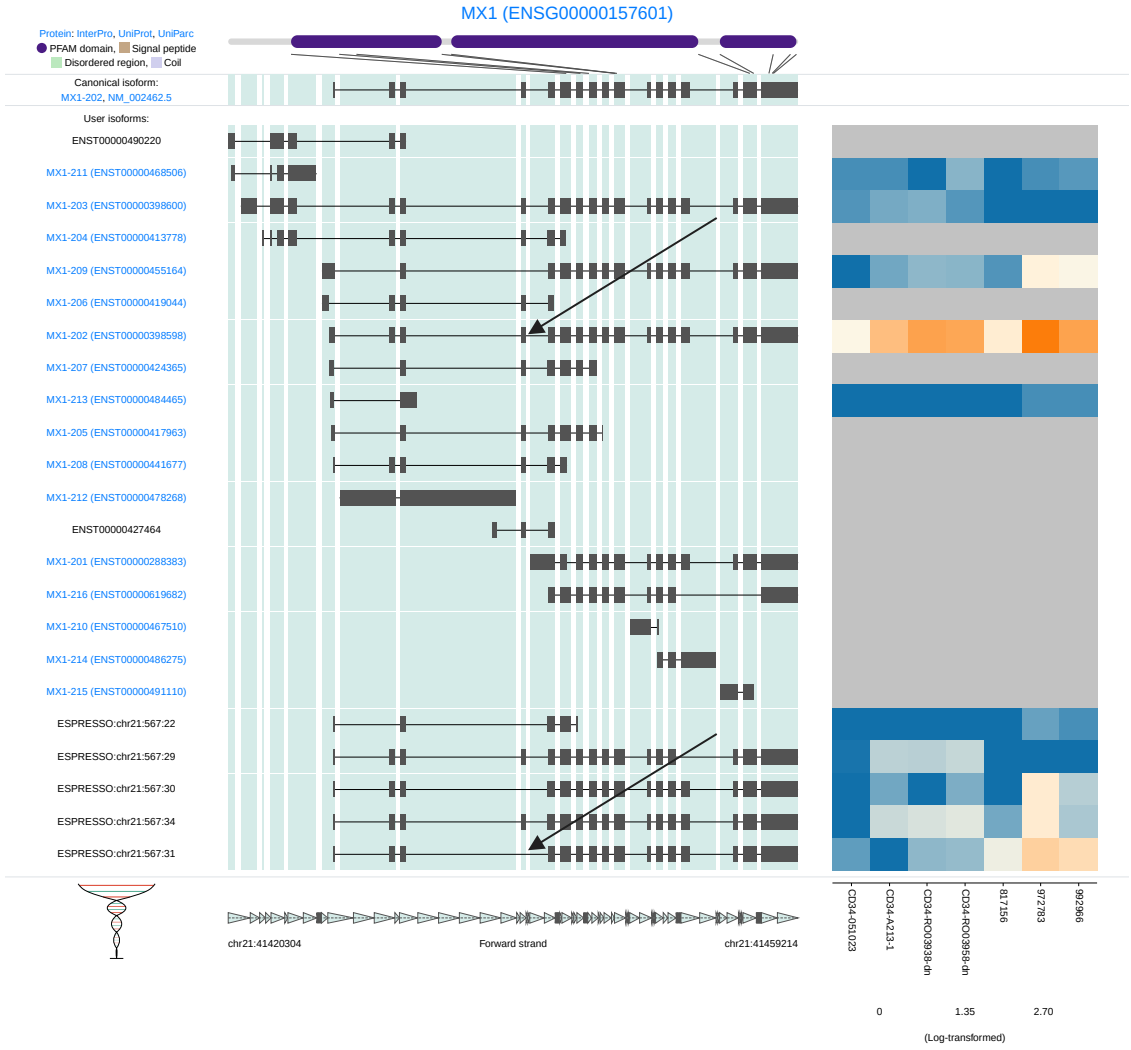


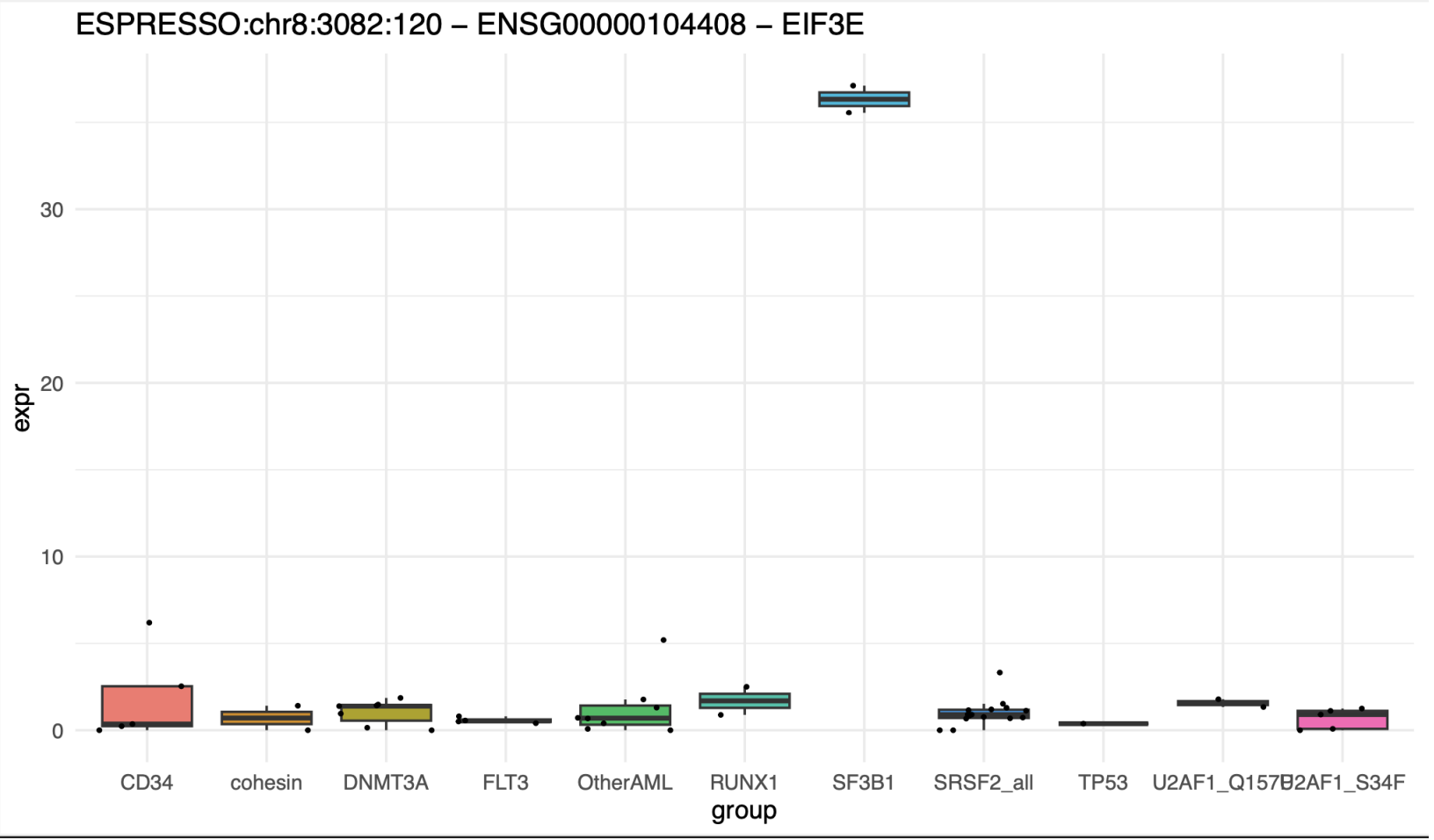
NIC = Novel In Catalog

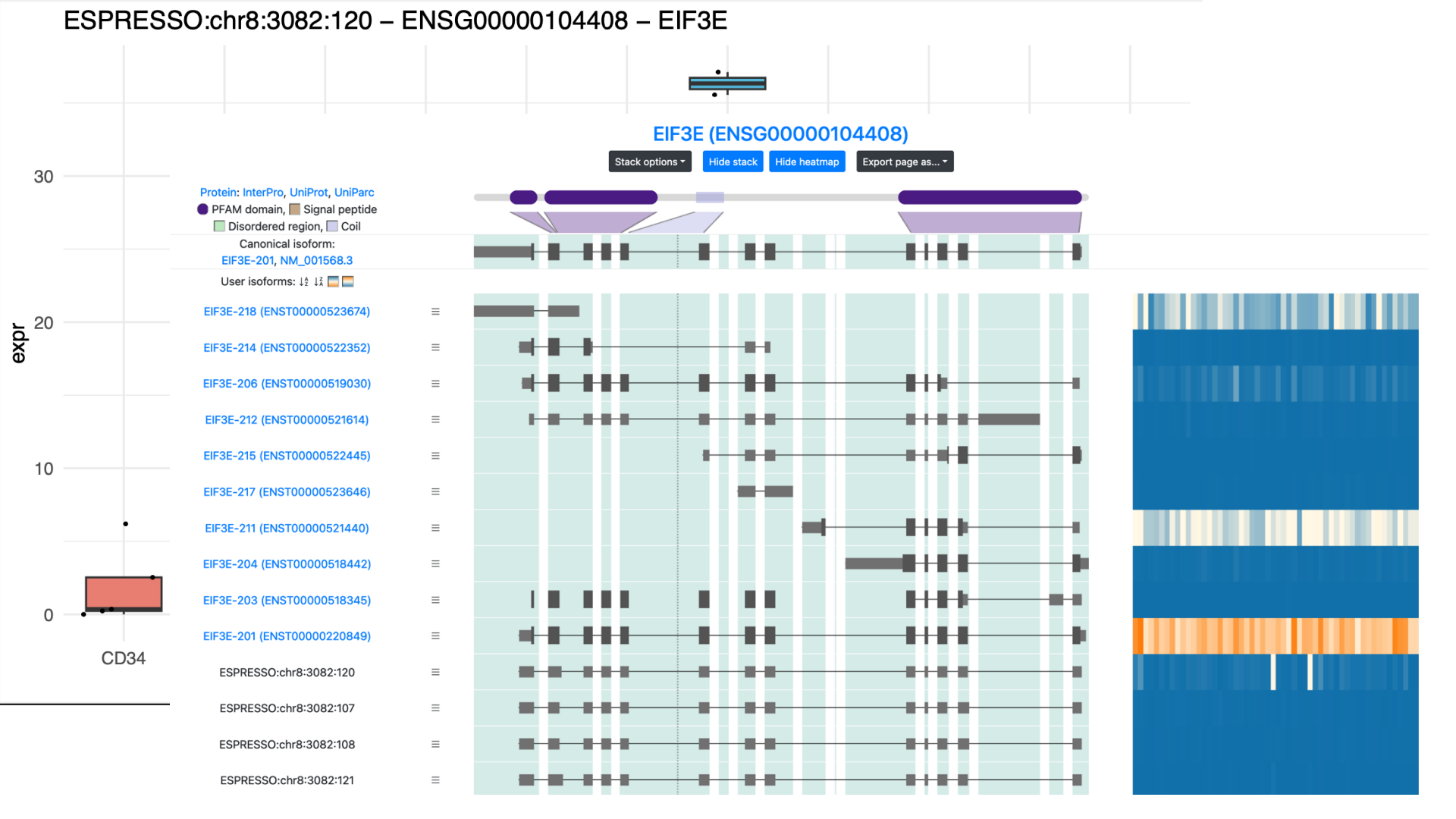
NNC = Novel Not in Catalog

ISM = Incomplete Splice Match

Exon Skipping: isoform expression up in SRSF2 AML




ESPRESSO:chr8:3082:120 – ENSG00000104408 – EIF3E
expr
30
20
10
0
CD34
cohesin
DNMT3A
FLT3
OtherAML
RUNX1
SF3B1
SRSF2_all
TP53
U2AF1_Q157R
U2AF1_S34F
group


ESPRESSO:chr8:3082:120 – ENSG00000104408 – EIF3E
EIF3E (ENSG00000104408)
Stack options
Hide stack
Hide heatmap
Export page as...
Protein: InterPro, UniProt, UniParc
PFAM domain, Signal peptide
Disordered region, Coil
Canonical isoform:
EIF3E-201, NM_001568.3
User isoforms:
EIF3E-218 (ENST00000523674)
EIF3E-214 (ENST00000522352)
EIF3E-206 (ENST00000519030)
EIF3E-212 (ENST00000521614)
EIF3E-215 (ENST00000522445)
EIF3E-217 (ENST00000523646)
EIF3E-211 (ENST00000521440)
EIF3E-204 (ENST00000518442)
EIF3E-203 (ENST00000518345)
EIF3E-201 (ENST00000220849)
ESPRESSO:chr8:3082:120
ESPRESSO:chr8:3082:107
ESPRESSO:chr8:3082:108
ESPRESSO:chr8:3082:121
expr
30
20
10
0
CD34

Assignment

- Start with some long-read RNAseq data from a cell line
- QC the data, trim adapters
- Align the reads
- Examine a few genes to see how the data looks