Long Read Sequencing

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

How to sequence a human genome: Sanger method

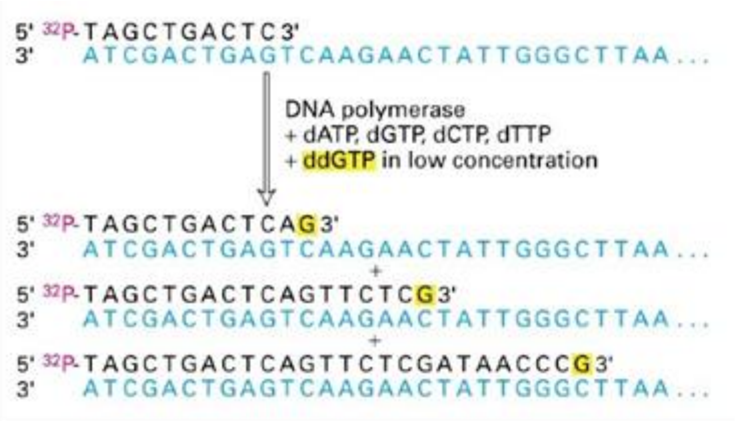
Key points:

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



5' ^{32}P-TAGCTGACTC3'
3' ATCGACTGAGTCAAGAACTATTGGGCTTAA...

DNA polymerase
+dATP, dGTP, dCTP, dTTP
+ddGTP in low concentration

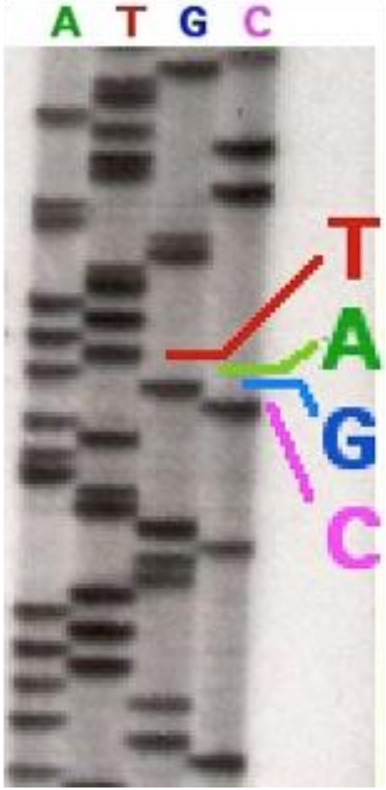
5' ^{32}P-TAGCTGACTCAG3'
3' ATCGACTGAGTCAAGAACTATTGGGCTTAA...

+

5' ^{32}P-TAGCTGACTCAGTTCTCG3'
3' ATCGACTGAGTCAAGAACTATTGGGCTTAA...

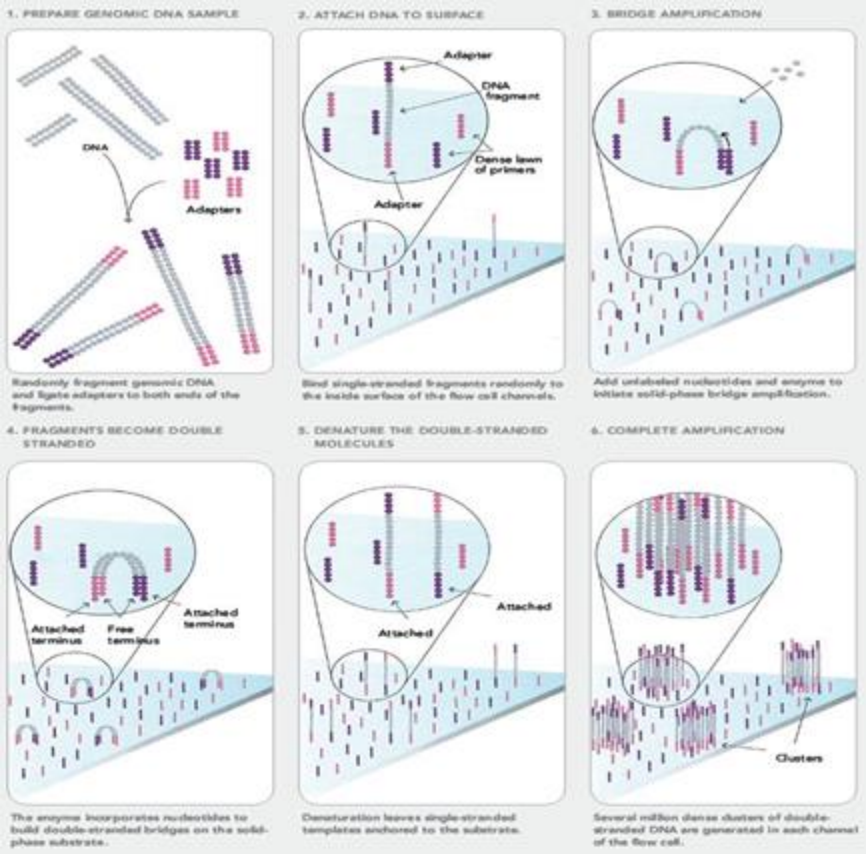
+

5' ^{32}P-TAGCTGACTCAGTTCTCGATAACCCG3'
3' ATCGACTGAGTCAAGAACTATTGGGCTTAA...



Solexa (Illumina) sequencing (2006)

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



1. PREPARE GENOMIC DNA SAMPLE

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

4. FRAGMENTS BECOME DOUBLE STRANDED

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES

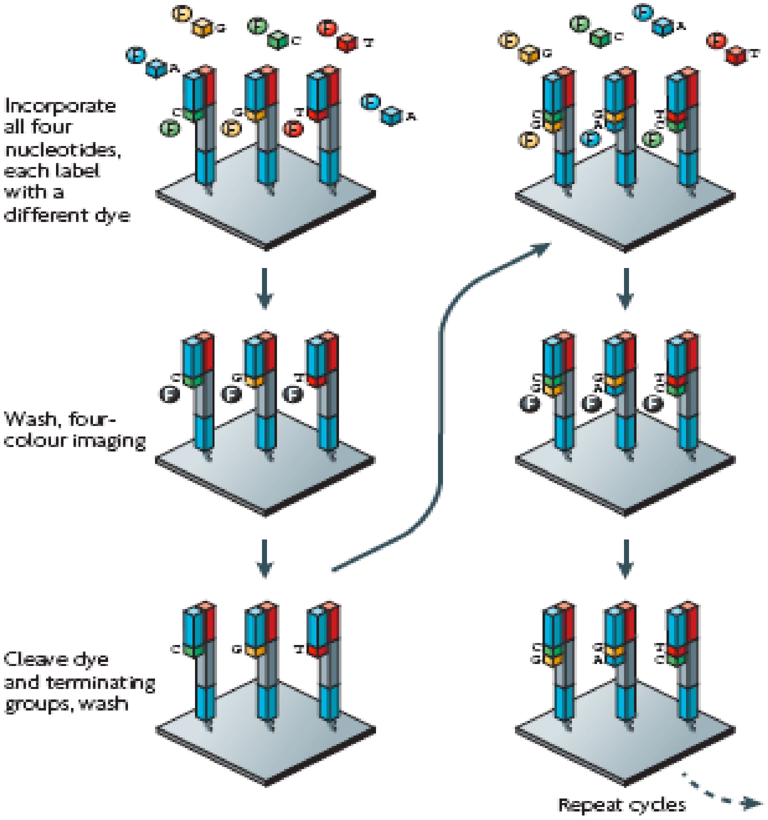
Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION

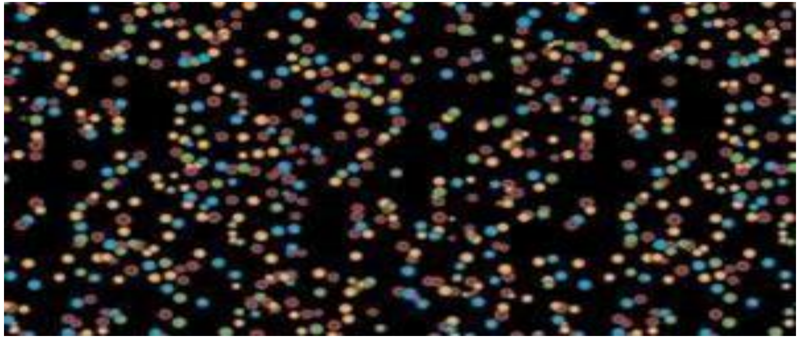
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Slide from Ira Hall

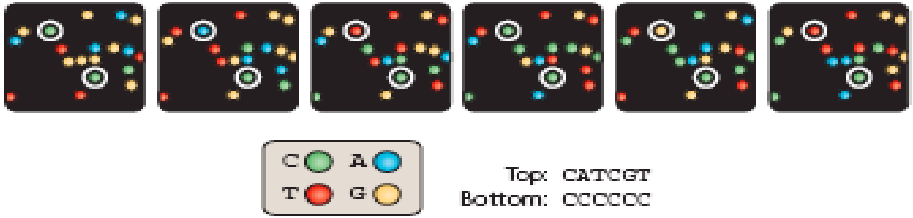
Illumina sequencing (2005)


Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
Repeat cycles
G
C
T
A
F

4 different images merged

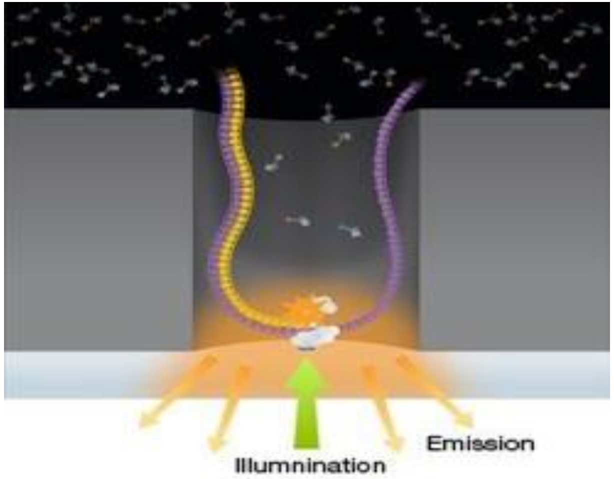


6 cycles w/ base-calling



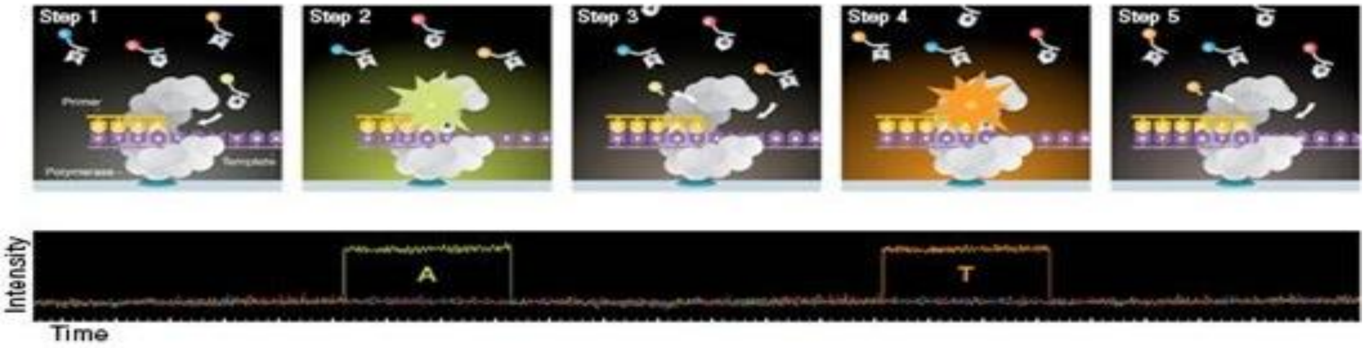
Slide from Ira Hall

Pacific Biosciences



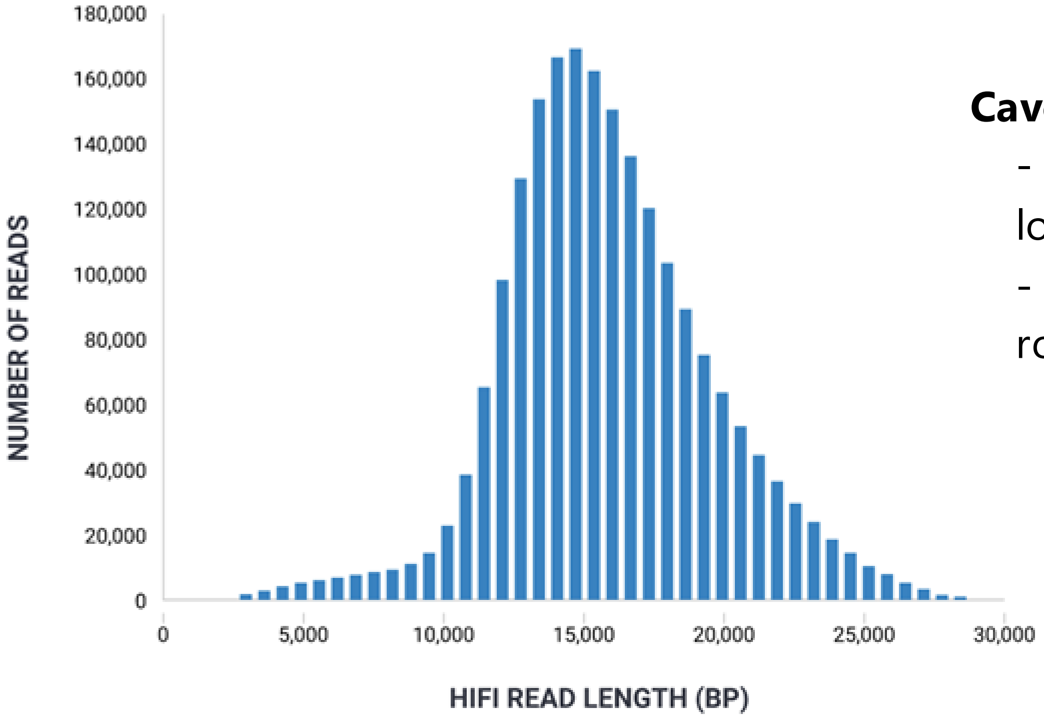
Key Points:

- 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



Slide from Ira Hall

Pacific Biosciences: long reads. Great for genome assembly

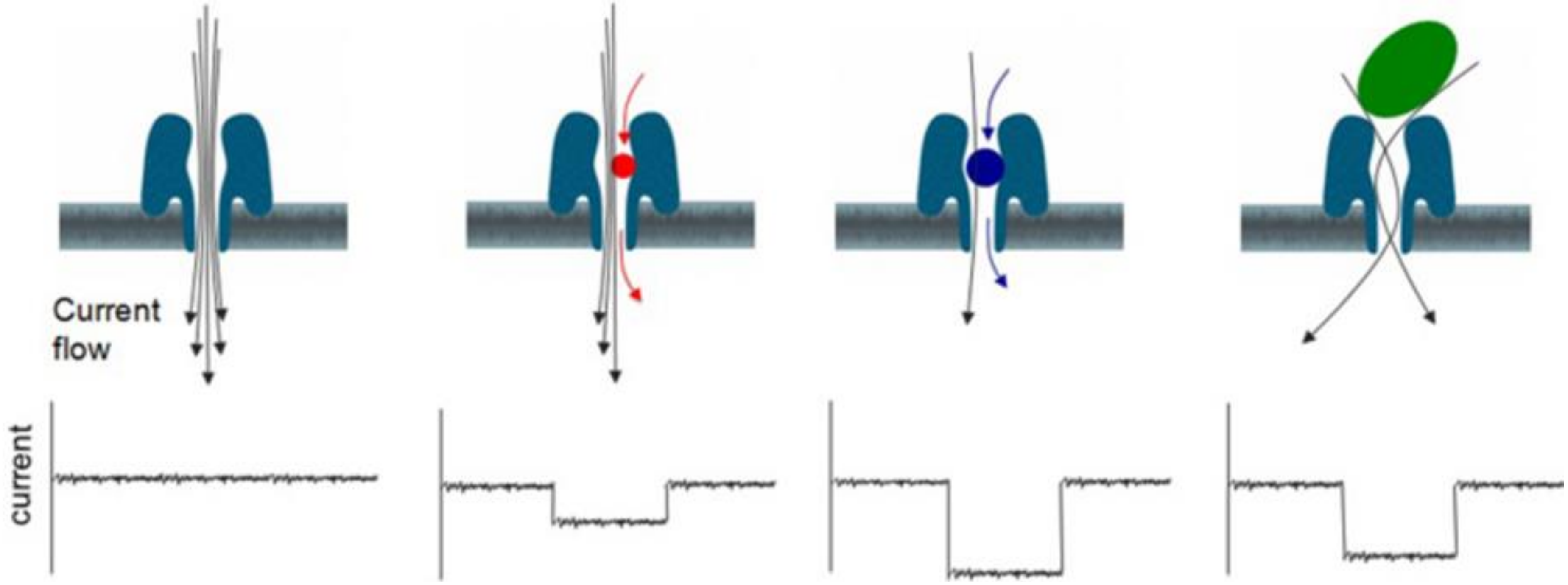


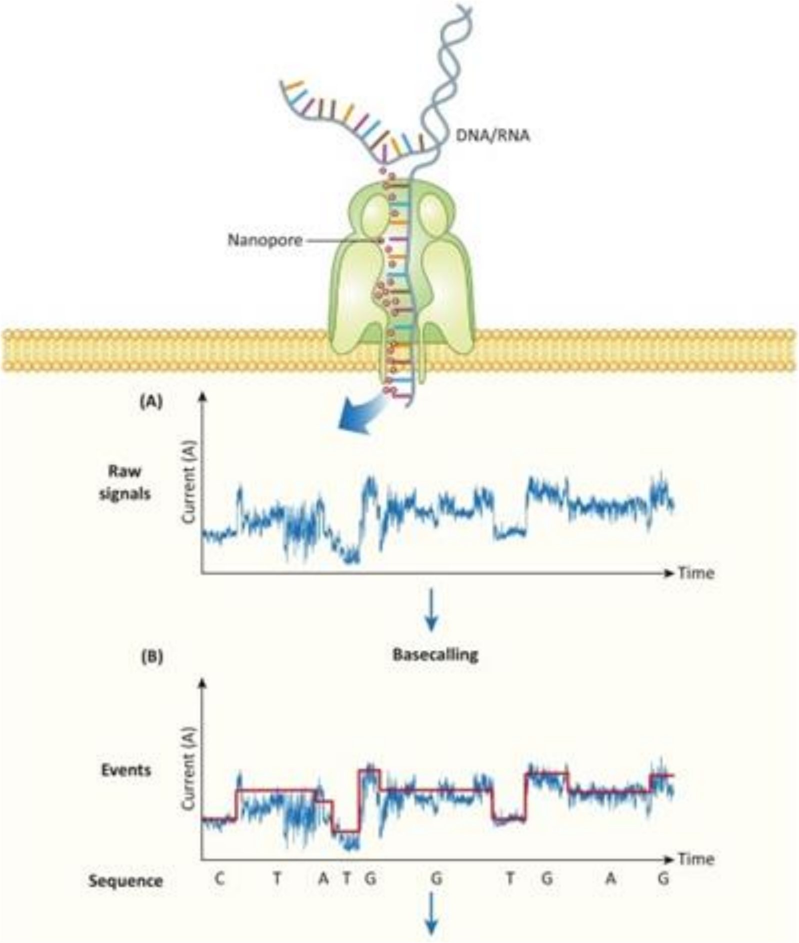
Caveats:

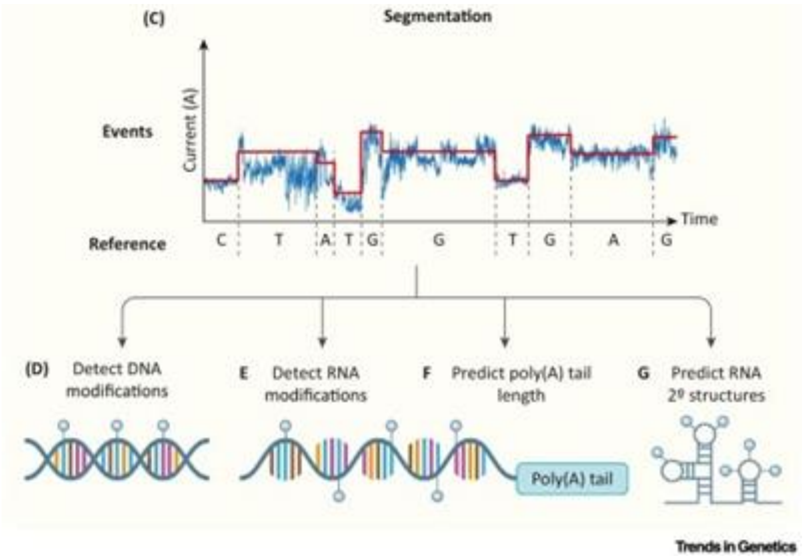
- higher error rate (1-2%), lower with Duplex runs
- lower throughput : roughly 90 gigabases per run

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

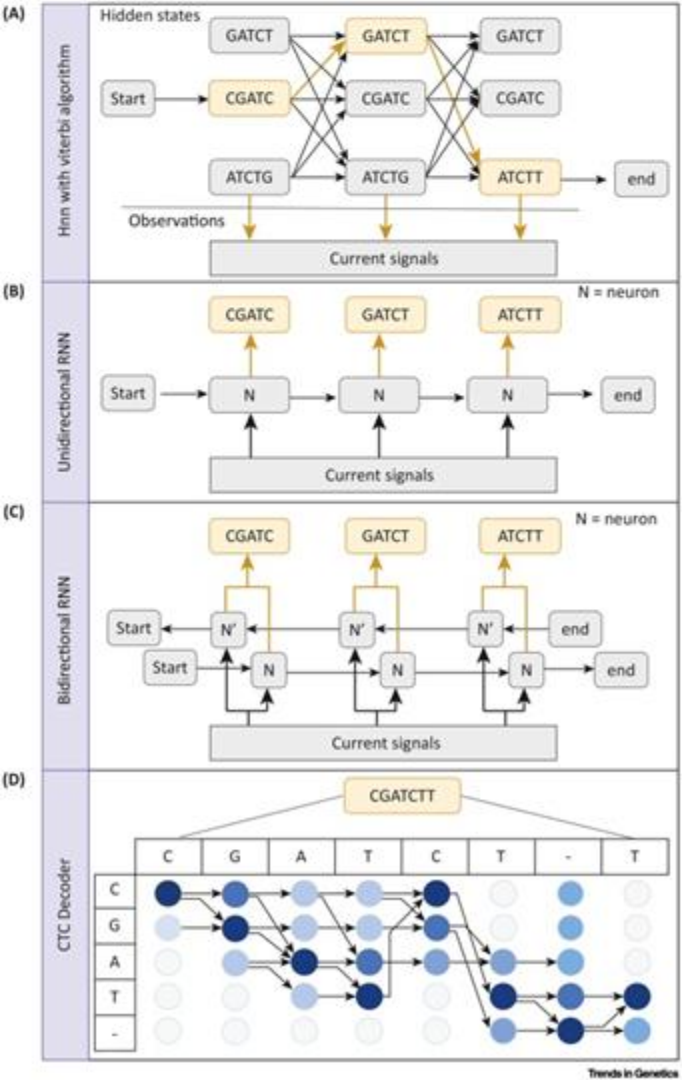
Oxford Nanopore Technologies






(C)
Segmentation
Events
Current (A)
Time
Reference
C T A T G G T G A G
(D) Detect DNA modifications
E Detect RNA modifications
F Predict poly(A) tail length
G Predict RNA 2º structures
Poly(A) tail
Trends in Genetics

doi.org/10.1016/j.tig.2021.09.001



Neural networks to translate signal into base calls

- Guppy (many versions)

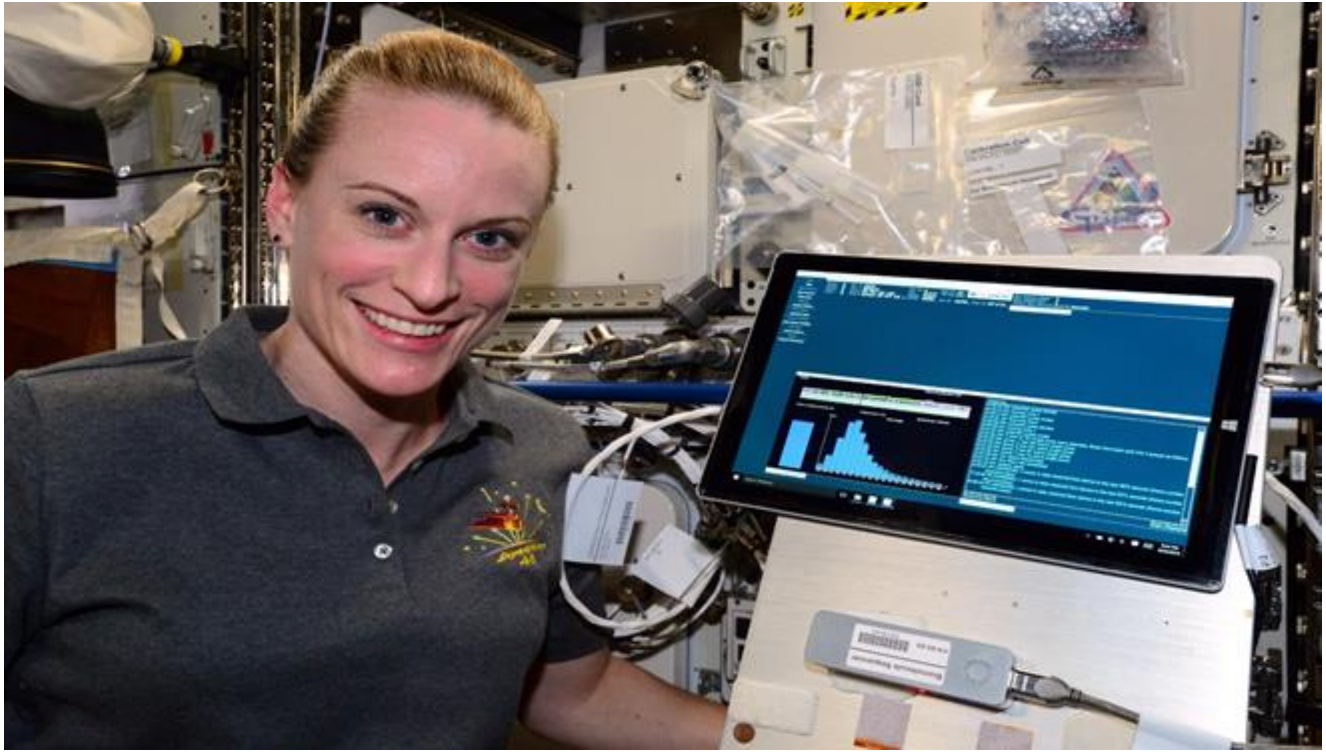
- Dorado (v0.4, eventual guppy replacement)

- many others

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

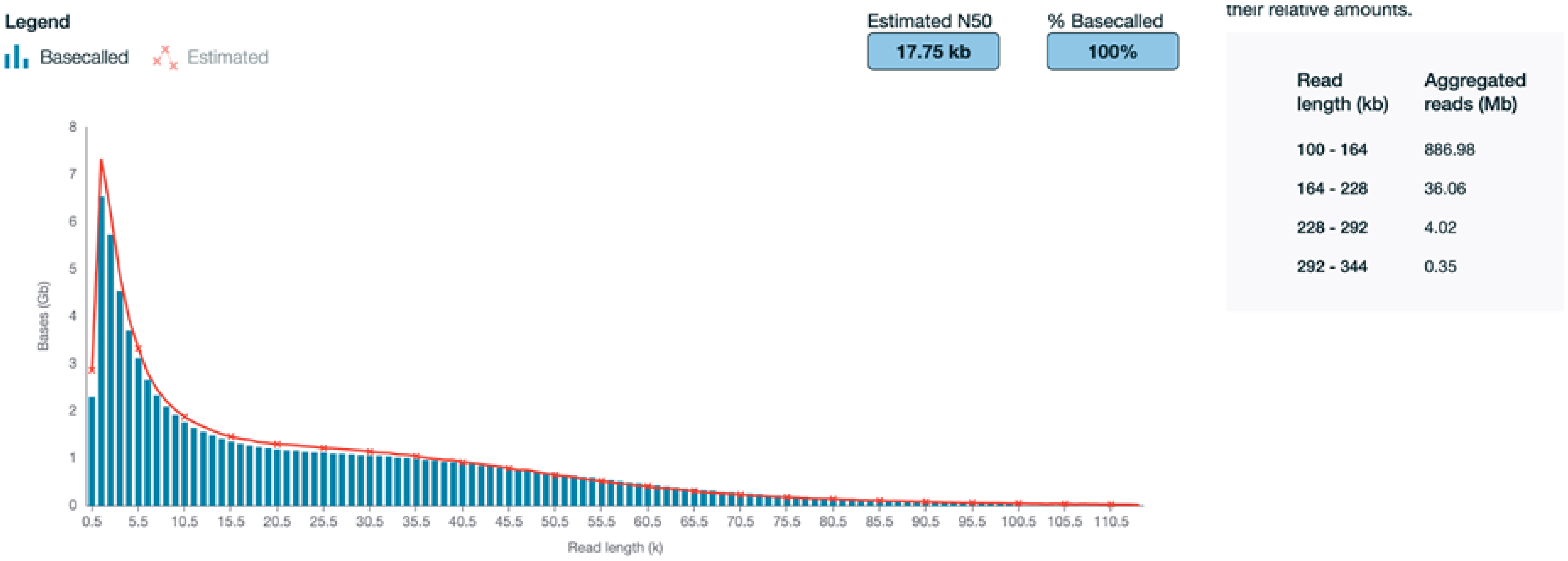
doi.org/10.1016/j.tig.2021.09.001

Nanopore sequencing is *extremely* portable



Kate Rubins sequencing DNA on the ISS

ONT sequence length distribution



Estimated N50

17.75 kb

% Basecalled

100%

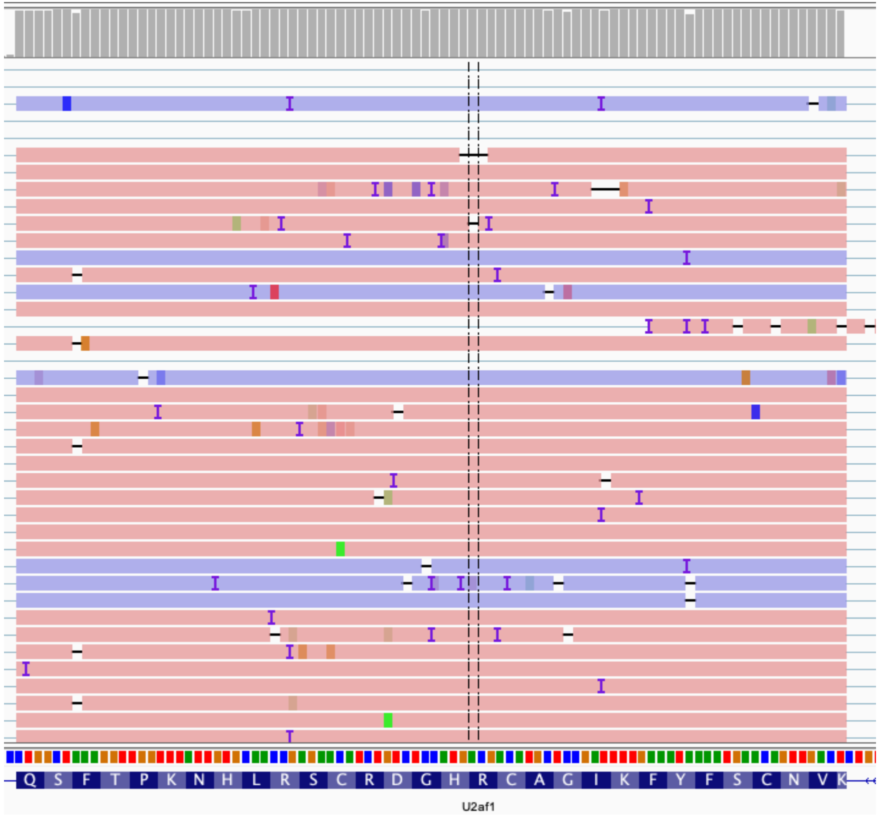
their relative amounts.

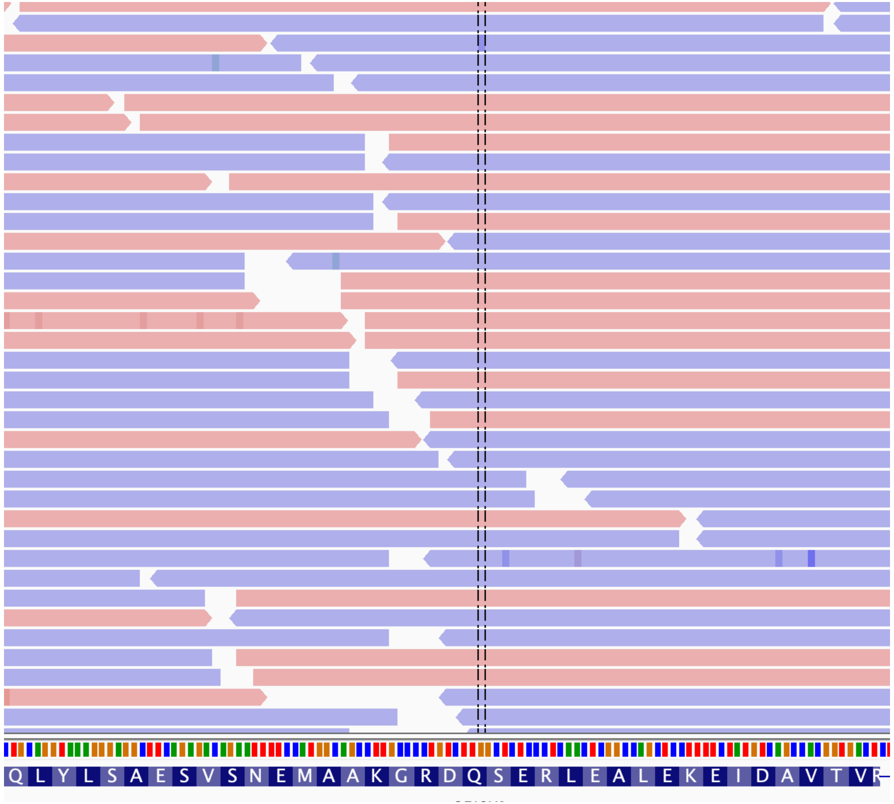
Read length (kb)	Aggregated reads (Mb)
100 - 164	886.98
164 - 228	36.06
228 - 292	4.02
292 - 344	0.35

Recent run of a tumor sample

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

What does the data look like?



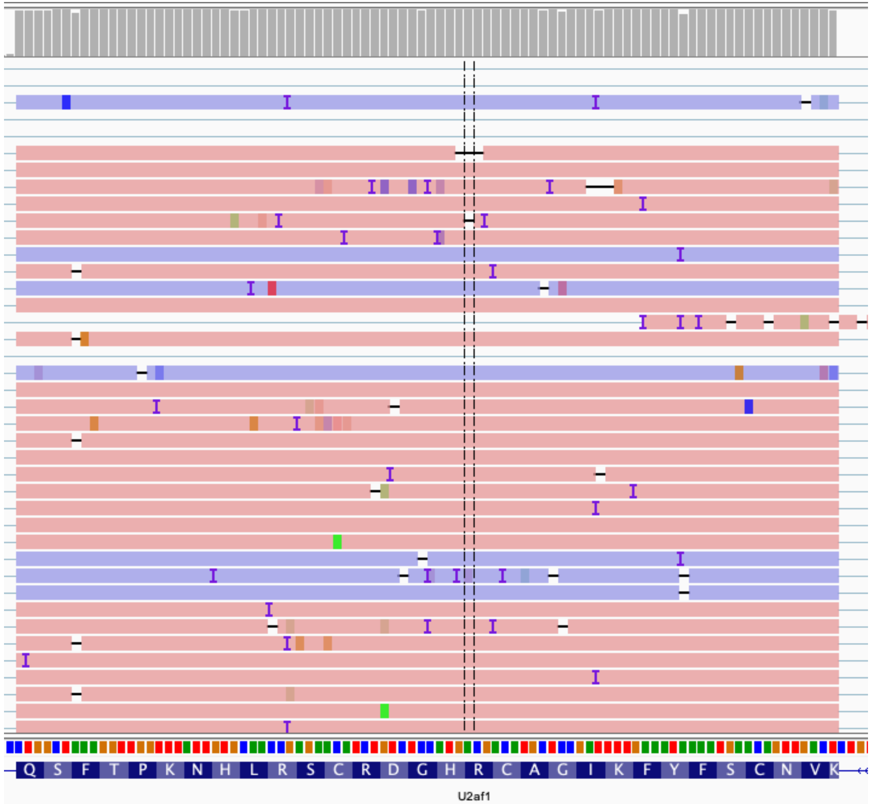

Q L Y L S A E S V S N E M A A K G R D Q S E R L E A L E K E I D A V T V R

Long-read ONT ~5% base error rate

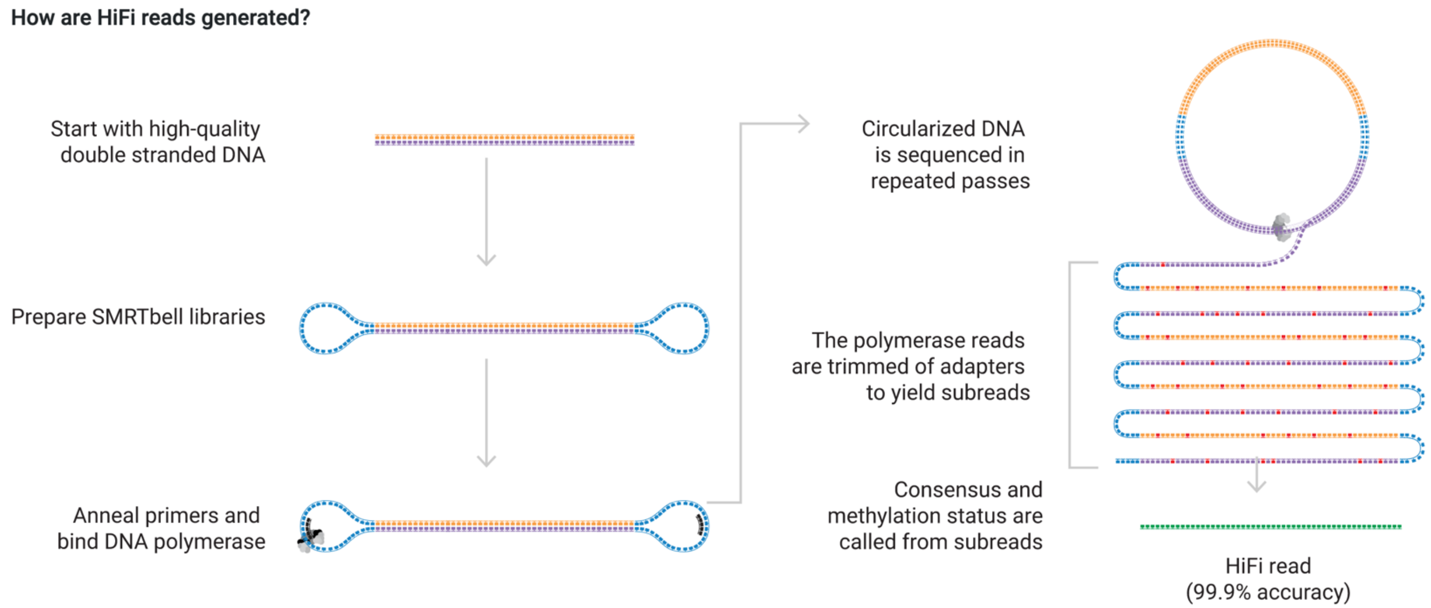
Short-read Illumina ~0.3% base error rate

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

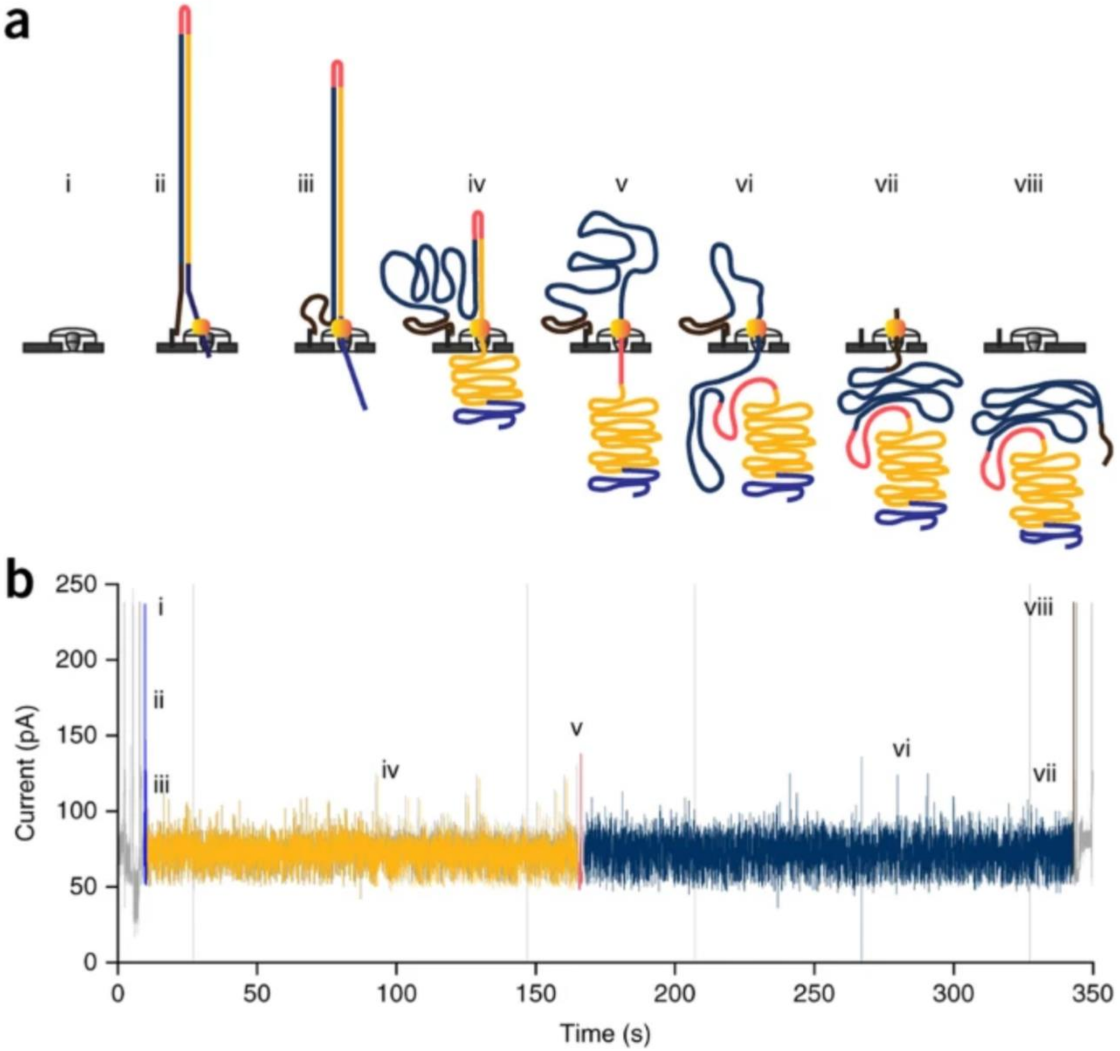

How are HiFi reads generated?
Start with high-quality double stranded DNA
Prepare SMRTbell libraries
Anneal primers and bind DNA polymerase
Circularized DNA is sequenced in repeated passes
The polymerase reads are trimmed of adapters to yield subreads
Consensus and methylation status are called from subreads
HiFi read (99.9% accuracy)

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

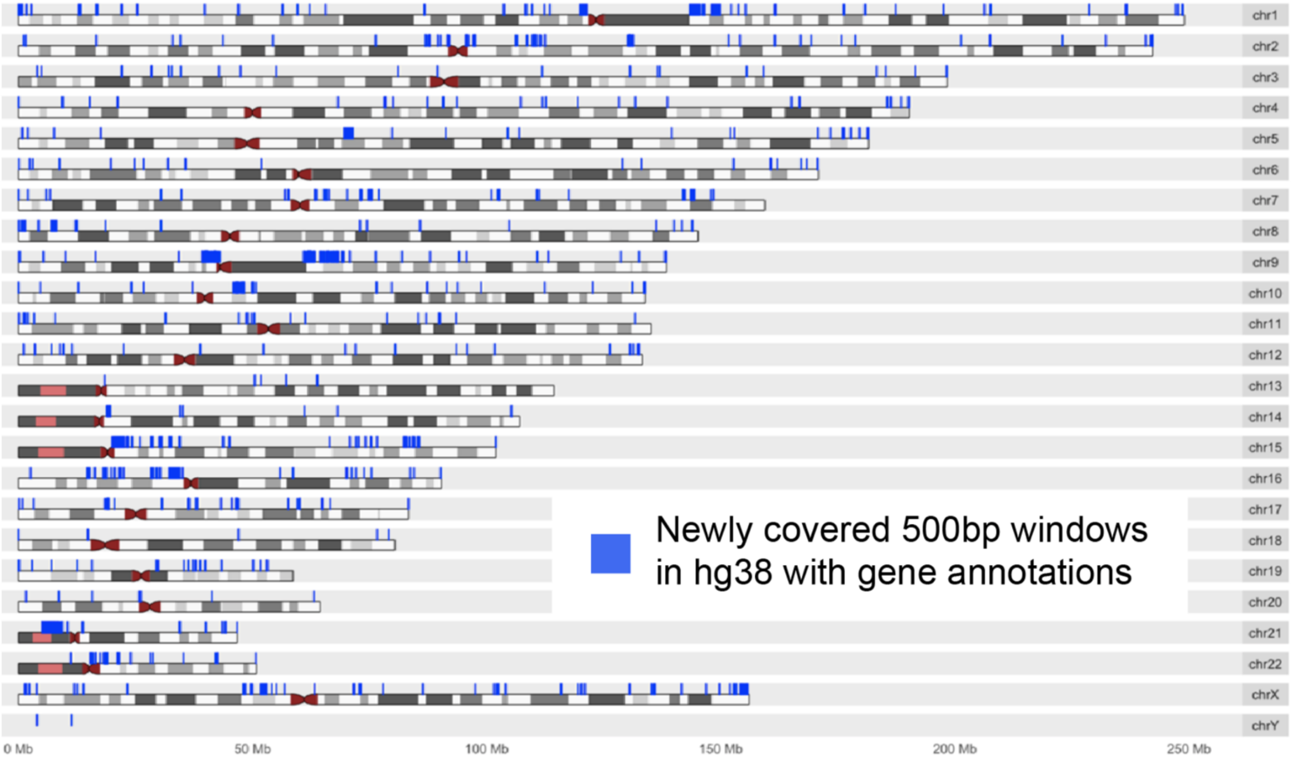
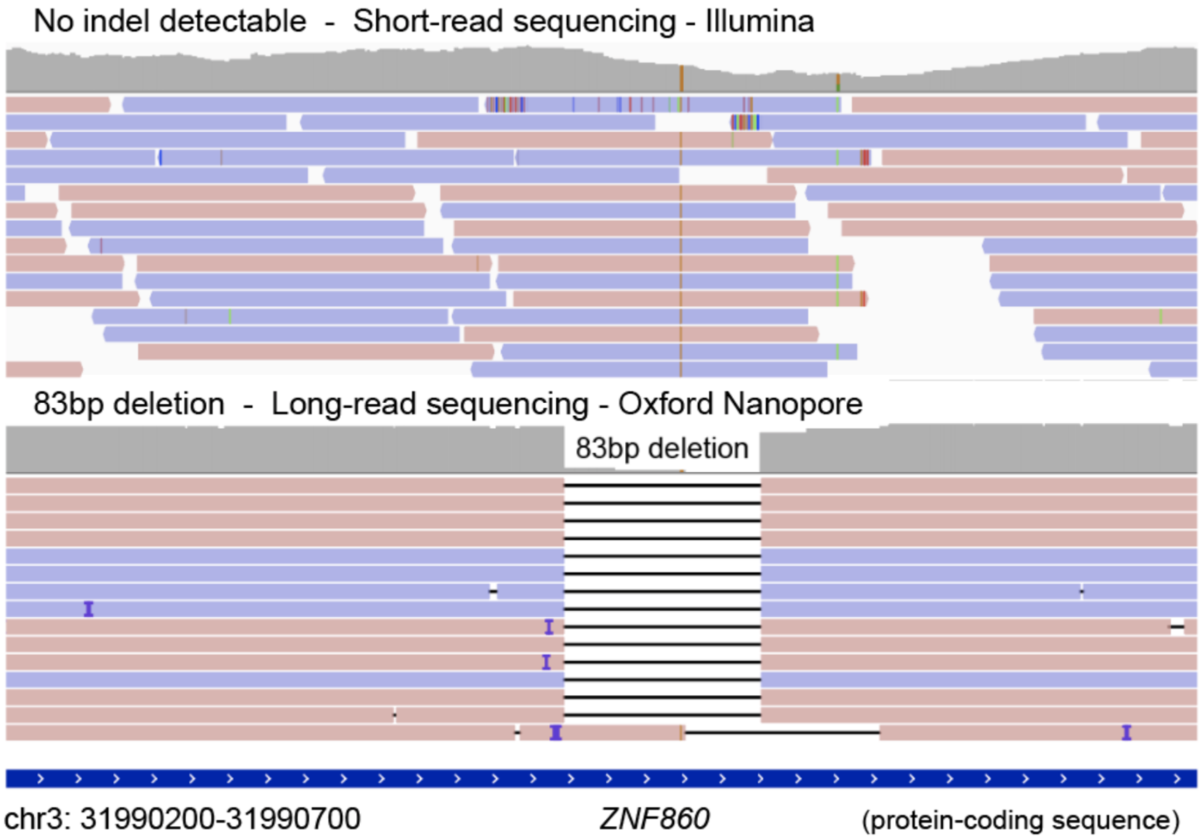


Figure 1: Blue-labeled genomic regions are accessible to long reads but not short, and have functional annotations (e.g. genes or enhancers)

David Spencer

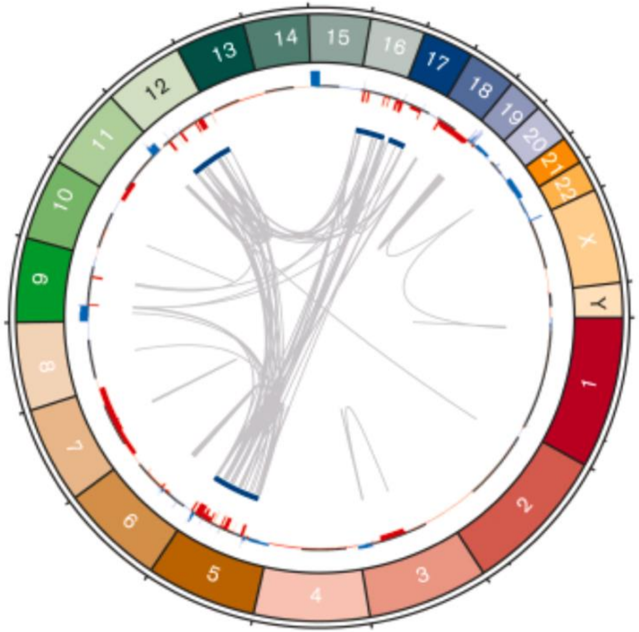
Large Indel detection


No indel detectable - Short-read sequencing - Illumina
83bp deletion - Long-read sequencing - Oxford Nanopore
83bp deletion
chr3: 31990200-31990700
ZNF860
(protein-coding sequence)

Haley Abel

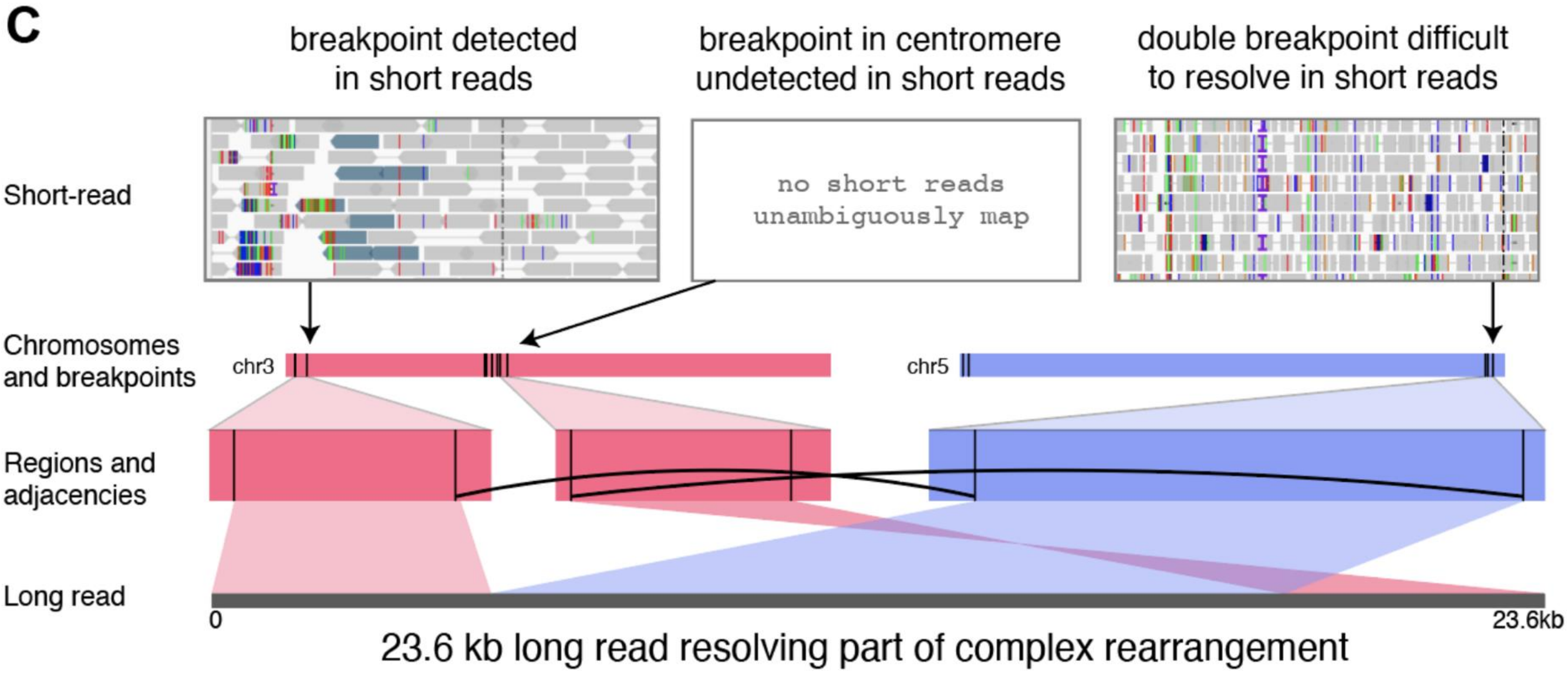
Structural variant resolution

TP53-mutated AML



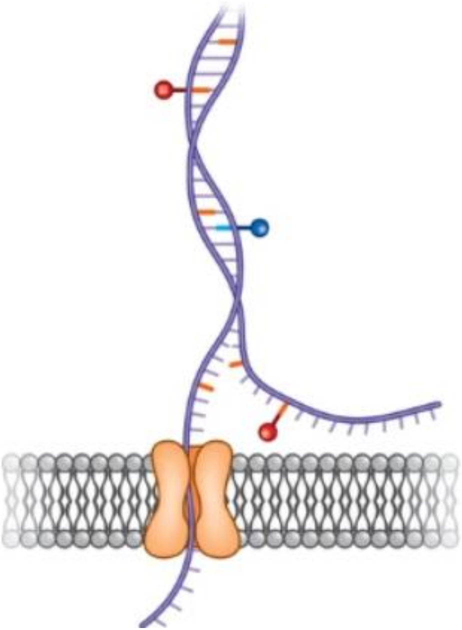
doi: 10.1182/bloodadvances.2023010156

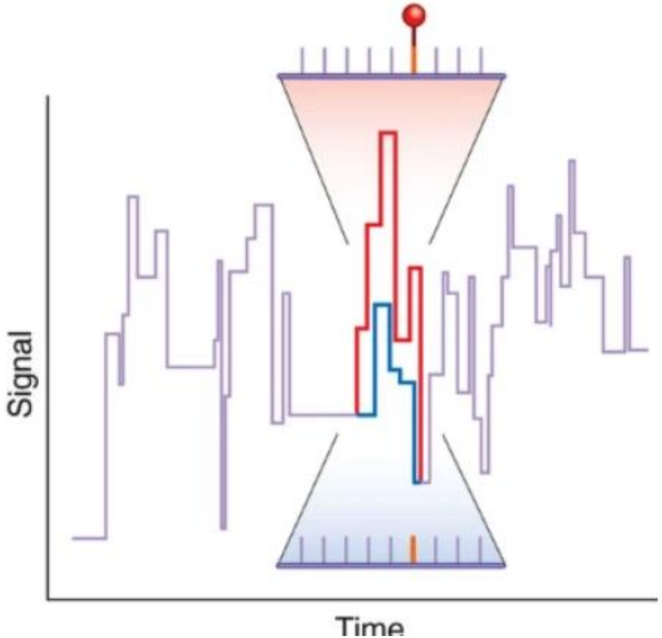
Structural variant resolution


C
breakpoint detected in short reads
breakpoint in centromere undetected in short reads
double breakpoint difficult to resolve in short reads
Short-read
no short reads unambiguously map
Chromosomes and breakpoints
chr3
chr5
Regions and adjacencies
Long read
0
23.6kb
23.6 kb long read resolving part of complex rearrangement

Haley Abel

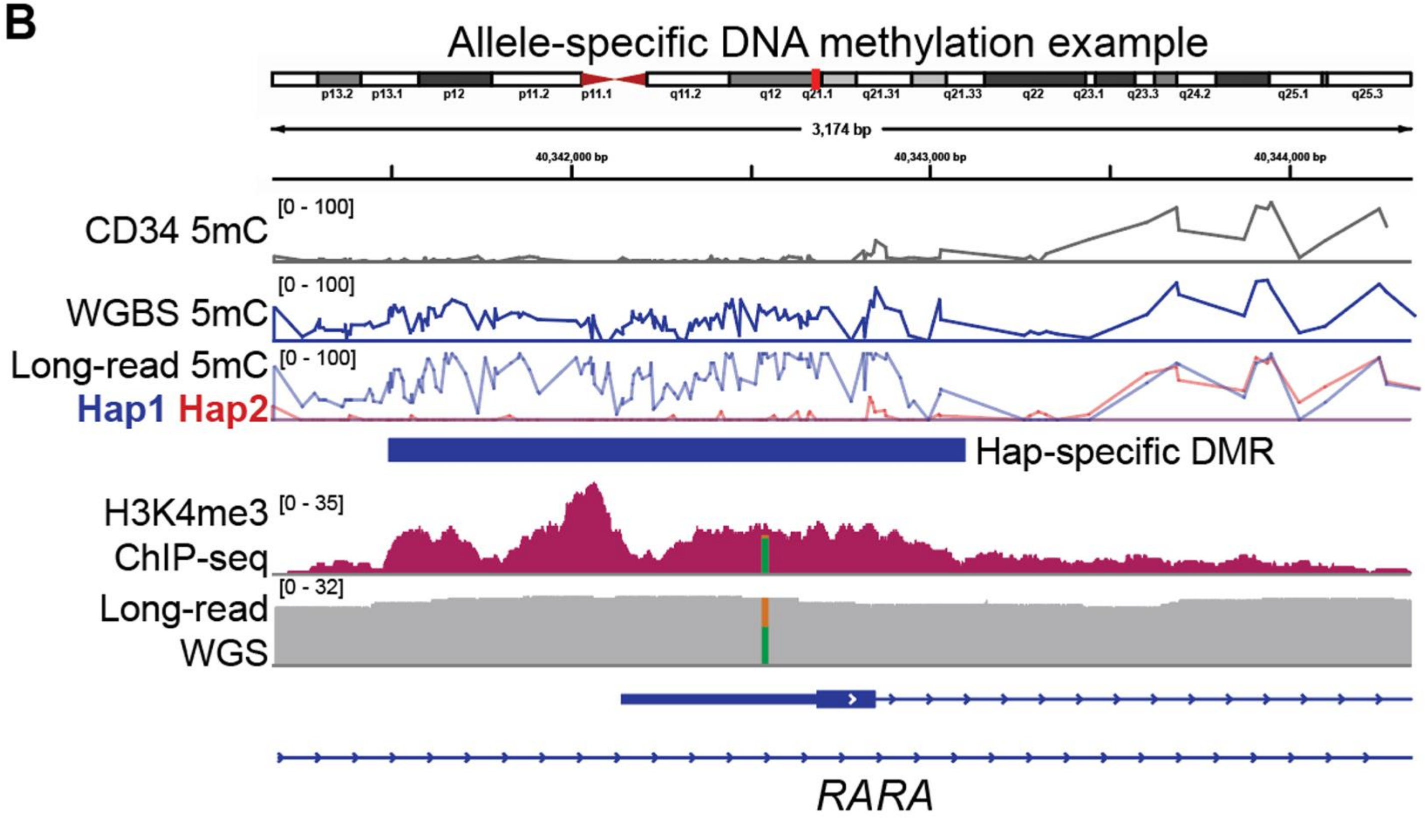
Base modification detection





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

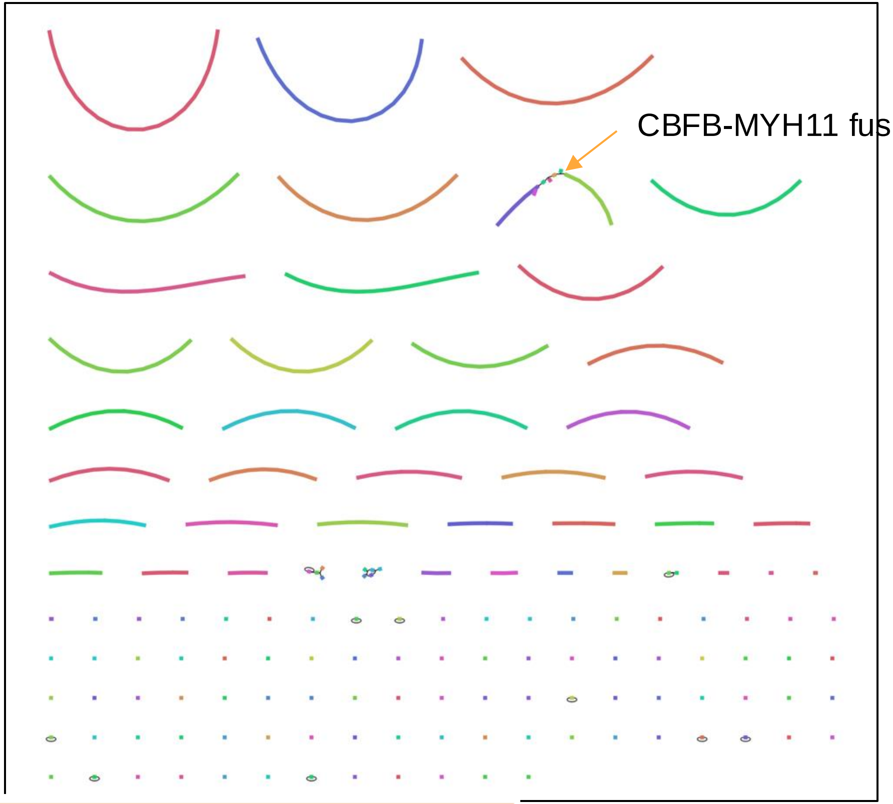
Phasing of reads/modifications



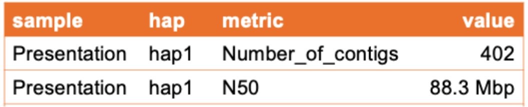
Dave Spencer

Genome assembly

- Assembly of personal genomes



CBFB-MYH11 fusion



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

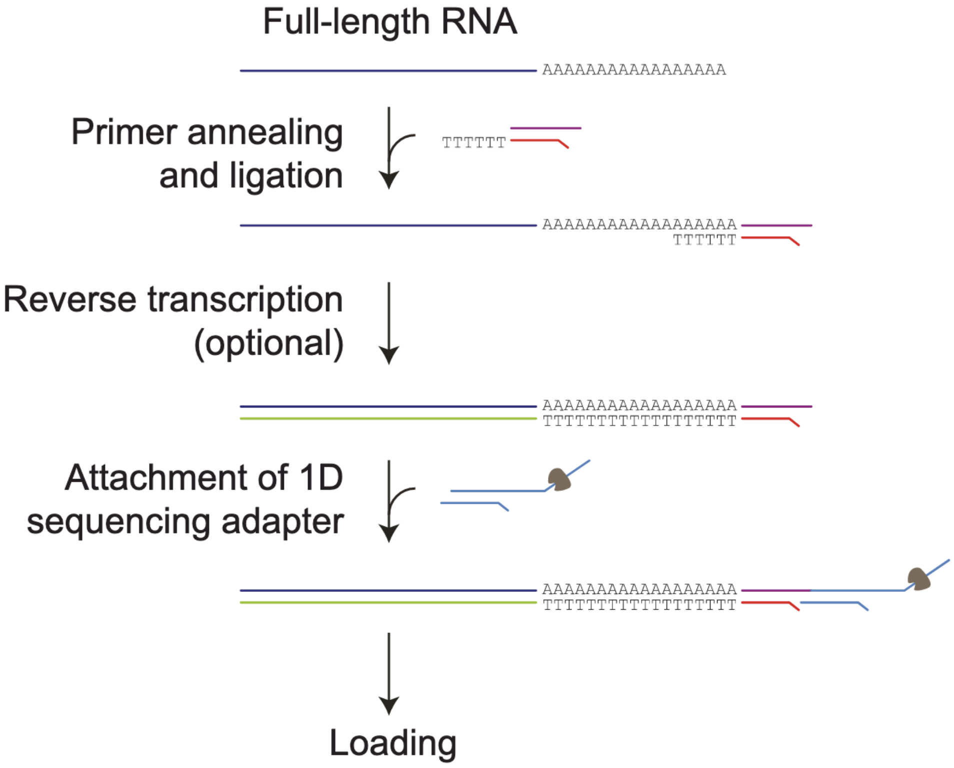
Dave Spencer, Haley Abel

Long-read RNA sequencing

- Direct RNA

- No amplification, less bias

- Preserves base modifications (m6a, etc)



Full-length RNA

Primer annealing and ligation

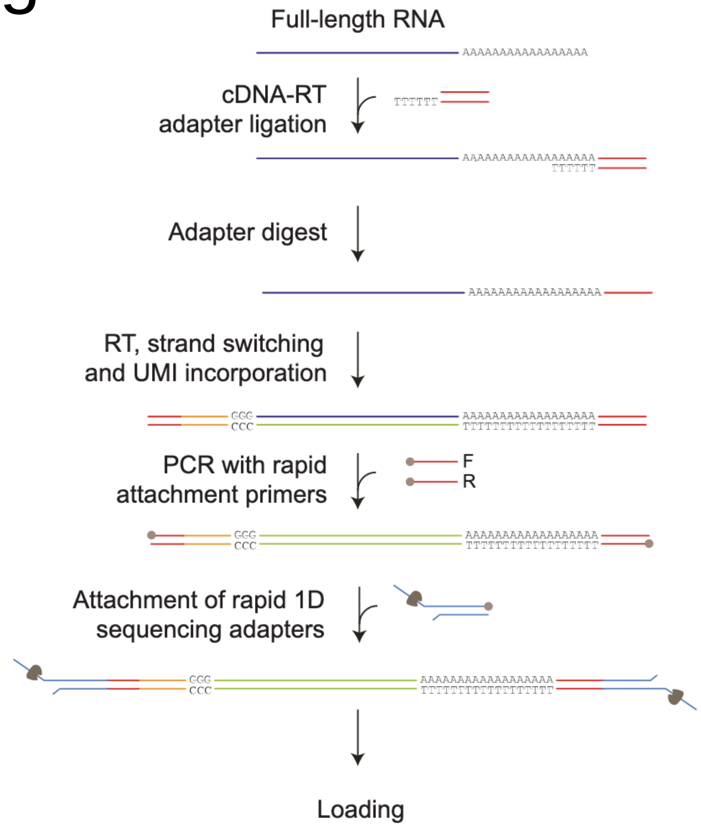
Reverse transcription (optional)

Attachment of 1D sequencing adapter

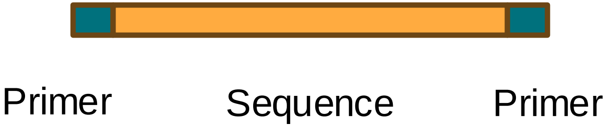
Loading

Long-read RNA sequencing

- cDNA sequencing
- much higher yields

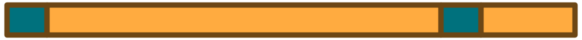


Pychopper

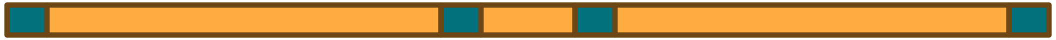


Primer Sequence Primer

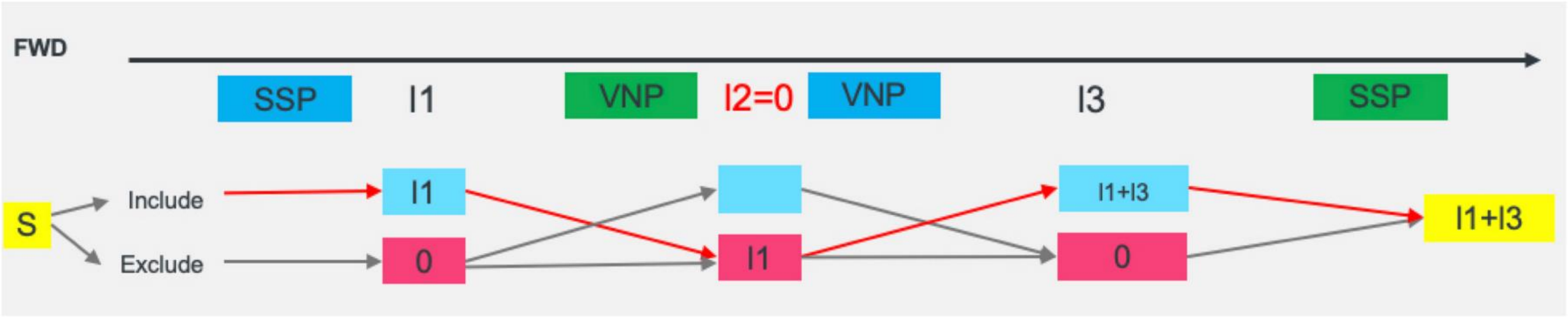
Pychopper



Pychopper



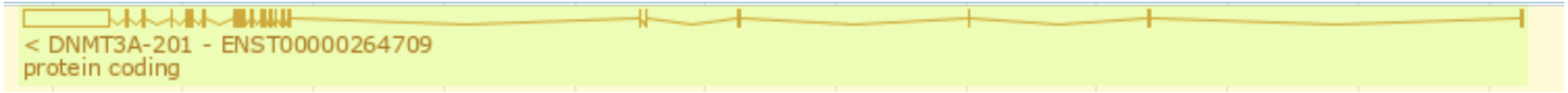
Primer Sequence Primer



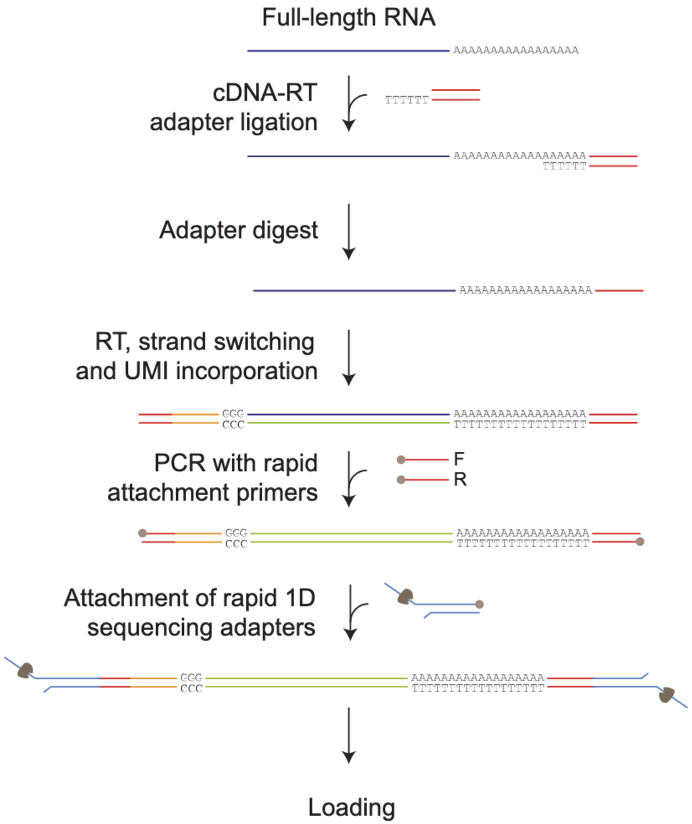
How to estimate duplication rates

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

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

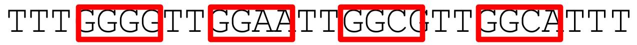


UMIs


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

UMIs

- UMI at the 3' end of the read



TTTGGGGTTGGAATTGGCCTTGGCATTTT

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

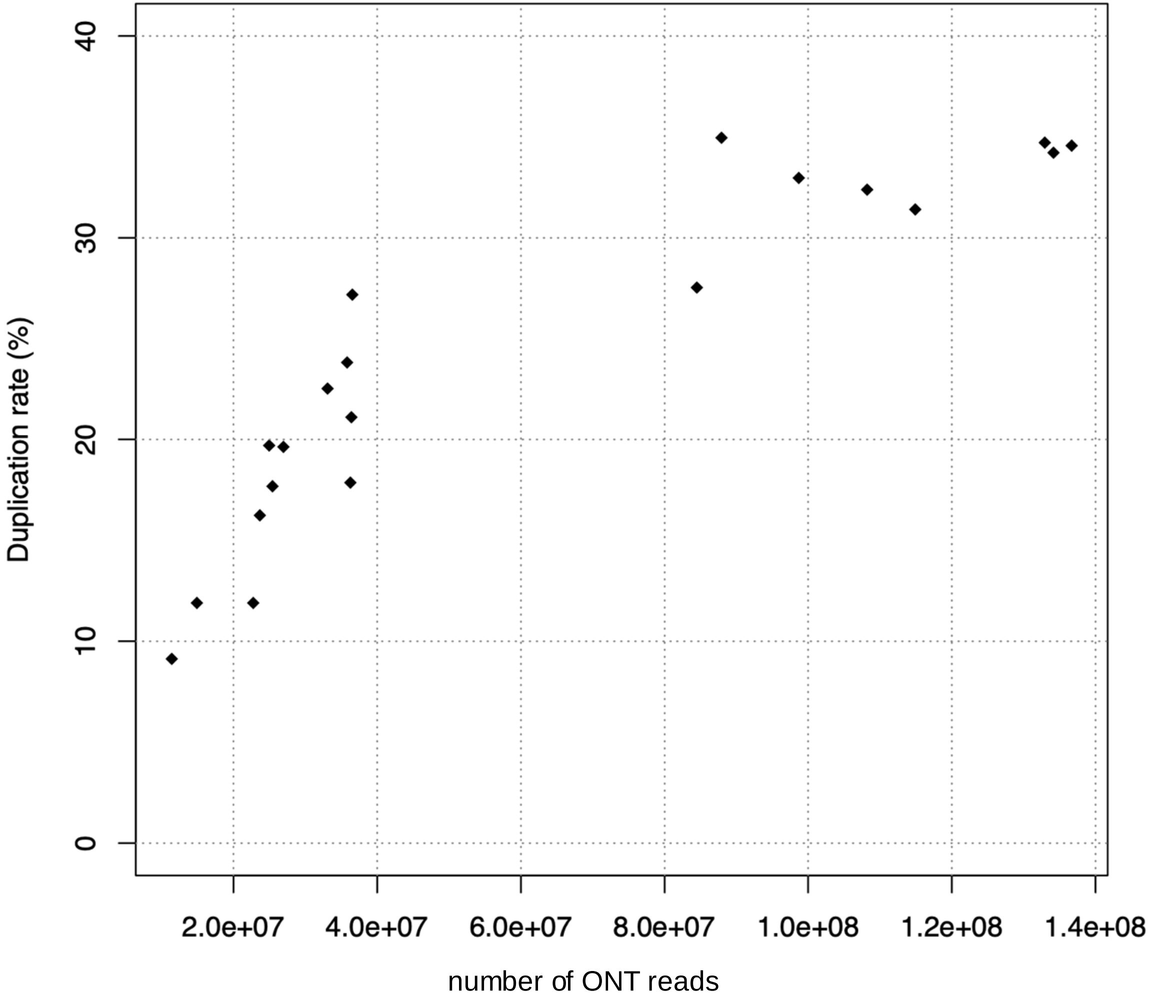
- Different lengths indicative of high error rate

- only 47% of reads have fully intact UMI
- 7% have no UMI at all

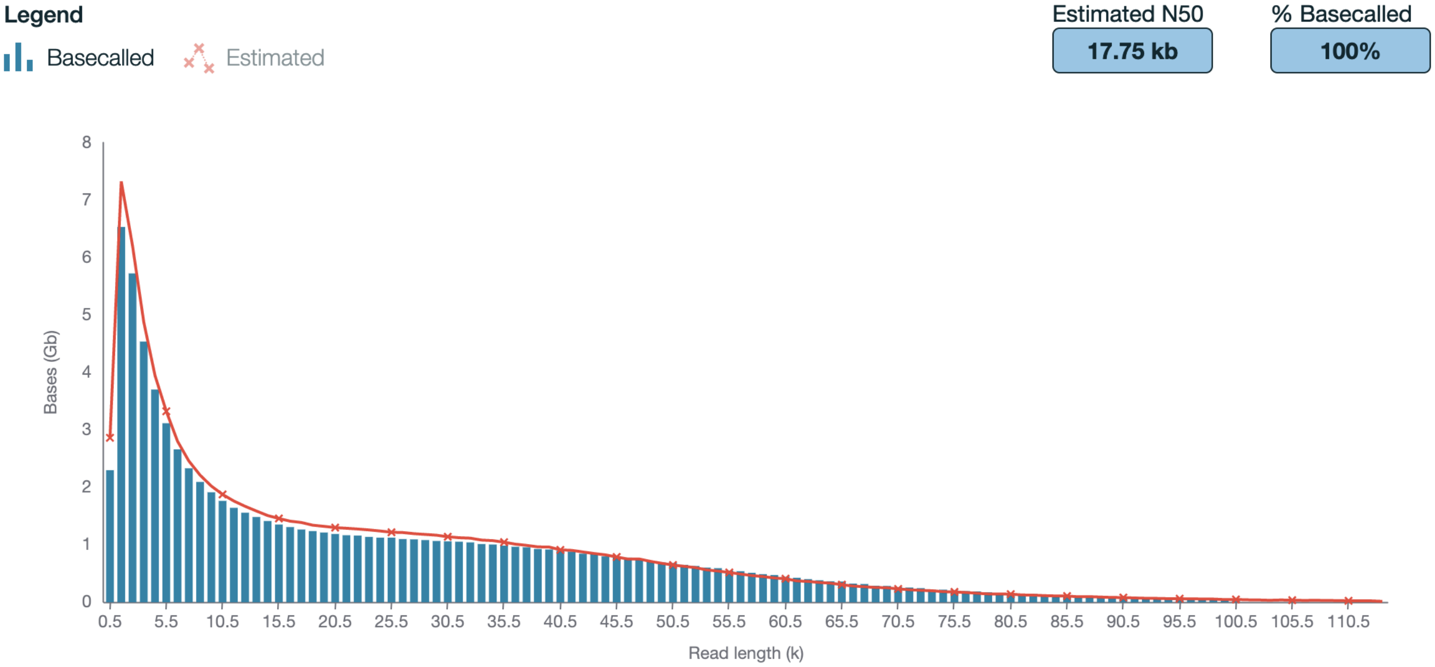
- Even using some error correction with Levenshtein distance, it's ugly

```
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
```

0.98^28 = 0.56


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?

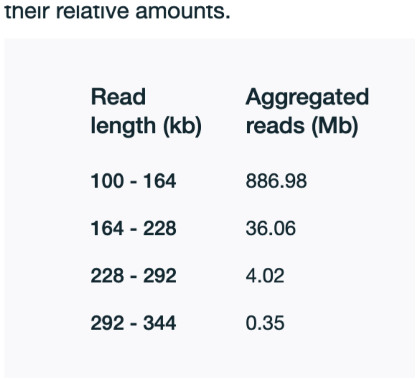


Estimated N50

17.75 kb

% Basecalled

100%

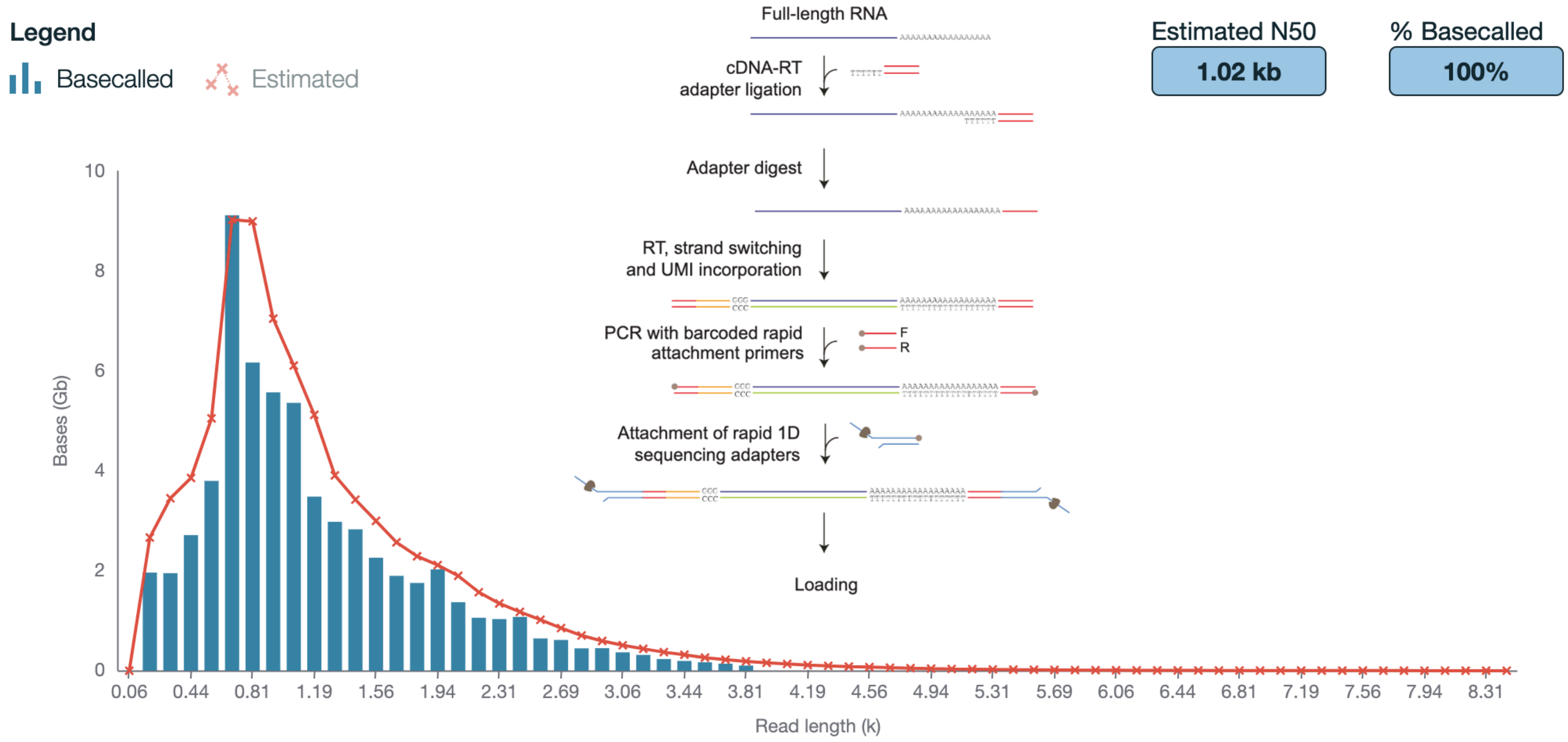


their relative amounts.

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?



Full-length RNA

cDNA-RT adapter ligation

Adapter digest

RT, strand switching and UMI incorporation

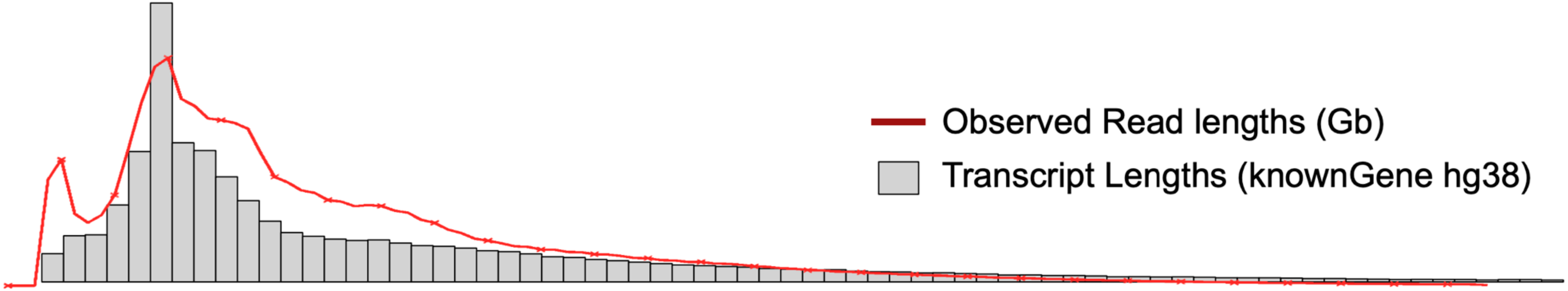
PCR with barcoded rapid attachment primers

Attachment of rapid 1D sequencing adapters

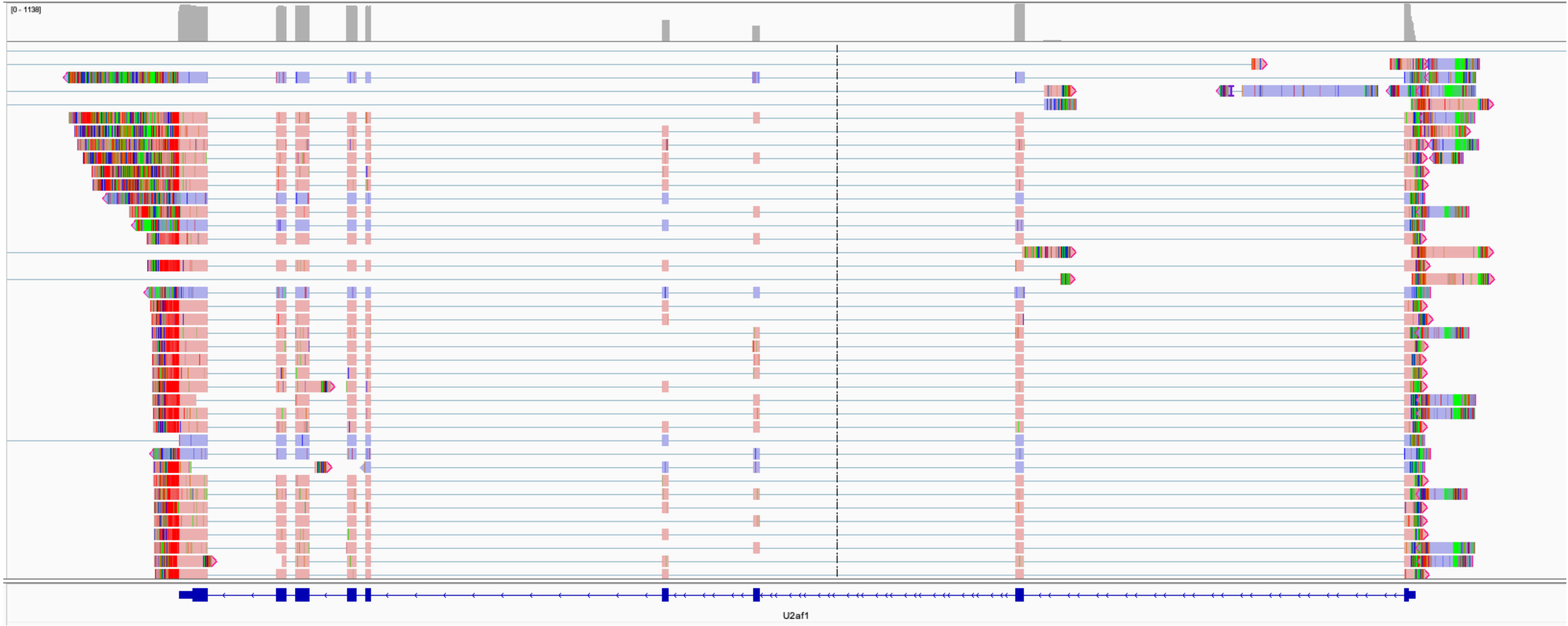
Loading

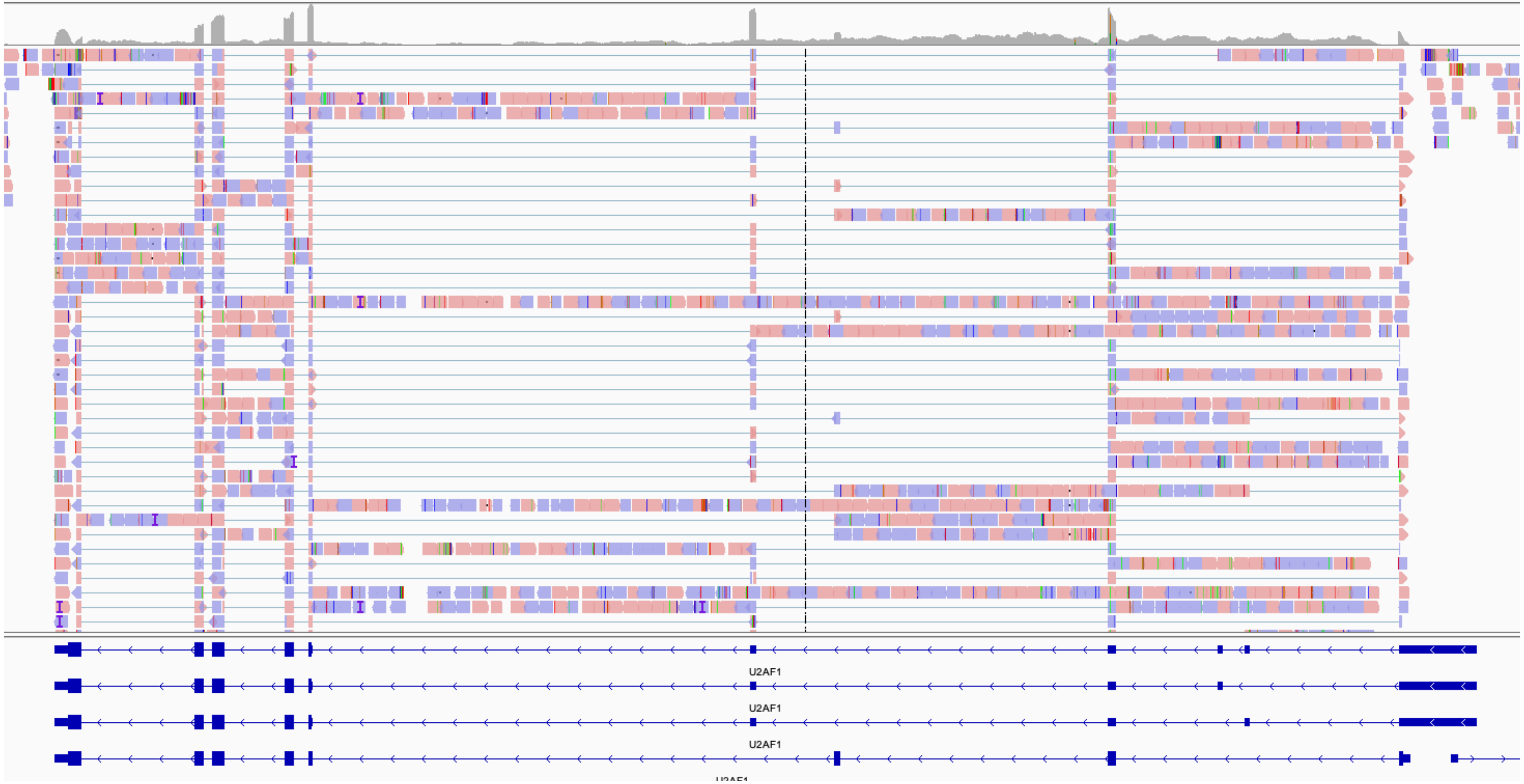
RNA/cDNA – standard prep

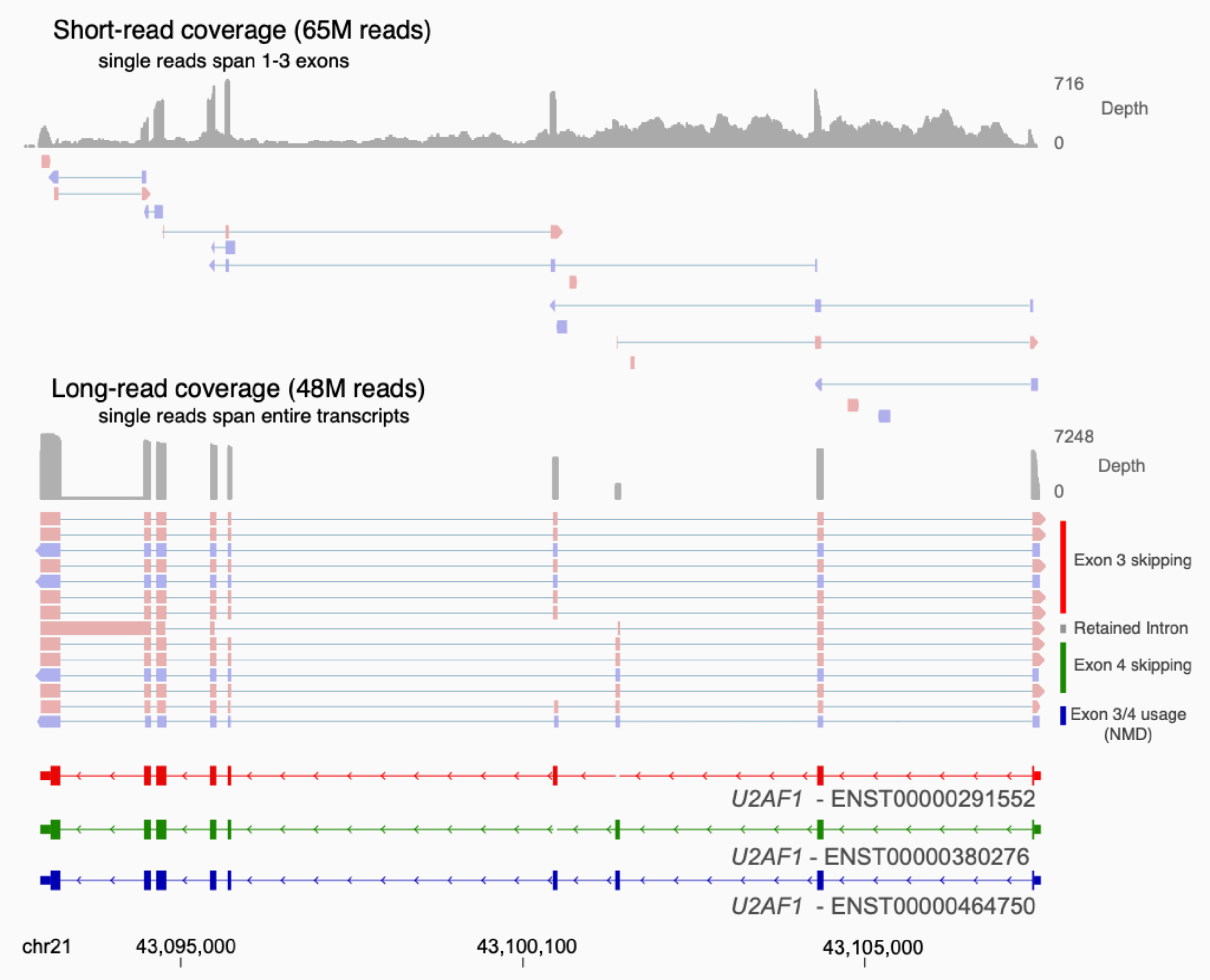
What does the data look like?



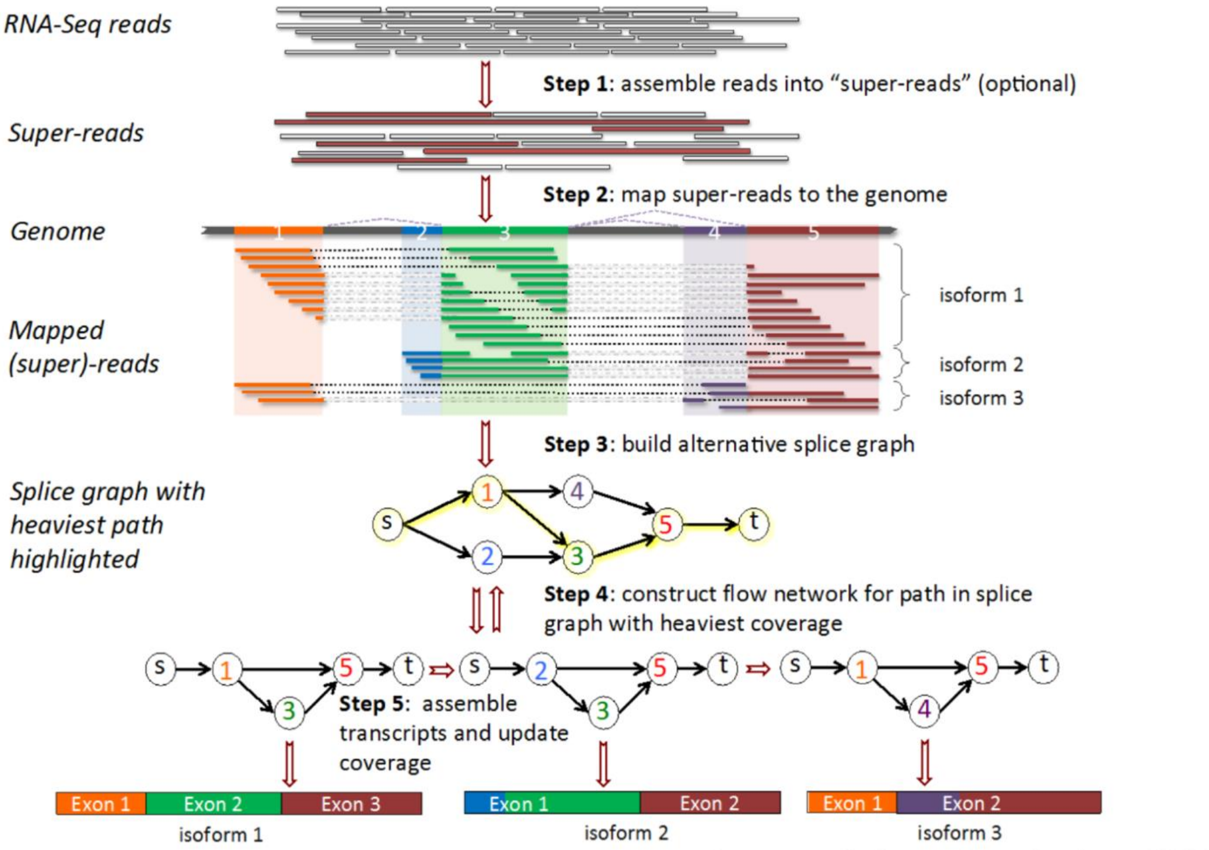
cDNA – standard prep


[0 - 1138]
U2af1


U2AF1
U2AF1
U2AF1
U2AF1

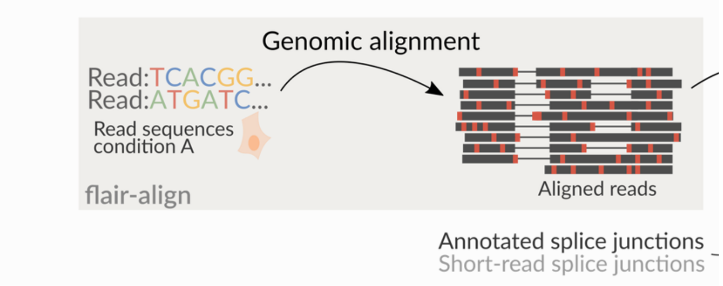

Short-read coverage (65M reads)
single reads span 1-3 exons
716
Depth
0
Long-read coverage (48M reads)
single reads span entire transcripts
7248
Depth
0
Exon 3 skipping
Retained Intron
Exon 4 skipping
Exon 3/4 usage
(NMD)
U2AF1 - ENST00000291552
U2AF1 - ENST00000380276
U2AF1 - ENST00000464750
chr21
43,095,000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

Estimating transcript abundance – short-read, Stringtie


RNA-Seq reads
Step 1: assemble reads into "super-reads" (optional)
Super-reads
Step 2: map super-reads to the genome
Genome
1
2
3
4
5
Mapped (super)-reads
isoform 1
isoform 2
isoform 3
Step 3: build alternative splice graph
Splice graph with heaviest path highlighted
s
1
4
5
t
2
3
Step 4: construct flow network for path in splice graph with heaviest coverage
s
1
5
t
3
s
2
5
t
3
s
1
5
t
4
Step 5: assemble transcripts and update coverage
Exon 1
Exon 2
Exon 3
isoform 1
Exon 1
Exon 2
isoform 2
Exon 1
Exon 2
isoform 3

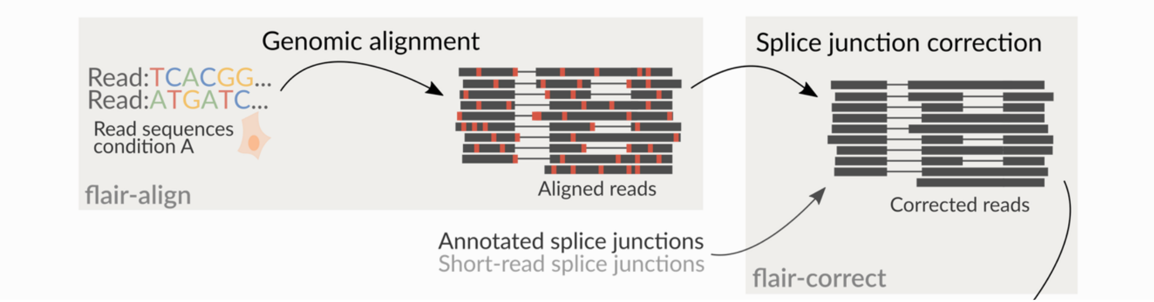
Pertea et al. Nature Biotechnology, 2015

Estimating transcript abundance – long read


Genomic alignment
Read:TCACGG...
Read:ATGATC...
Read sequences condition A
flair-align
Aligned reads
Annotated splice junctions
Short-read splice junctions

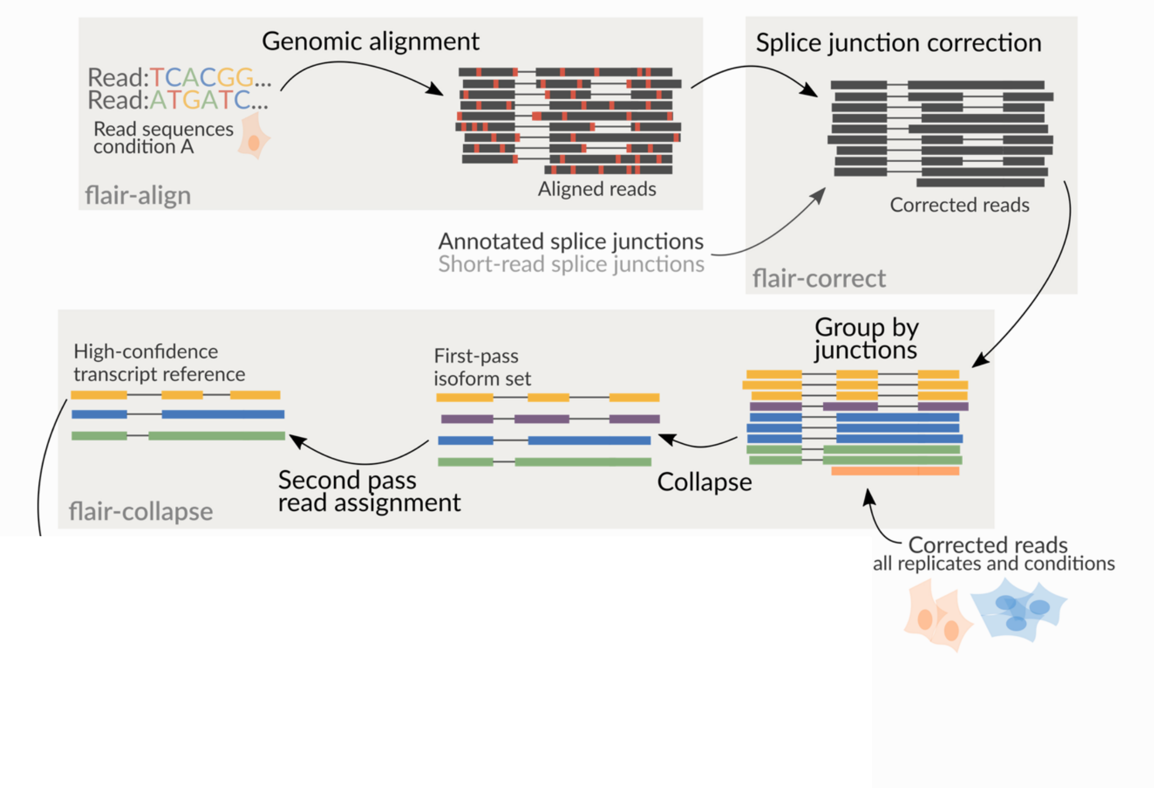
https://github.com/BrooksLabUCSC/flair

Estimating transcript abundance – long read


Genomic alignment
Read:TCACGG...
Read:ATGATC...
Read sequences condition A
flair-align
Aligned reads
Splice junction correction
Corrected reads
Annotated splice junctions
Short-read splice junctions
flair-correct

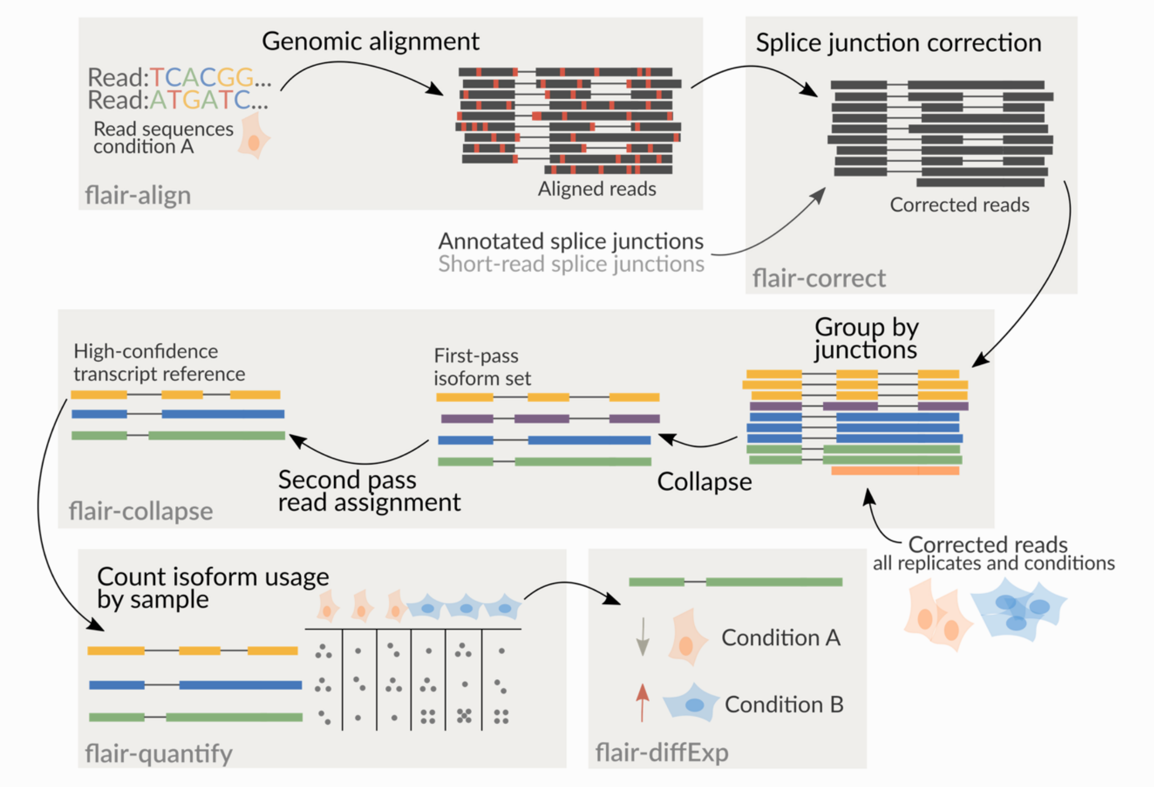
https://github.com/BrooksLabUCSC/flair

Estimating transcript abundance – long read



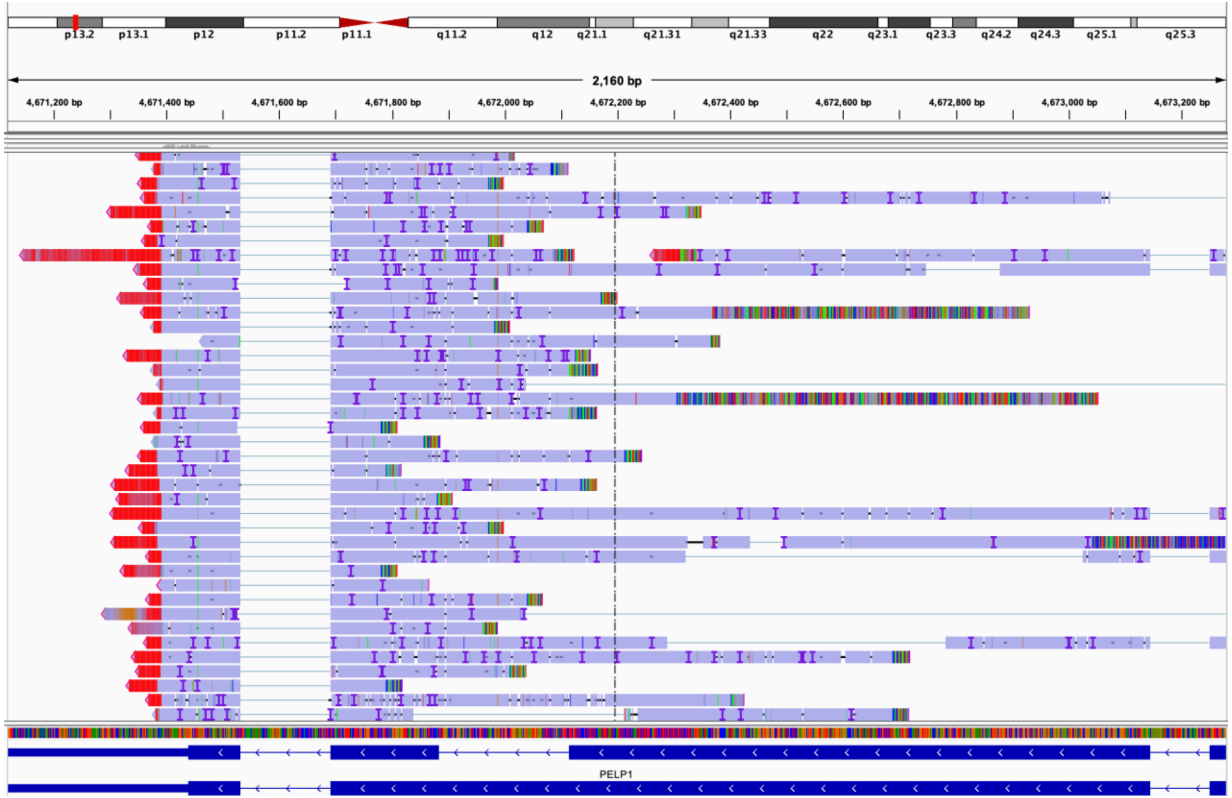
https://github.com/BrooksLabUCSC/flair

Estimating transcript abundance – long read


Genomic alignment
Read:TCACGG...
Read:ATGATC...
Read sequences condition A
Aligned reads
flair-align
Splice junction correction
Corrected reads
Annotated splice junctions
Short-read splice junctions
flair-correct
Group by junctions
High-confidence transcript reference
First-pass isoform set
Second pass read assignment
Collapse
flair-collapse
Corrected reads
all replicates and conditions
Count isoform usage by sample
flair-quantify
Condition A
Condition B
flair-diffExp

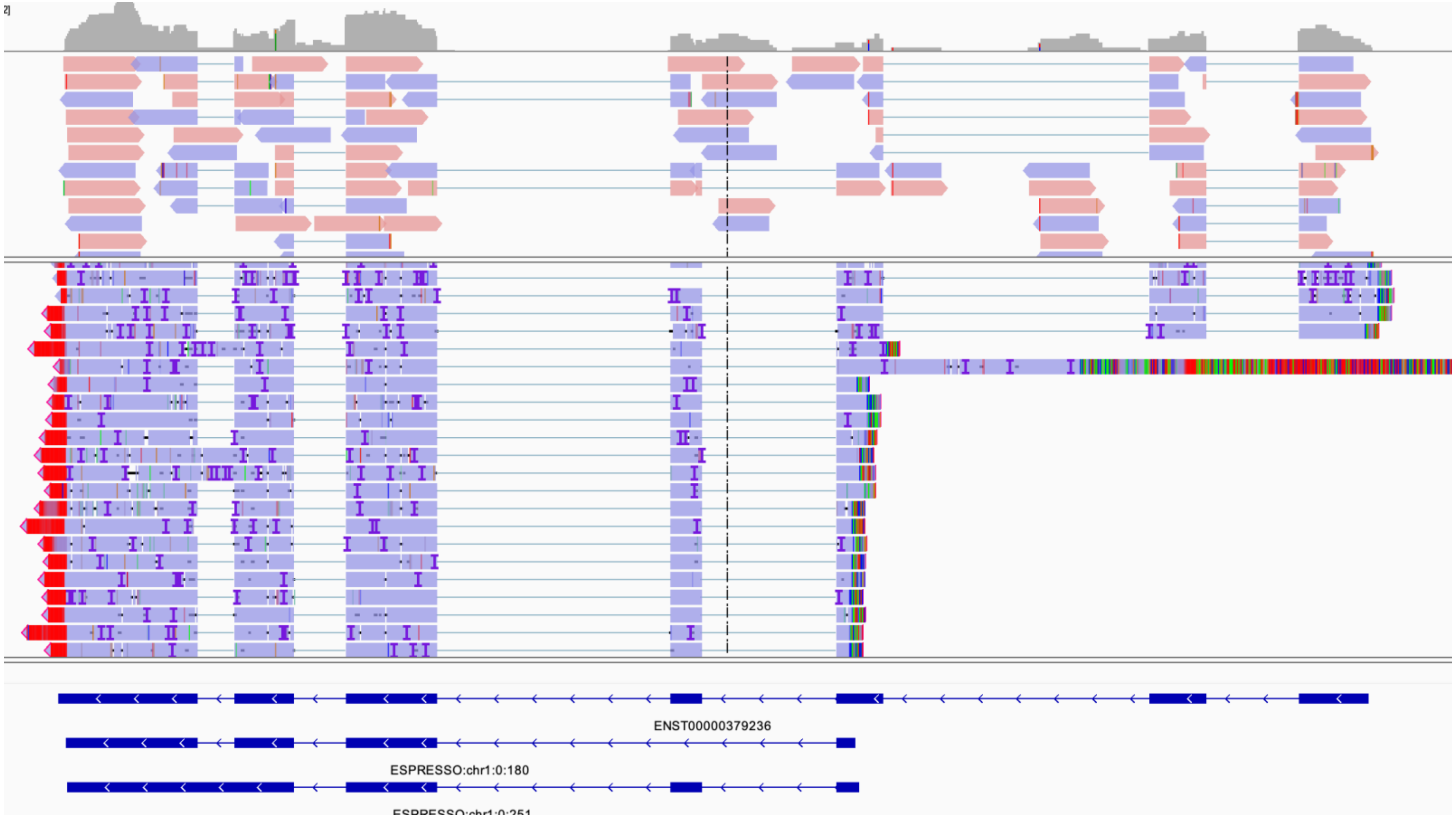
https://github.com/BrooksLabUCSC/flair

Technical artifacts



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

"Full-length" reads



Short-read

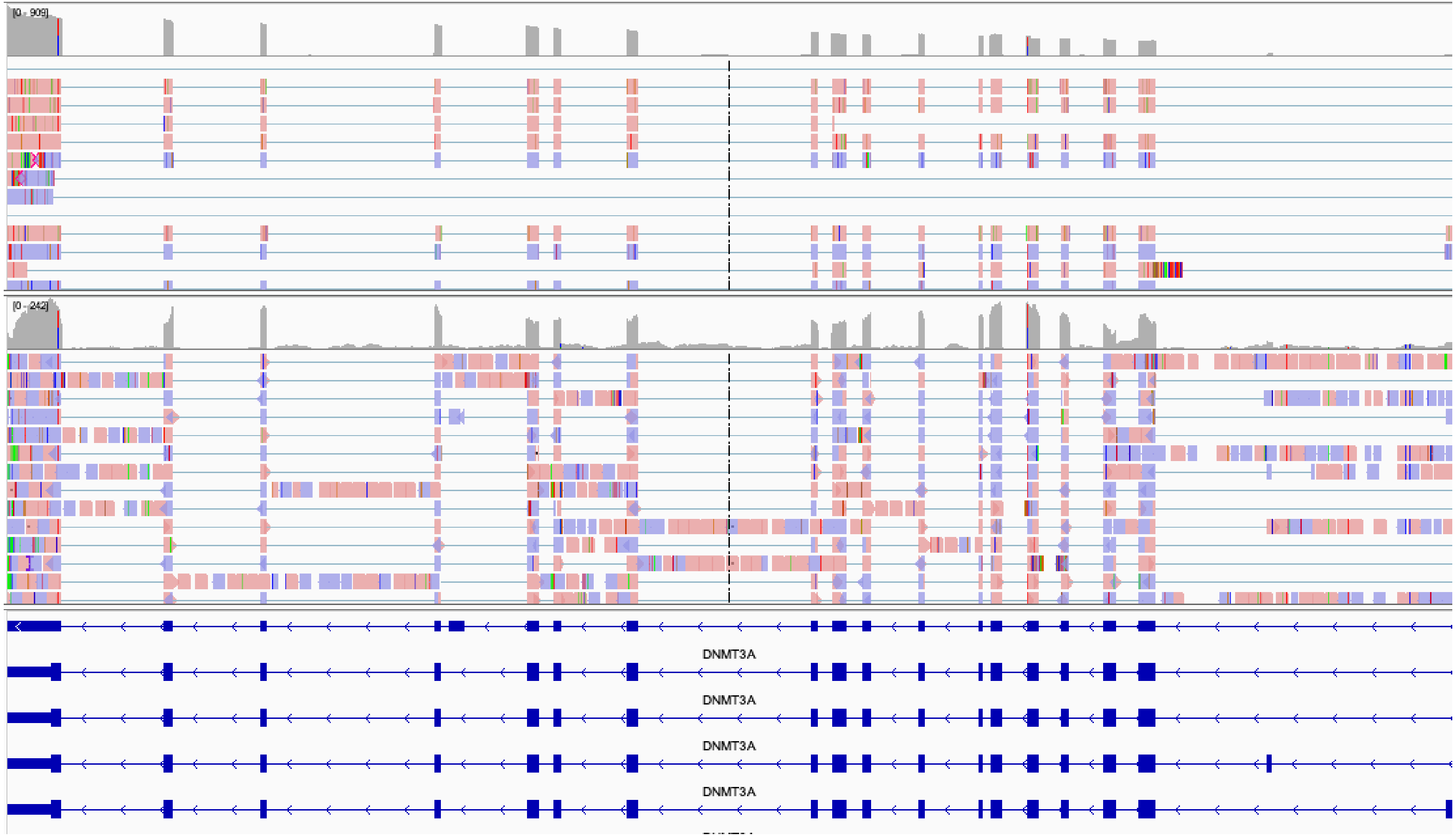
Long-read

ESPRESSO transcript assembly

human DNMT3A

ONT

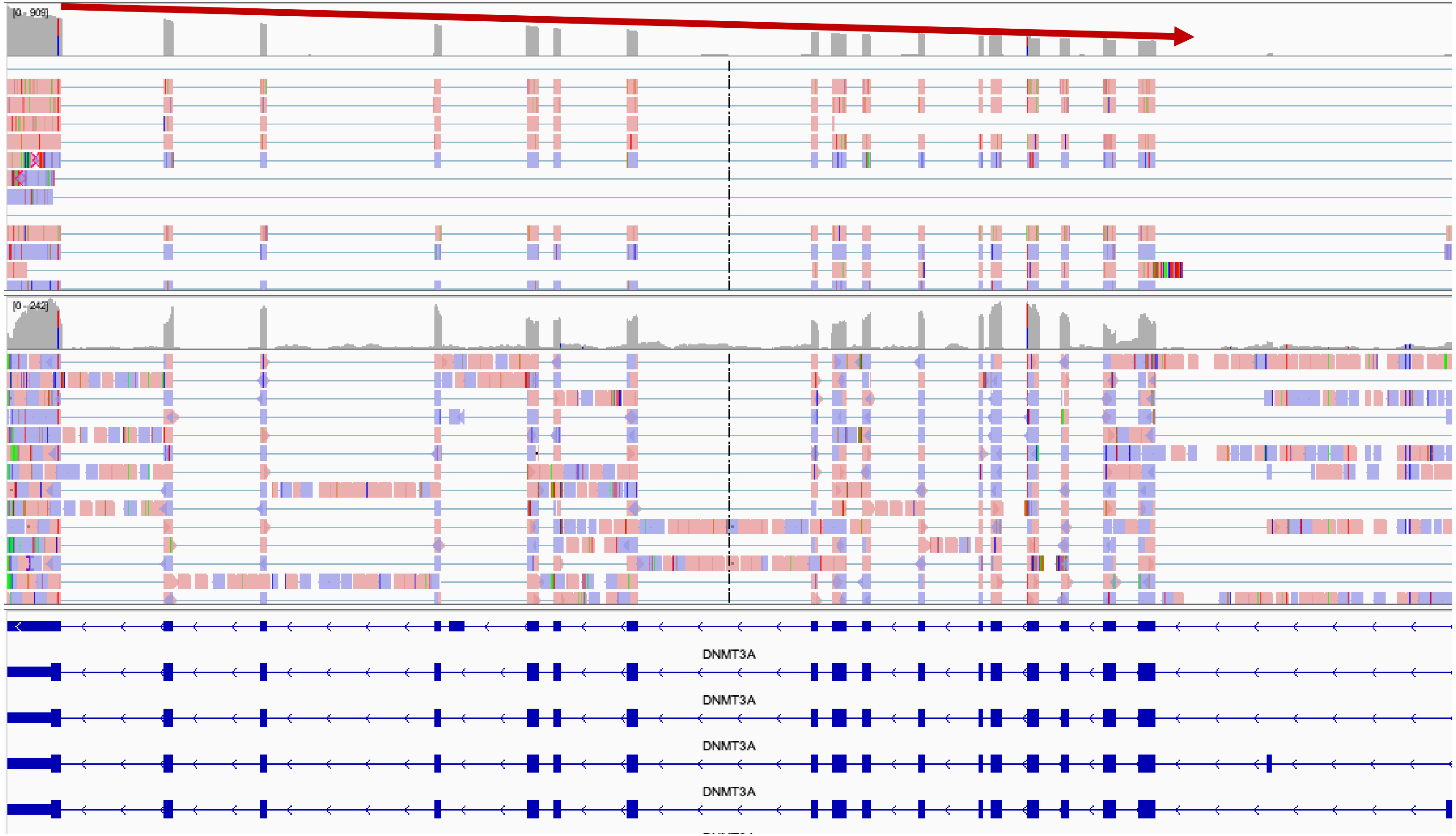
Short-read



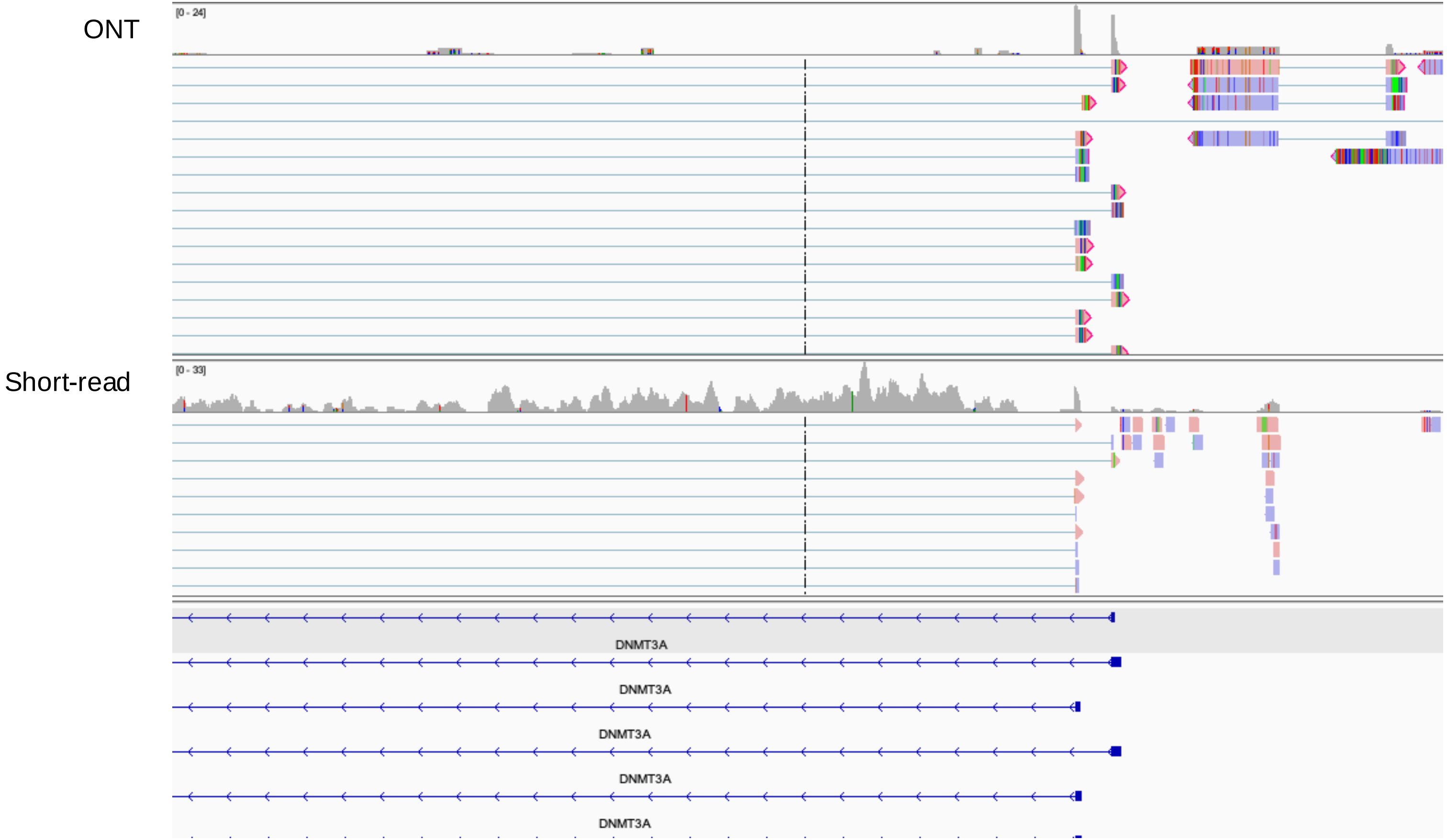
human DNMT3A

ONT

Short-read



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

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