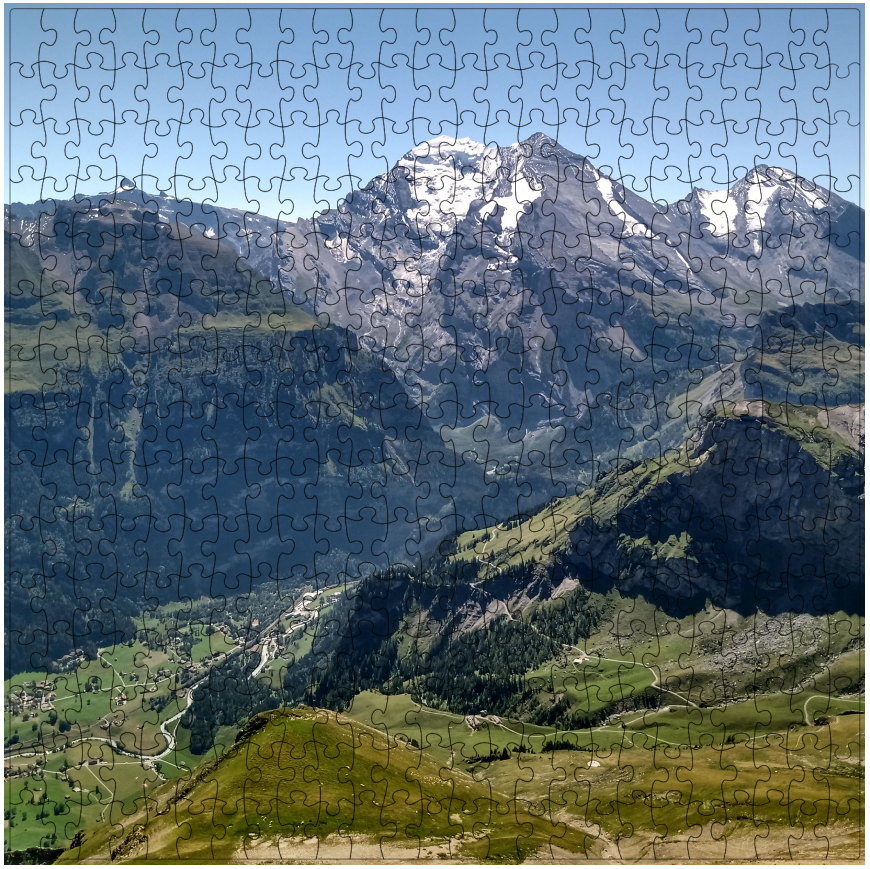
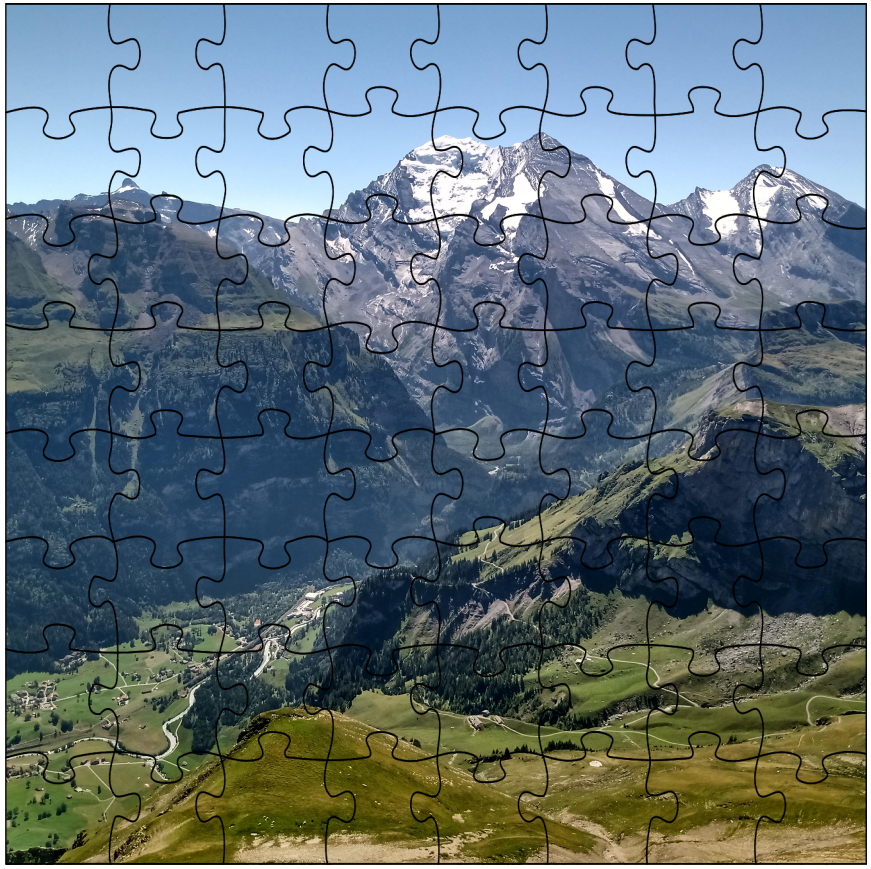
Long-read RNA-seq

What is a long read?

- Short read: 50-300 bp, often paired-end
- Long read: > 1kb, up to 20 Mb

Why long reads?

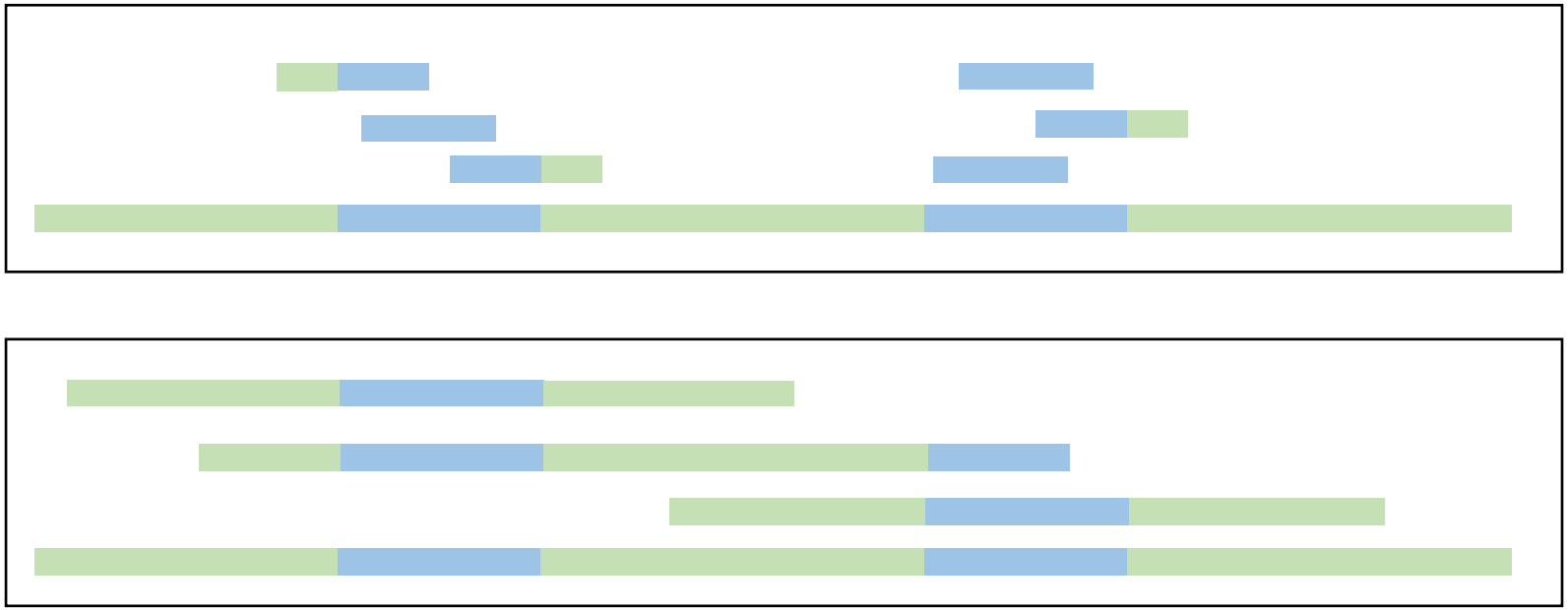




Why long reads?

- Many 2nd generation data analyses have an ‘assembly’ or ‘uniqueness’ question:
  - Mapping quality?
  - Isoform variation?
  - Structural variation?
  - Haplotypes?
  - Genome/transcriptome sequence?

Mapping quality

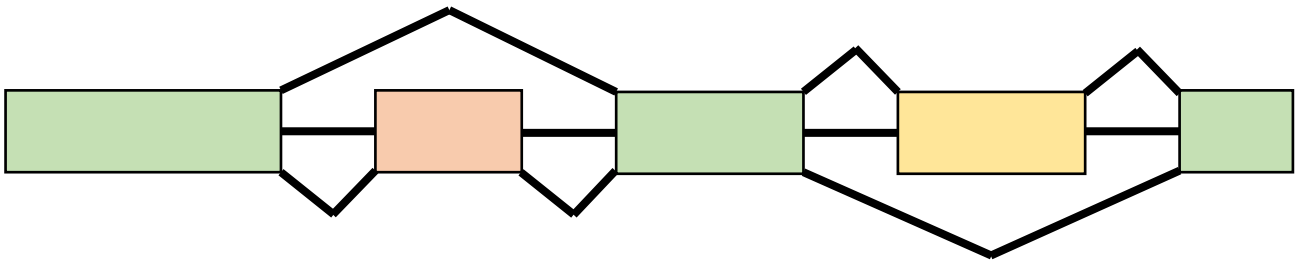


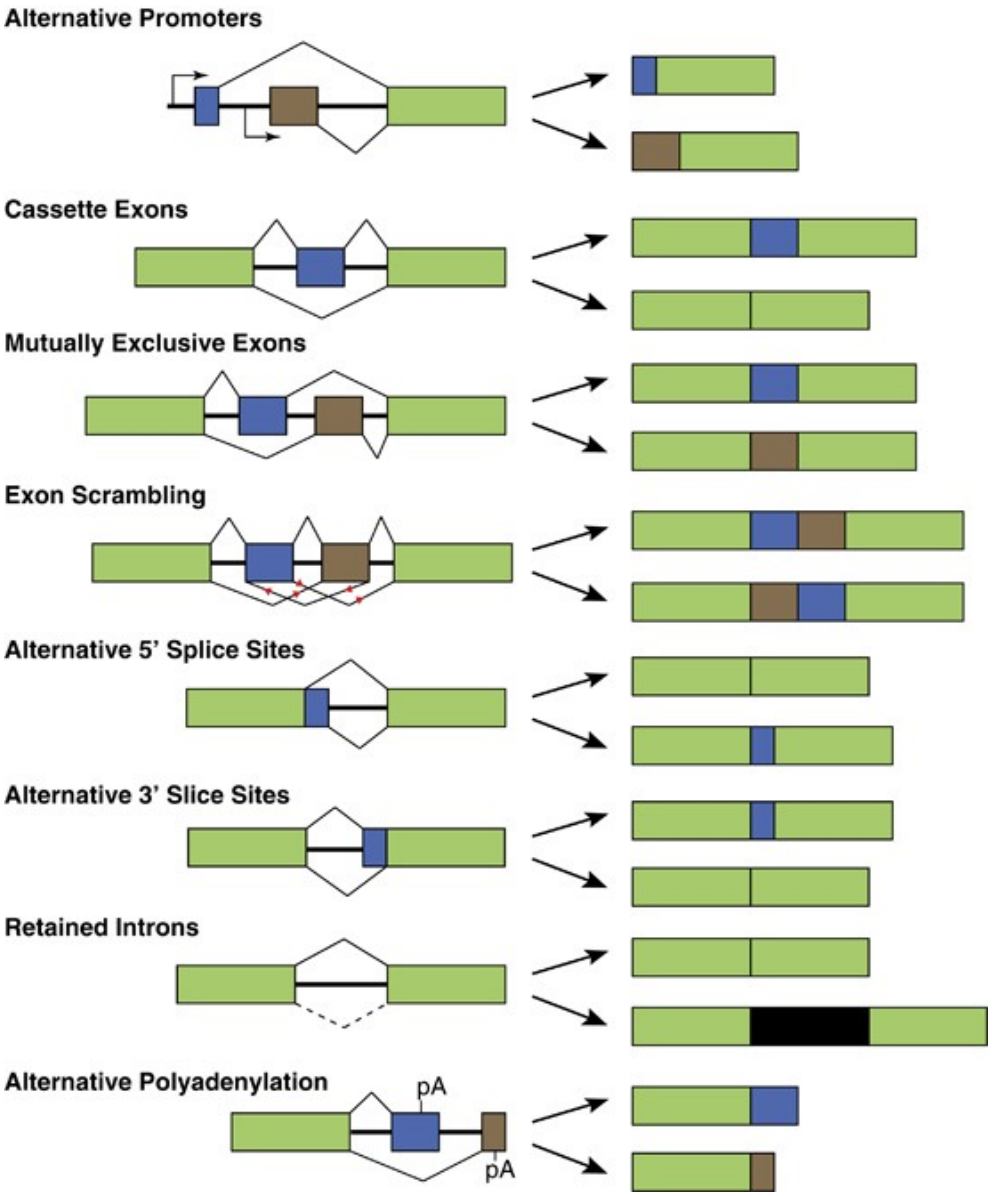
$$MAPQ = -10log_{10}\Pr\{mapping\ position\ is\ wrong\}$$

$$-10log_{10}\Pr\{0.01\} = 20$$
$$-10log_{10}\Pr\{0.5\} = 3$$

Isoform variation

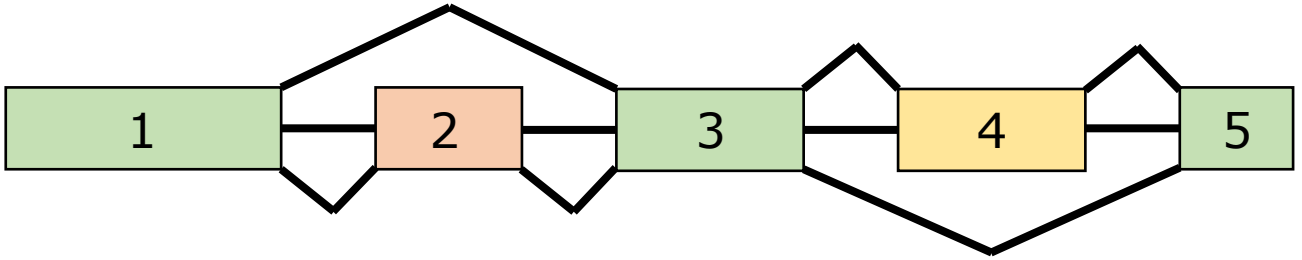
- > 95% of human genes are alternatively spliced
- Some with large phenotypic effects
- Isoforms can have different:
  - Amino acid sequence
  - Promoters
  - Polyadynelation

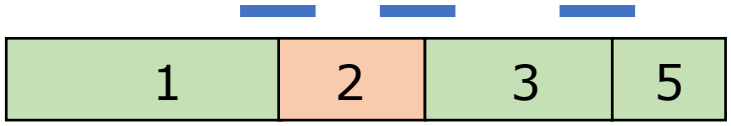


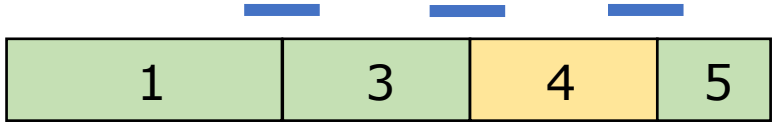


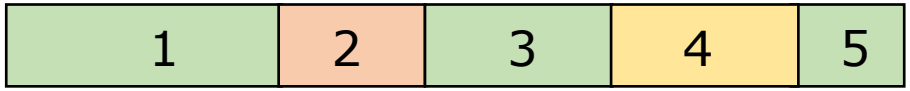
Chen, J., & Weiss, W. A. (2015). *Oncogene*, *34*(1), 1–14.

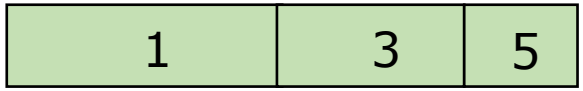
Isoform variation

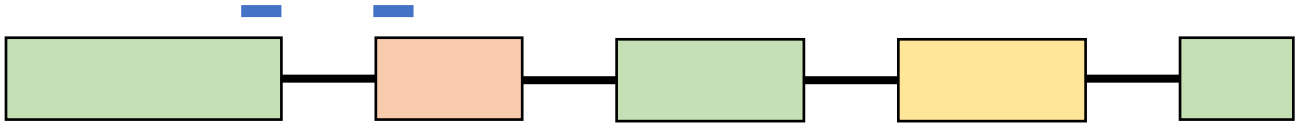






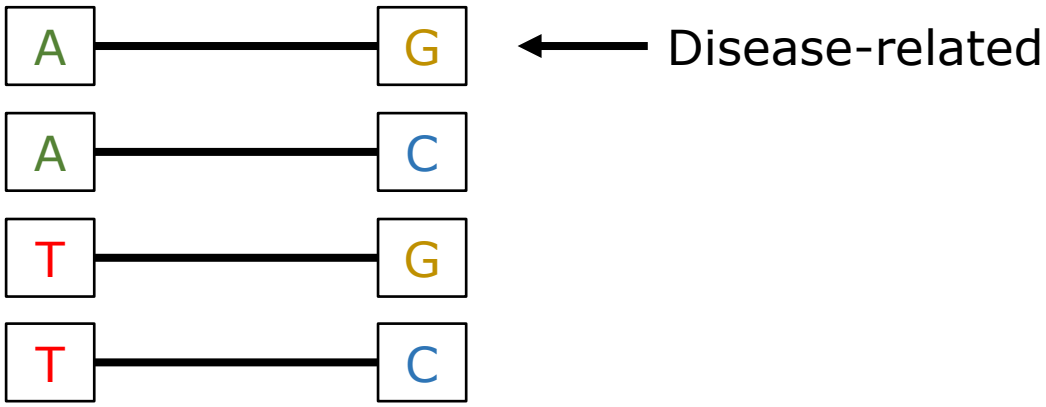




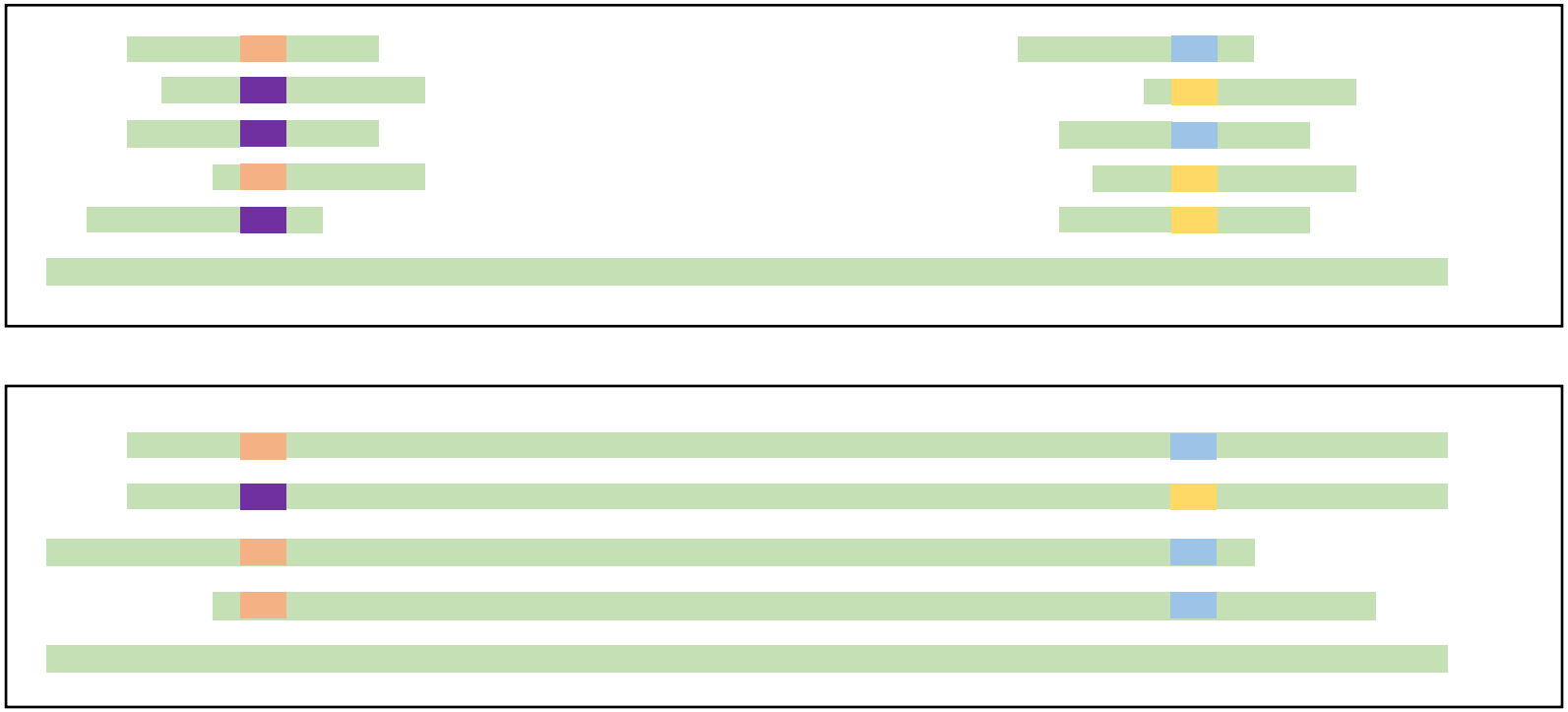


Haplotypes

- Genetic association studies often based on SNPs
- Most SNPs are bi-allelic e.g. [A/T], [G/C]
- Genetic variation is often multi-allelic



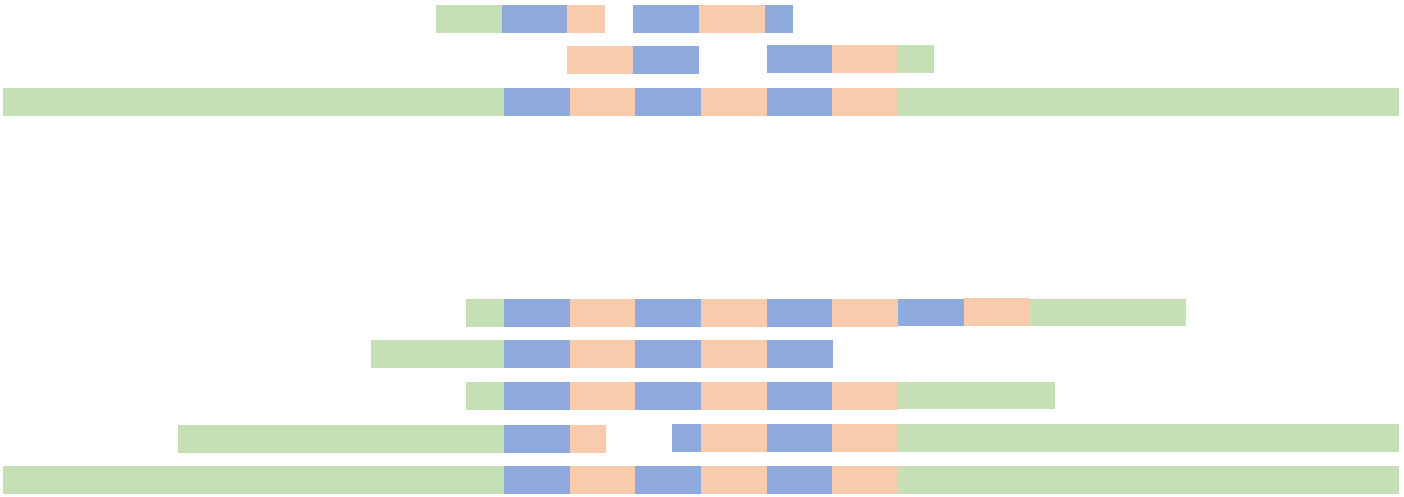
Haplotypes



Structural variation

- ~1kb – 3 Mb
- Associated with important traits
- Insertions/deletions
- Copy-number variation
- Translocations
- Inversions

Copy number variation

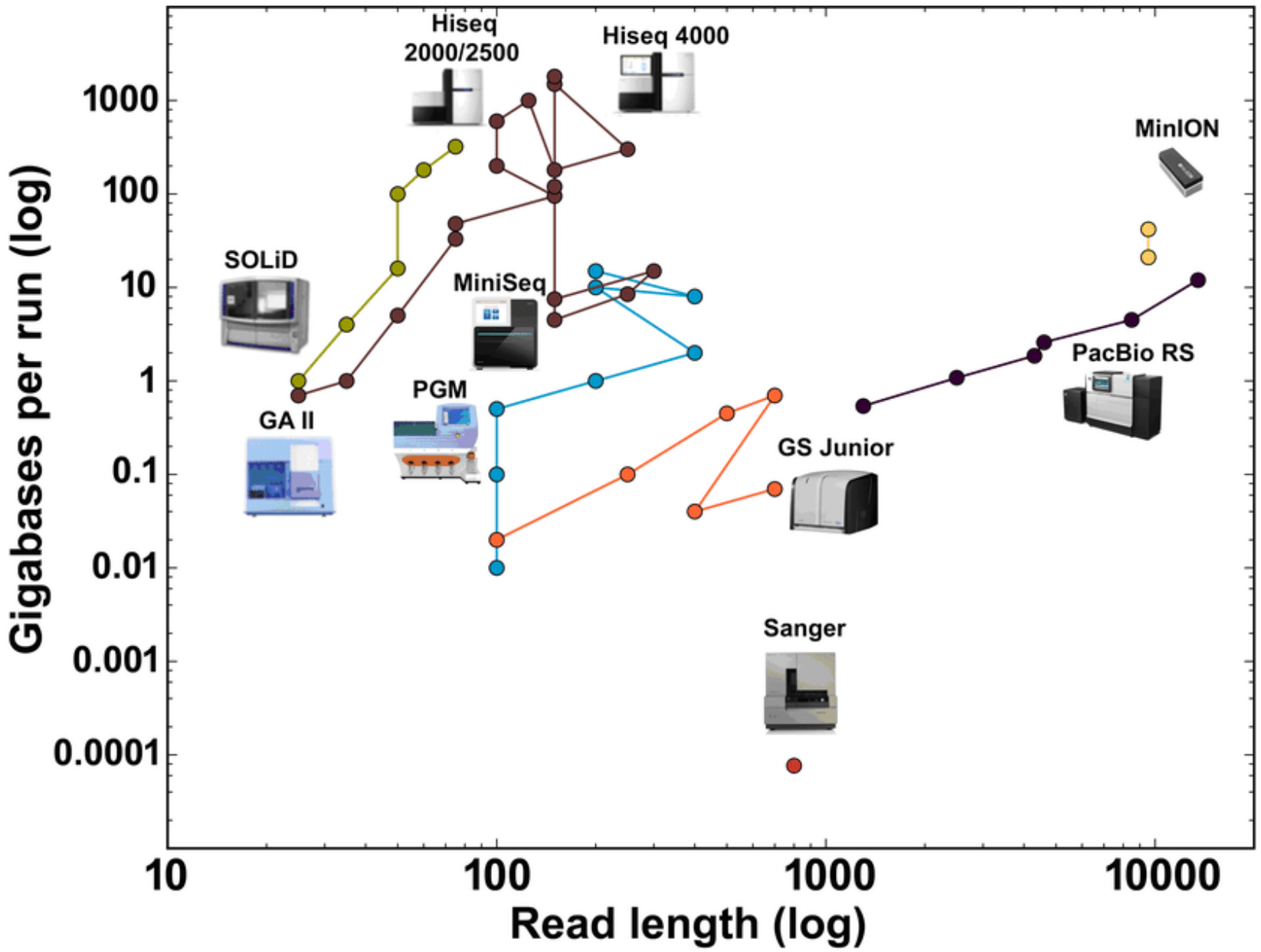


Why long-read RNA-seq?

- Genome annotation
- Transcriptome assembly
- Transcript & isoform discovery
- Differential isoform expression
- Poly(A)tail length (dRNAseq)
- RNA base modifications (dRNAseq)
- Transcript haplotypes
- Any other?

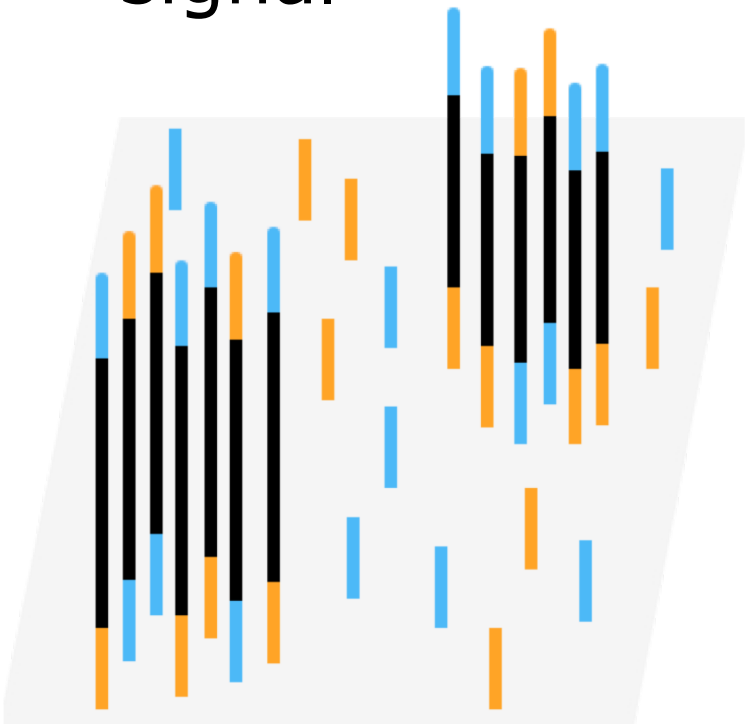
Sequencing (c)DNA

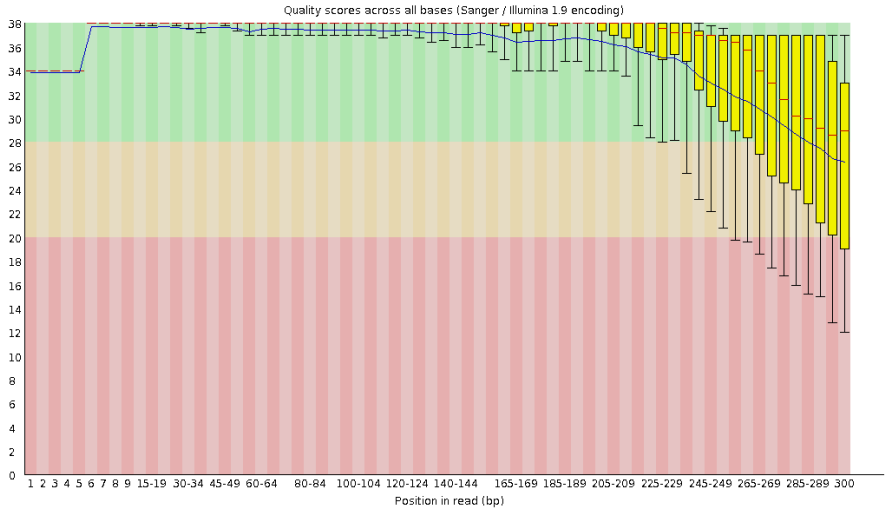
1st generation	2nd generation	3rd generation
Sanger	Illumina	PacBio/ONT
300-1000 bp	2x 50-300 bp	10->20 kb
High accuracy	High accuracy	Medium-High accuracy


Hiseq
2000/2500
Hiseq 4000
MinION
SOLiD
MiniSeq
PacBio RS
PGM
GA II
GS Junior
Sanger
1000
100
10
1
0.1
0.01
0.001
0.0001
Gigabases per run (log)
10
100
1000
10000
Read length (log)

Second generation

- Bridge amplification
- Lengths are limited by out-of-phase of signal



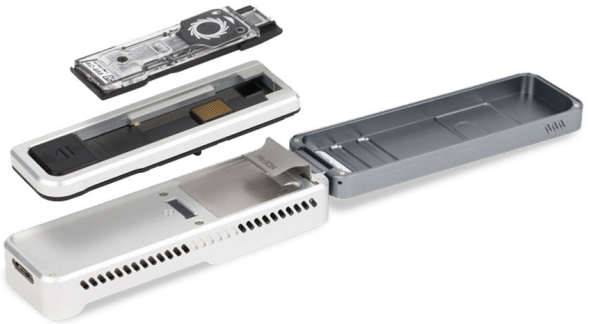


Third generation

- Other than 2nd generation:
  - single molecule -> No bridge amplification
- No out-of-phase of signal
- Virtually no limit in read length
- Most used platforms:
  - PacBio SMRT sequencing
  - Oxford nanopore technology

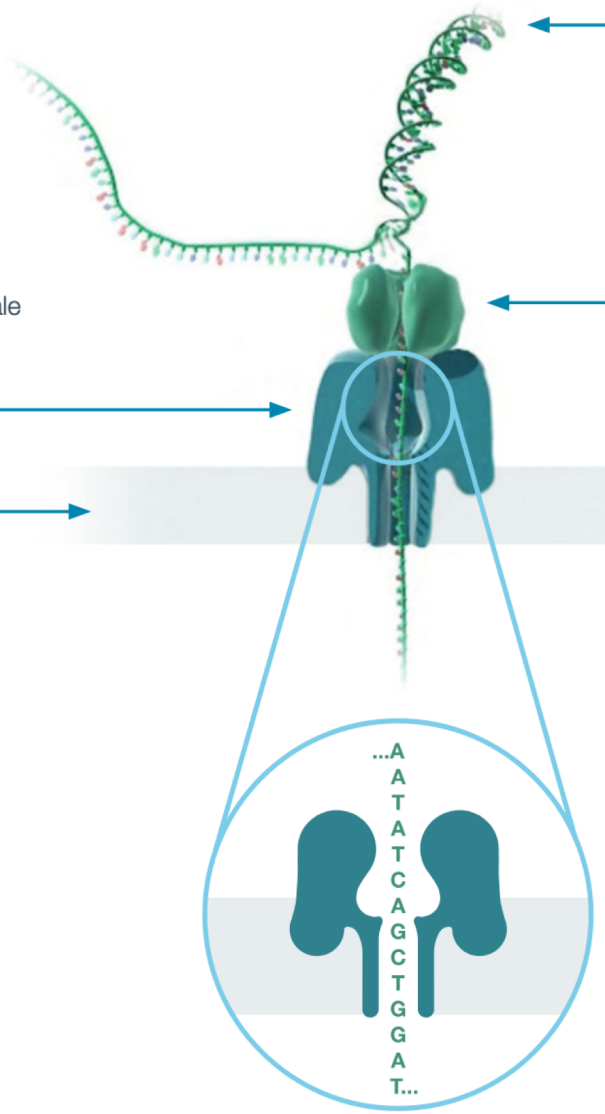
Oxford Nanopore technology

- Based on changes in electrical current
- Well-known for its scalability and portability
- ~95-97% accuracy



Nanopore reader
DNA or RNA passes through a nano-scale hole. The fluctuations in current during translocation are used to determine the DNA or RNA sequence.

An electrically resistant **membrane** means all current must pass through the nanopore, ensuring a clean signal.

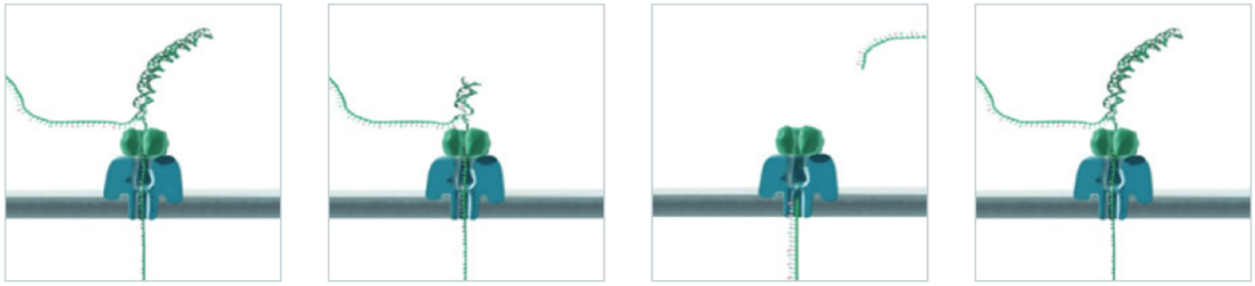


The nanopore processes the length of **DNA or RNA** presented to it. The user can control this through the library preparation protocol utilised. (e.g. >2 Mb DNA has been recorded.)

An **enzyme motor** controls the translocation of the DNA or RNA strand through the nanopore. Once the DNA or RNA has passed through, the motor protein detaches and the nanopore is ready to accept the next fragment.

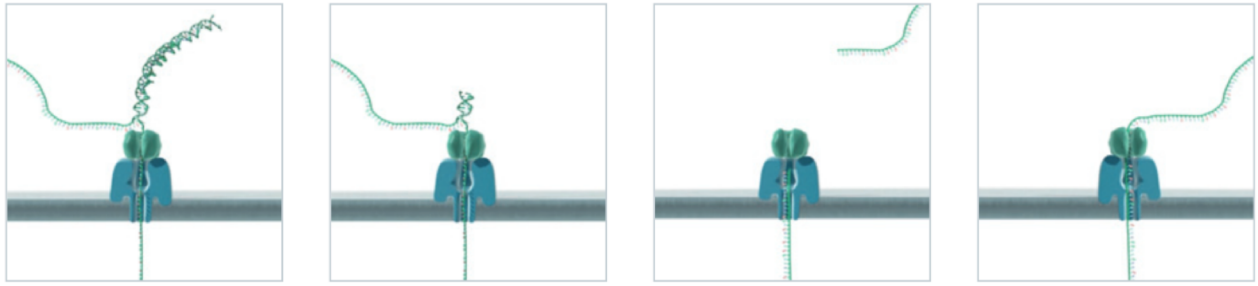
The **nanopore signal**, captured by the ASIC in the device, is characteristic of the sequence of the DNA or RNA fragment. Algorithms are used to convert the signal into basecalls.

1D



Template... ...Template... (Exit) Next molecule…

$1D^2$

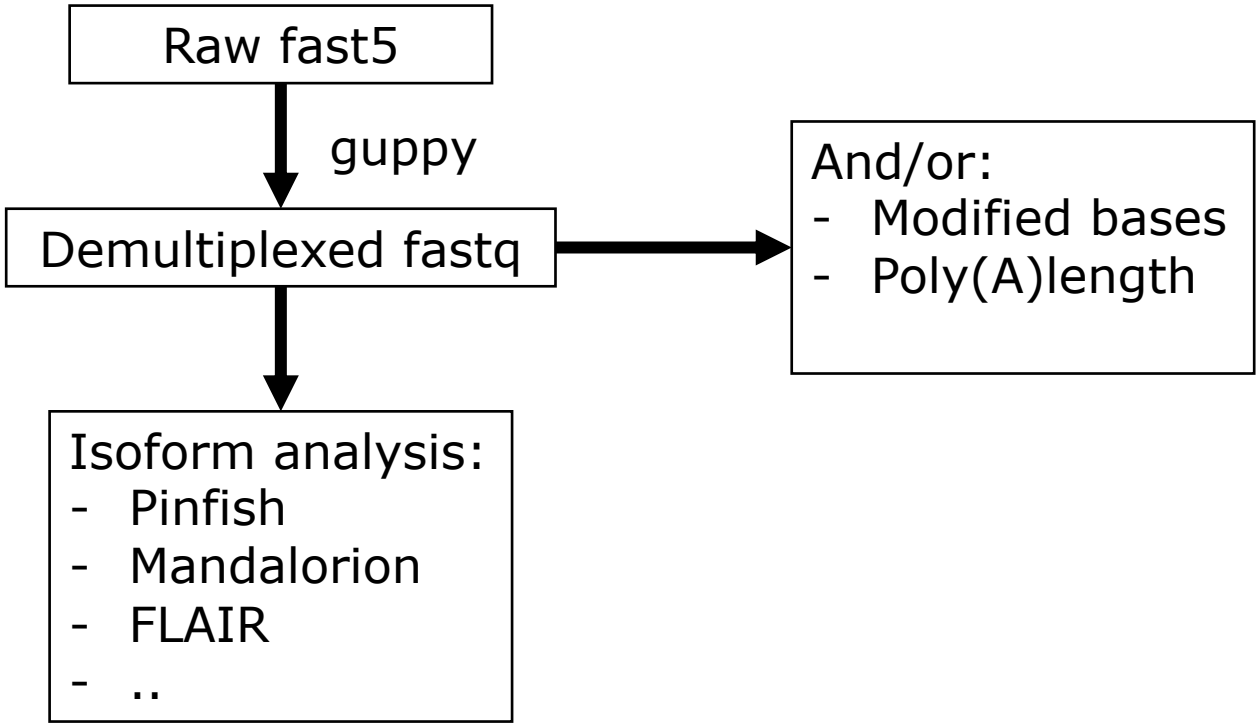


Template... ...Template... (Exit) ...Complement

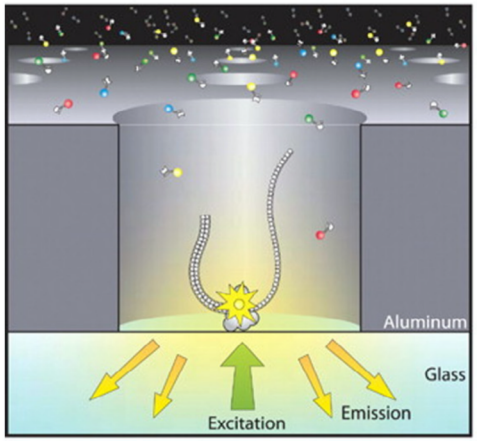
ONT RNA library prep

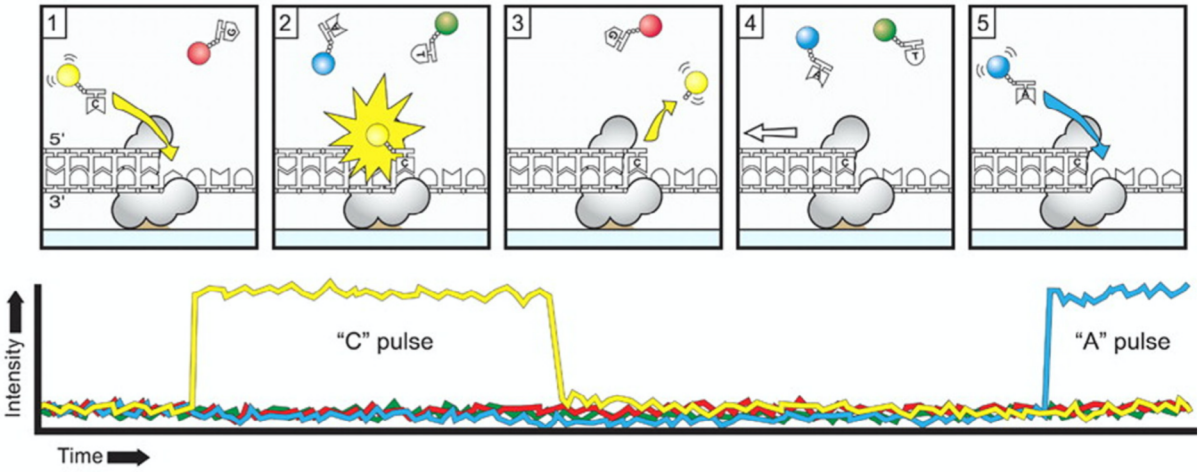
- cDNA PCR
  - Multiplexing
  - Low input
- Direct cDNA
  - Multiplexing
  - No PCR bias
- Direct RNA (dRNAseq)
  - No multiplexing
  - Detect base modifications
  - Polyadenylation

‘Standard’ transcript workflow

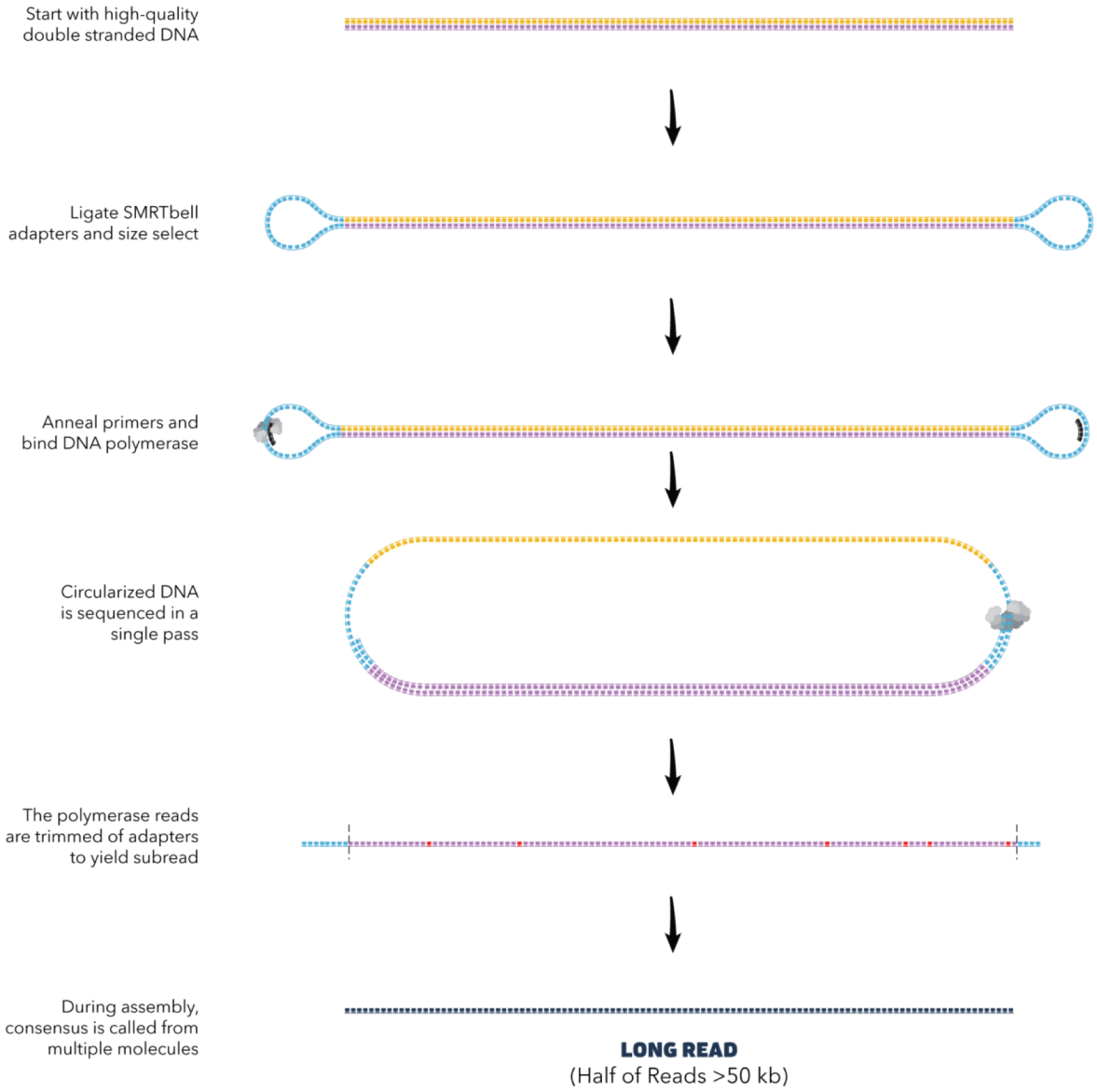


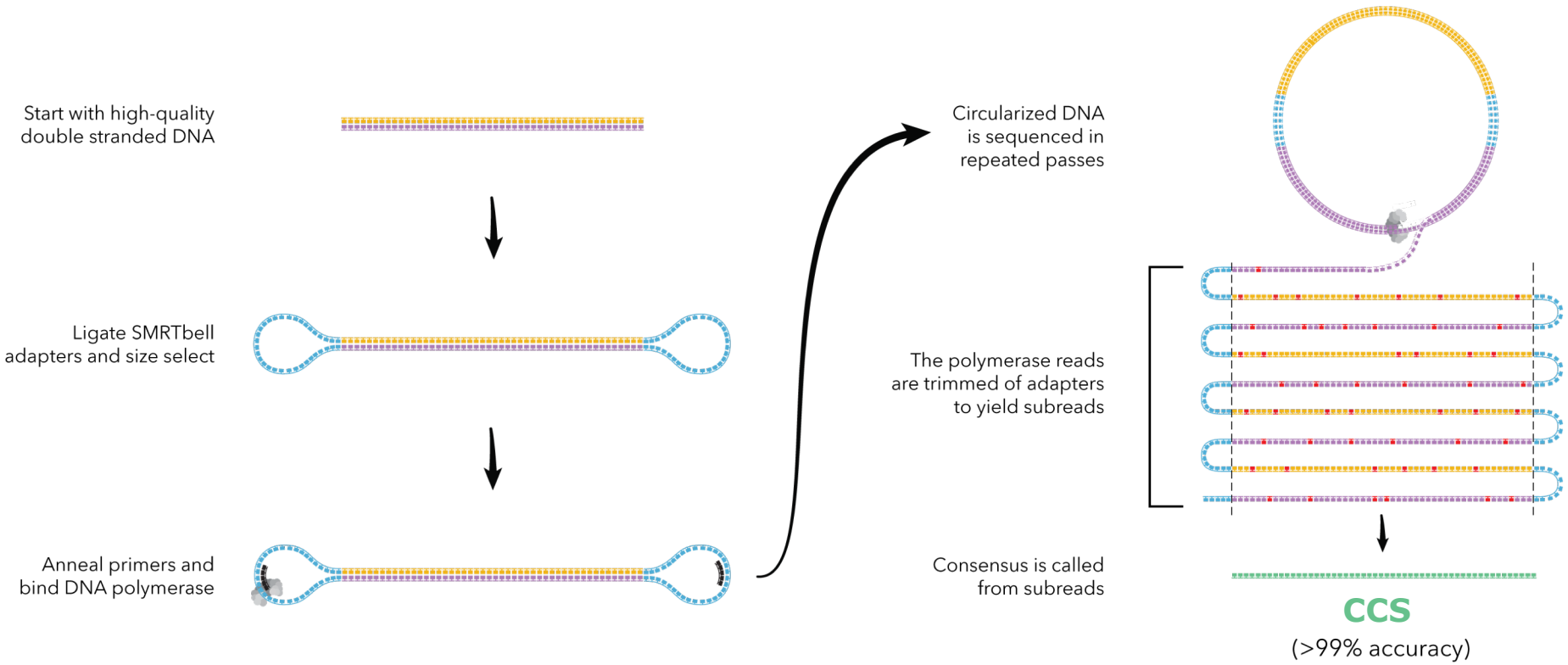
PacBio sequencing



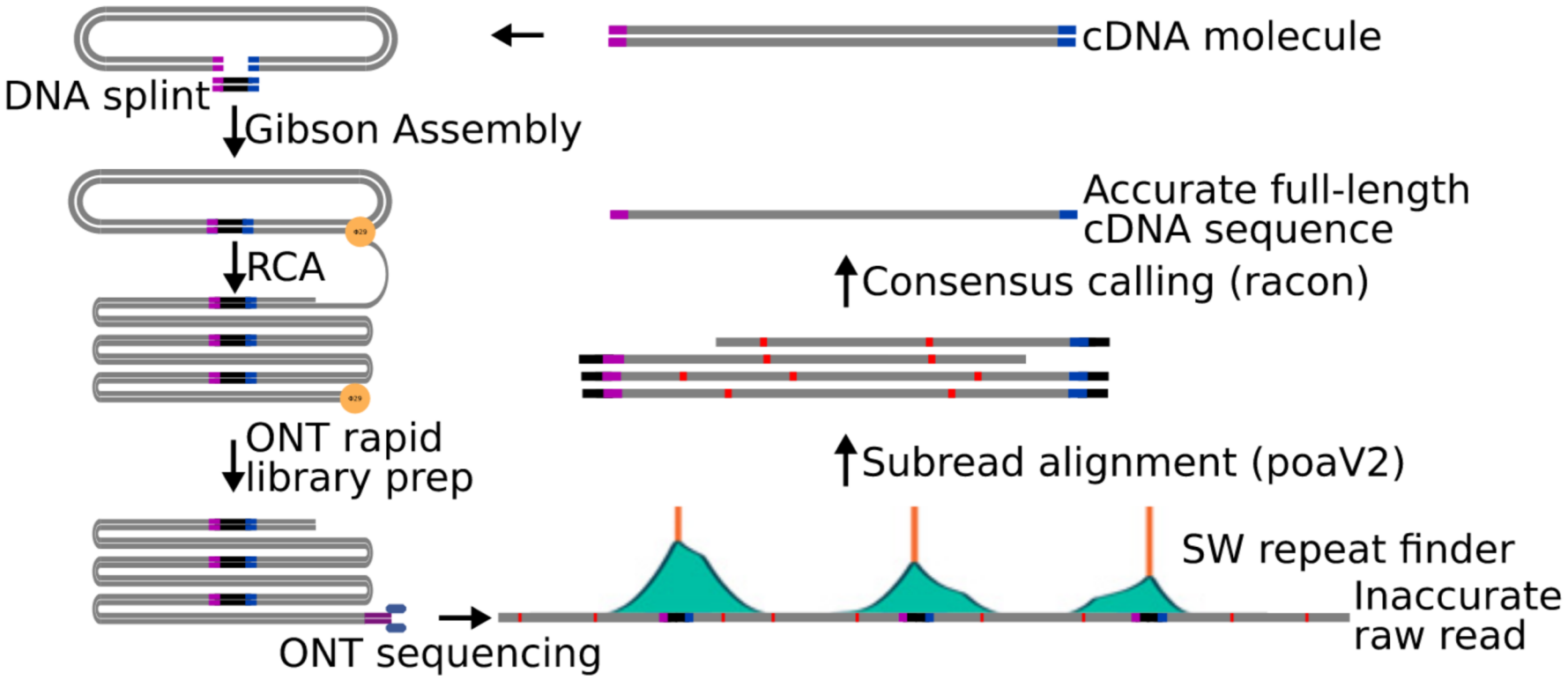


- Polymerase bound to ZMW bottom
- Circular molecules
- Single read out ~90% accuracy
- CCS (HiFi): single molecule sequenced multiple times


Start with high-quality double stranded DNA
Ligate SMRTbell adapters and size select
Anneal primers and bind DNA polymerase
Circularized DNA is sequenced in a single pass
The polymerase reads are trimmed of adapters to yield subread
During assembly, consensus is called from multiple molecules
LONG READ
(Half of Reads >50 kb)


Start with high-quality double stranded DNA
Ligate SMRTbell adapters and size select
Anneal primers and bind DNA polymerase
Circularized DNA is sequenced in repeated passes
The polymerase reads are trimmed of adapters to yield subreads
Consensus is called from subreads
CCS
(>99% accuracy)

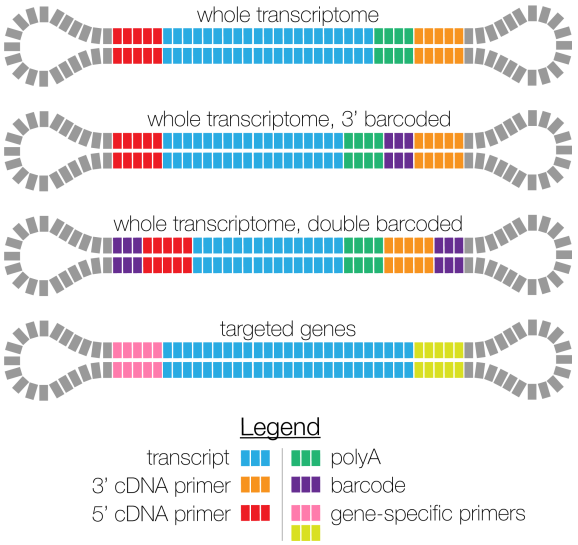
CCS for ONT? R2C2

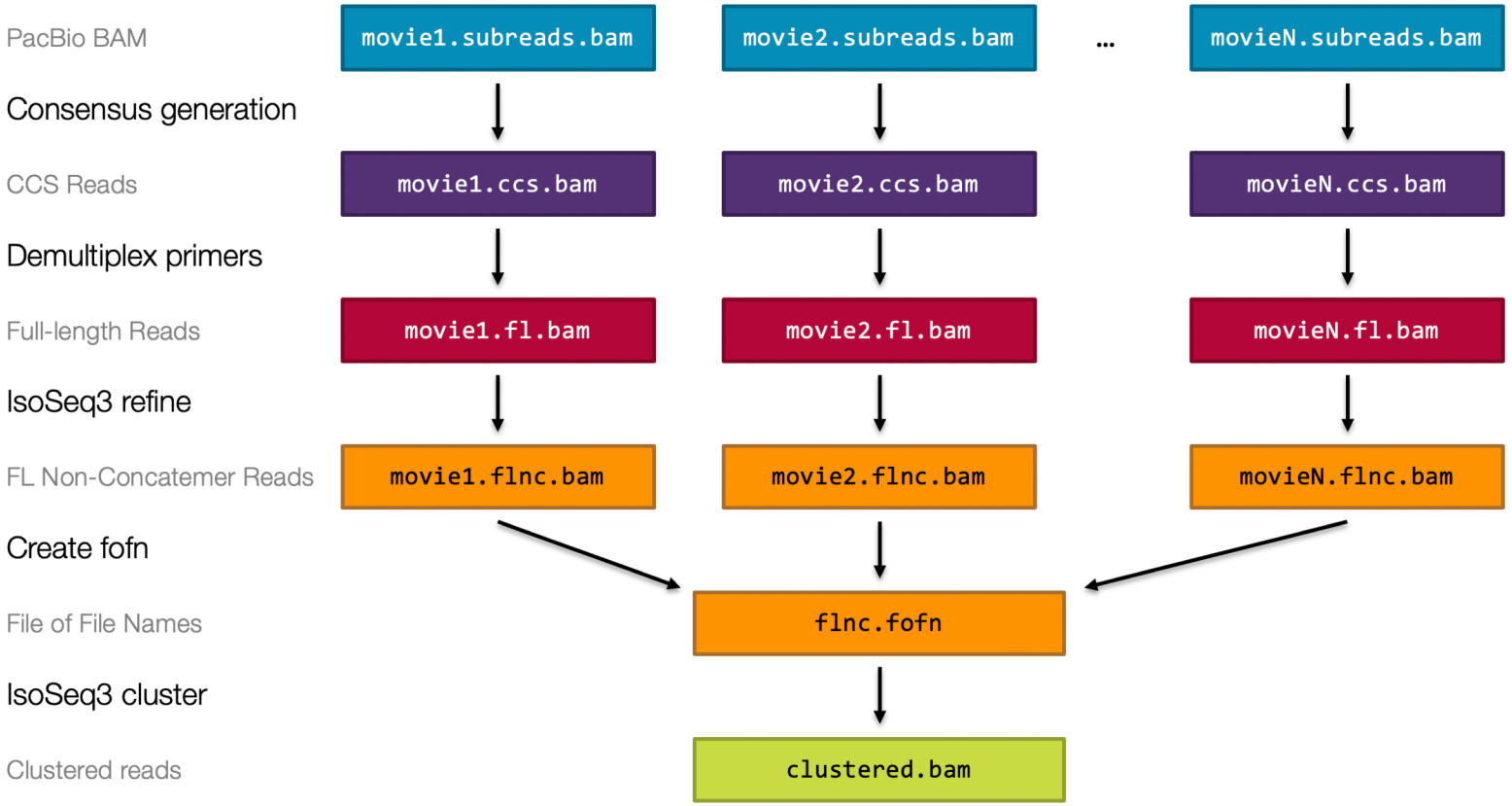


Volden, R. et al (2018). Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *PNAS*, *115*(39)

PacBio RNA library prep

- General workflow: IsoSeq
- cDNA & PCR amplification
- Multiplexing
- github.com/PacificBiosciences/IsoSeq




PacBio BAM
movie1.subreads.bam
movie2.subreads.bam
...
movieN.subreads.bam
Consensus generation
CCS Reads
movie1.ccs.bam
movie2.ccs.bam
movieN.ccs.bam
Demultiplex primers
Full-length Reads
movie1.fl.bam
movie2.fl.bam
movieN.fl.bam
IsoSeq3 refine
FL Non-Concatemer Reads
movie1.flnc.bam
movie2.flnc.bam
movieN.flnc.bam
Create fofn
File of File Names
flnc.fofn
IsoSeq3 cluster
Clustered reads
clustered.bam

	ONT	**PacBio**
Read accuracy	~97%	~90% (>99% CCS)
Read length	Transcript length	Transcript length
Transcript quantification	Yes, PCR & cDNA free possible	Yes, with PCR
RNA base modifications	Yes (m6A)[1]	No
DNA base modifications	Yes (m5C, m6A)[2]	Yes (m5C, m6A, hm5C)[3]
Throughput (BIF)	~250M reads/run	~30M CCS reads/run

1. Liu, H., et al (2019). Accurate detection of m6A RNA modifications in native RNA sequences. *Nature Communications*, *10*(1), 1–9
2. Liu, Q., et al (2019). Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nature Communications*, *10*(1).
3. Flusberg, B. A., et al (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, *7*(6), 461–465

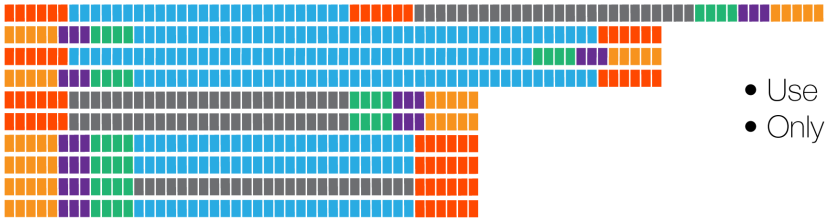
Iso-Seq Clustering

transcript gene A
transcript gene B
optional polyA

3' cDNA primer
5' cDNA primer
sample barcode

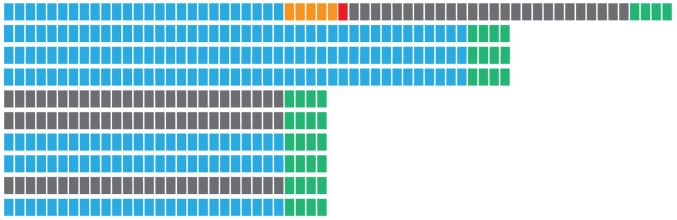
Version 8, Dr. Armin Töpfer

consensus



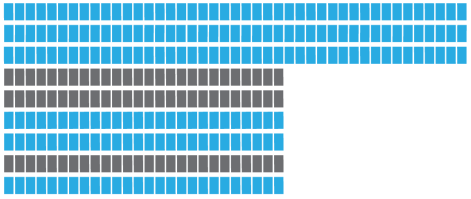
- Use **polished** CCS reads
- Only full-pass ZMWs

demultiplex



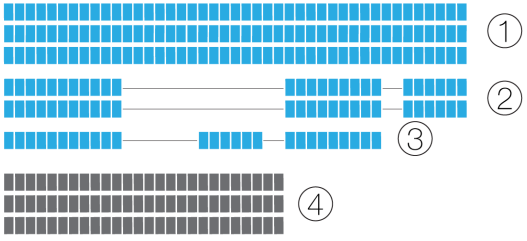
- Barcoded and unbarcoded cDNA primer removal
- Orientation
- Unwanted primer combination removal

refine



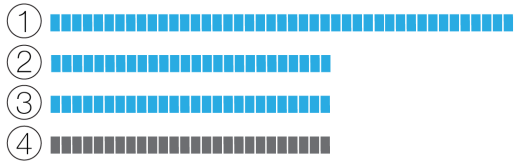
- PolyA tail trimming
- Concatemer removal

cluster



- Hierarchical, n*log(n) clustering, alignment of shorter to longer sequences
- Iterative cluster merging
- Generate consensus for each read cluster using QV guided PoA

fasta output



- Fasta output is split into HQ and LQ reads
- One consensus per read cluster

Isoform analysis software

- Pinfish
  - Developed by nanoporetech
  - github.com/nanoporetech/pinfish
- Mandalorion
  - C3PO/R2C2 only
  - github.com/rvolden/Mandalorion-Episode-II
- FLAIR
  - Optional polishing with Illumina reads
  - Also PacBio
  - github.com/BrooksLabUCSC/flair