NGS - quality control, alignment, visualisation

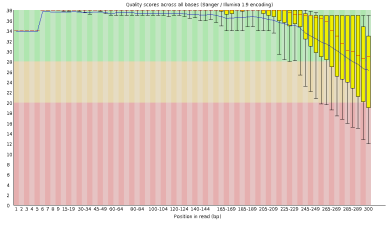
Geert van Geest
Interfaculty Bioinformatics Unit, UniBe
Training group, SIB

Major applications

- Transcriptome characterization
  - e.g. RNA-seq
- Epigenome characterization:
  - e.g. ATAC-seq
- DNA-protein interactions:
  - e.g. ChIP-seq
- Whole genome (assembly)
- Variant detection
- Metagenome characterization
- Any others?

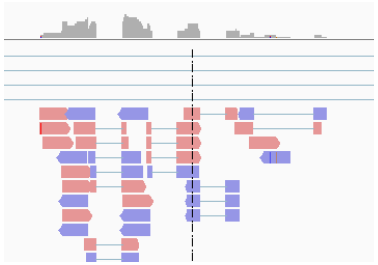
Sequencing

↓



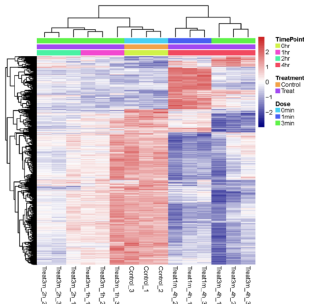
Quality control

↓



Alignment

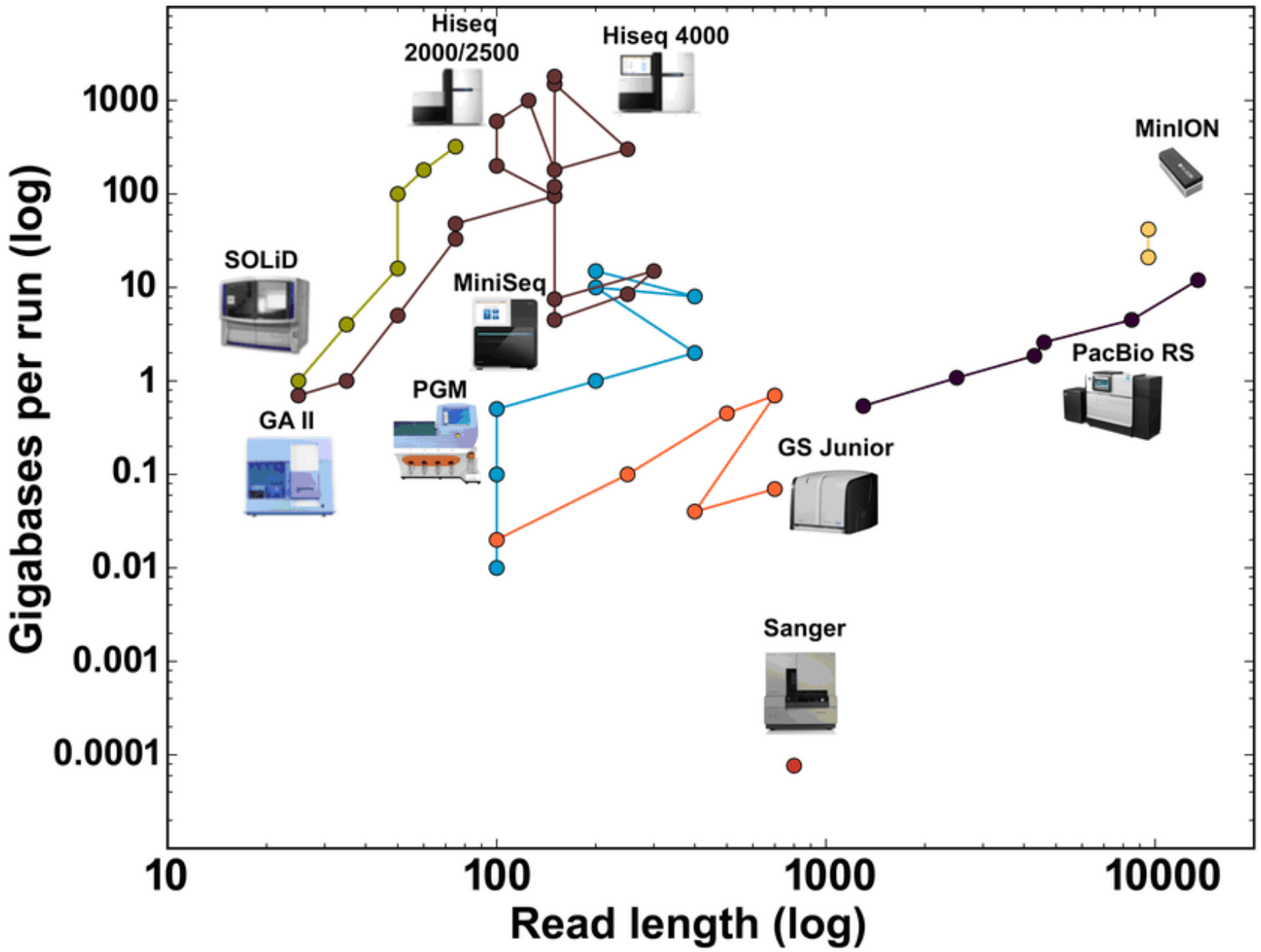
↓



Down-stream analysis

↓

Visualisation


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)

Question

Altough Sanger sequencing has been around since the 70s, it's still used by a lot of labs. Why is that?

A. Because the per-base sequence costs are very low
B. Because it's scalable
C. Because it's output quality is still unchallenged

Illumina sequencing

- Sequencing-by-synthesis: 2nd generation sequencing
- Massive throughput
- Most used platform today

illumina®

Illumina sequencing

- 50 – 300 bp
- Paired-end (or single-end)
- Multiplexing

Illumina libray prep

shear DNA

↓

Ligate adapters

↓

Universal primers + barcode + p5/p7 sites

↓

PCR: 8-16 cycles

↓

Sequencing

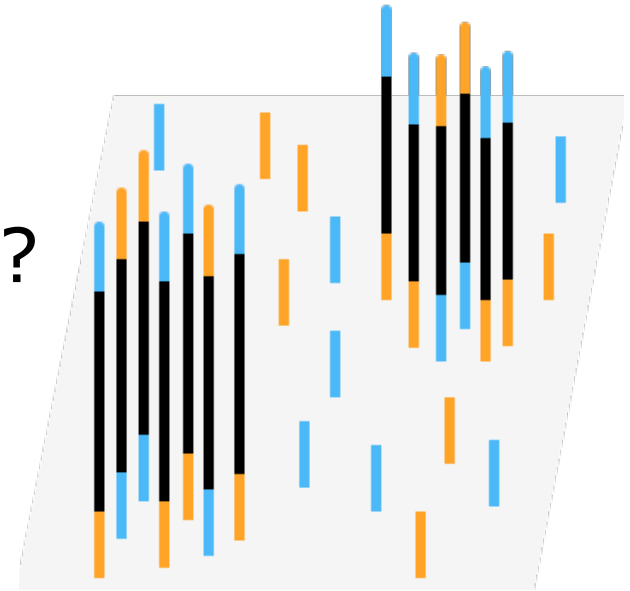
Question

Why is use of barcodes important for Illumina sequencing?

A. With barcodes, the fragment can anneal to the flow cell
B. The barcodes are used to add universal sequencing primers with PCR
C. With barcodes it is possible to sequence multiple samples in one flow cell

Illumina - limitations

- Maximum read length: 300 bp
- How to reconstruct:
  - Repeats?
  - Isoforms?
  - Structural variation?
  - Haplotypes?
  - Genomes?
- Why not longer read lengths?

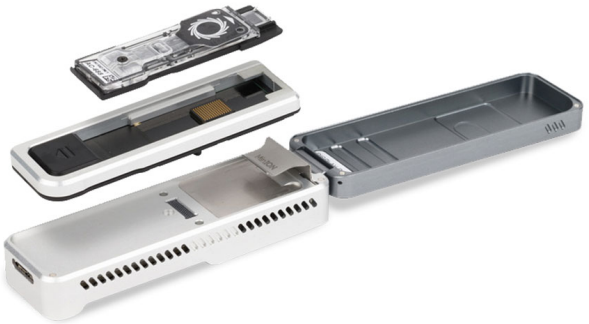


Long reads (3rd generation)

- Crux: maximizing signal from a single-molecule base read-out
- Single molecule, so no out-of-phase signal
- Two frequently used platforms:
  - PacBio SMRT sequencing
  - Oxford Nanopore Technology

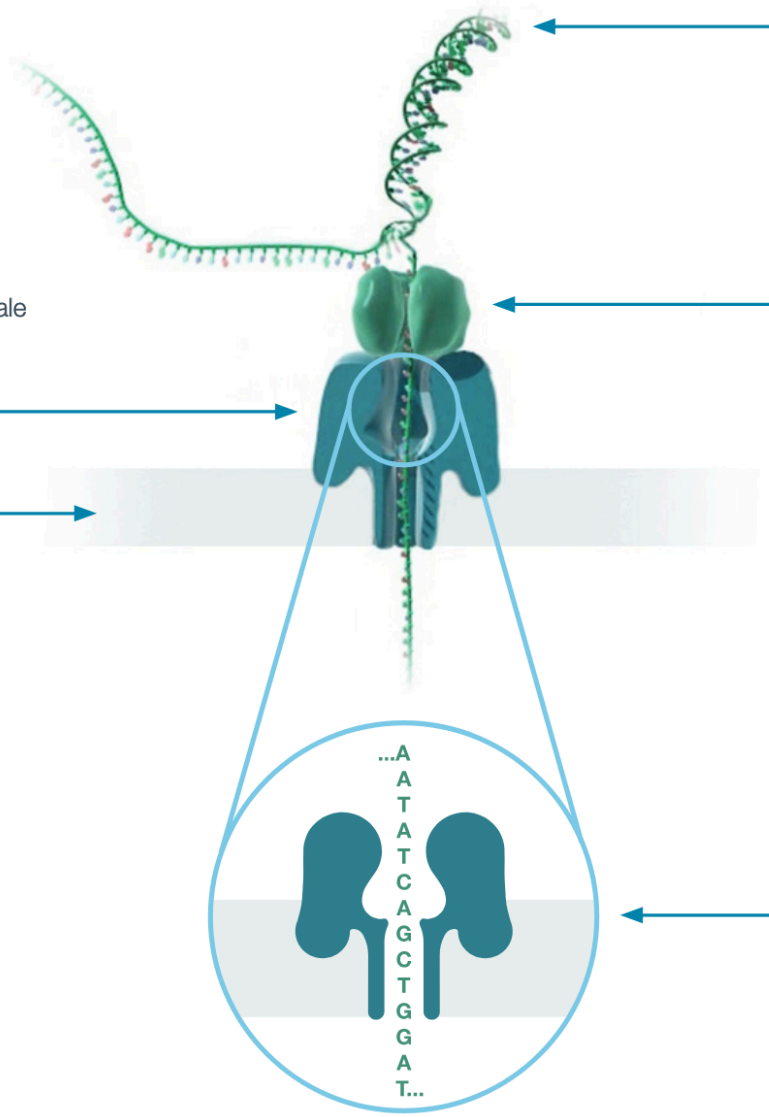
Oxford Nanopore technology

- 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.

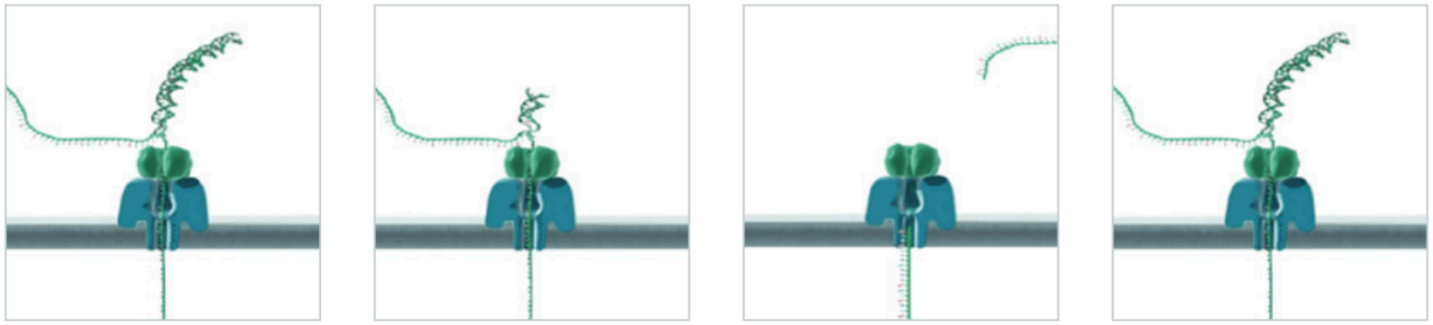


The nanopore processes the length of **DNA or RNA** presented to it. The user can control this through the library preparation protocol utilised. (e.g. >2 Mb DNA has been recorded.)

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.

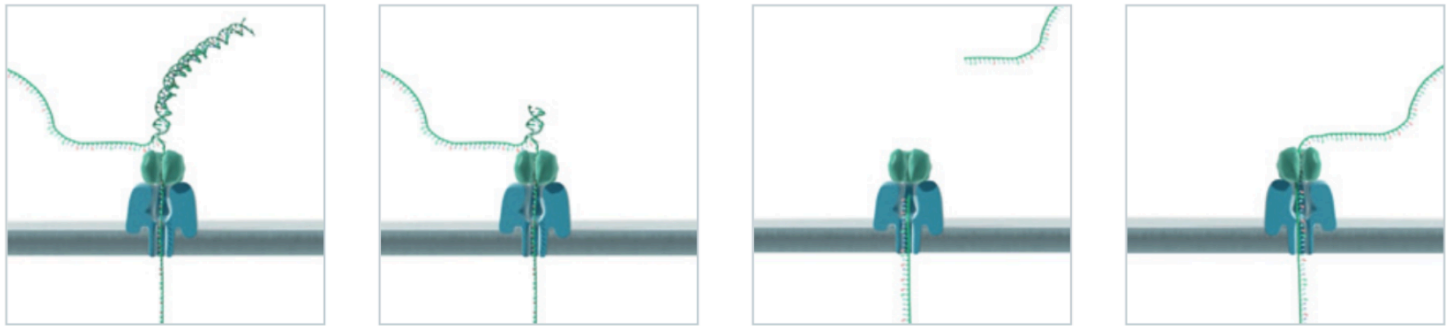
The **nanopore signal**, captured by the ASIC in the device, is characteristic of the sequence of the DNA or RNA fragment. Algorithms are used to convert the signal into basecalls.

1D



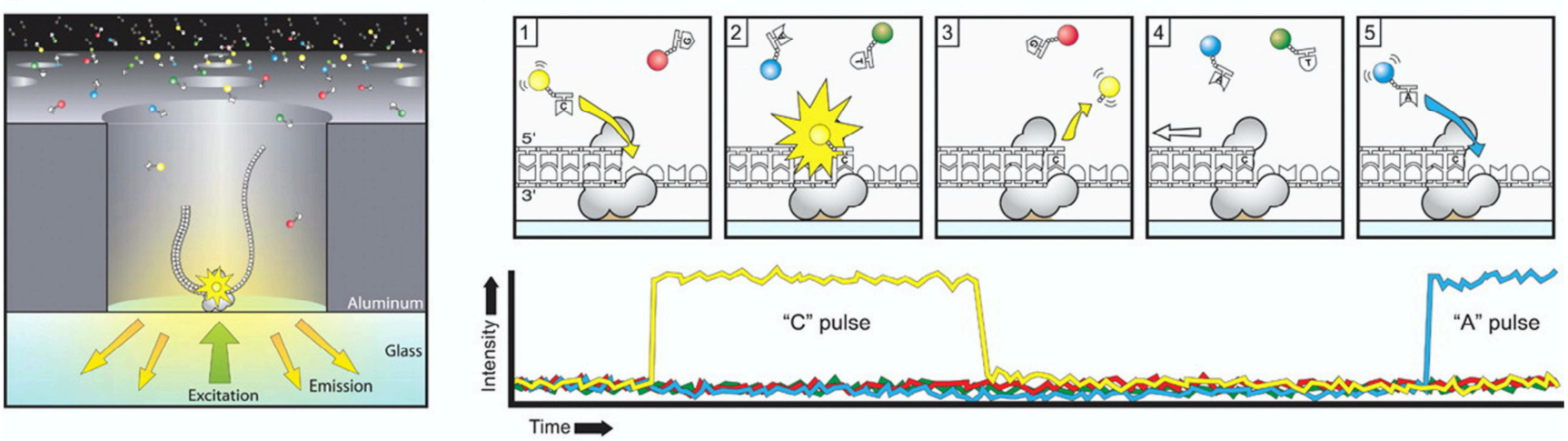
Template... ...Template... (Exit) Next molecule…

$1D^2$

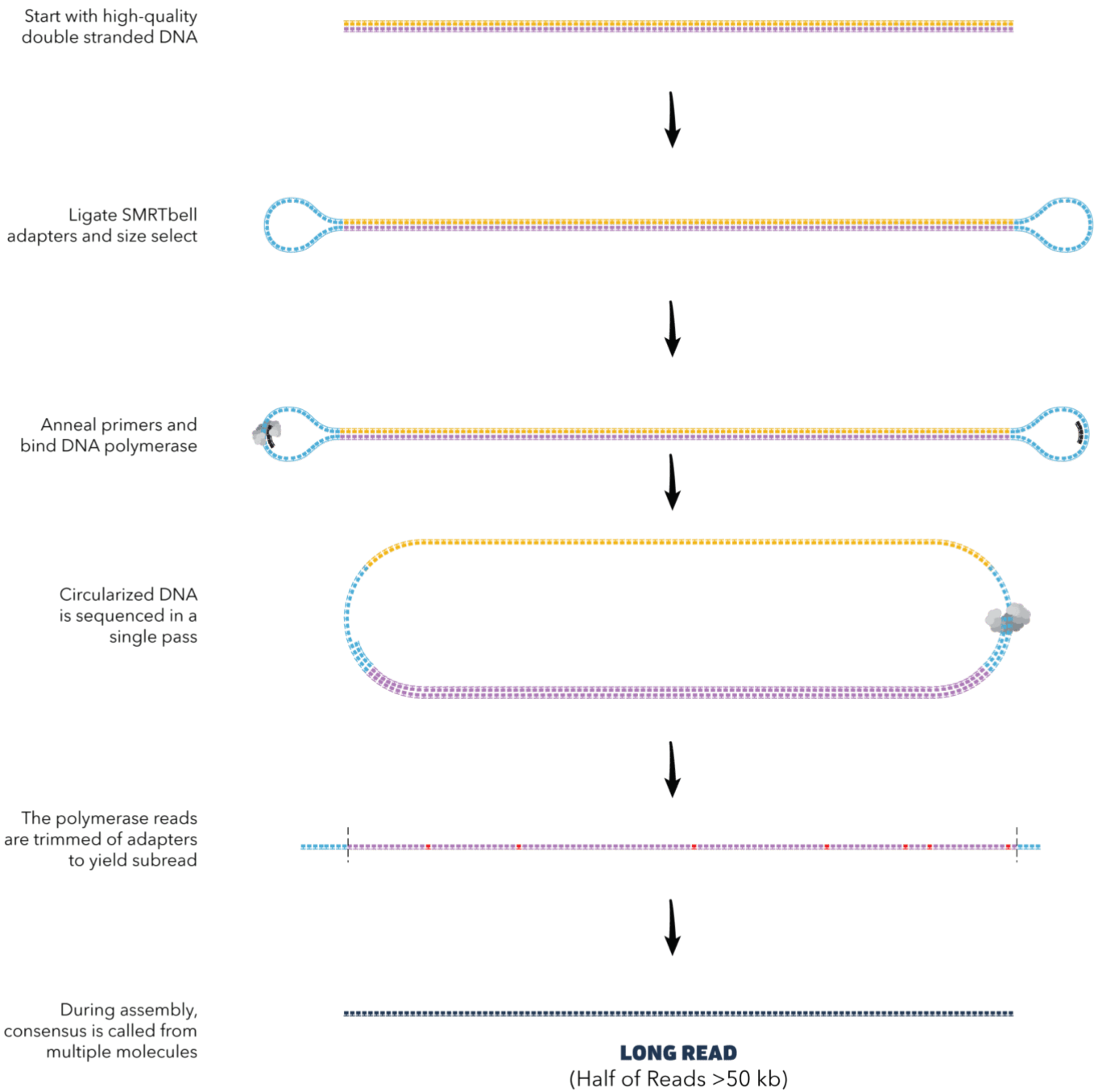


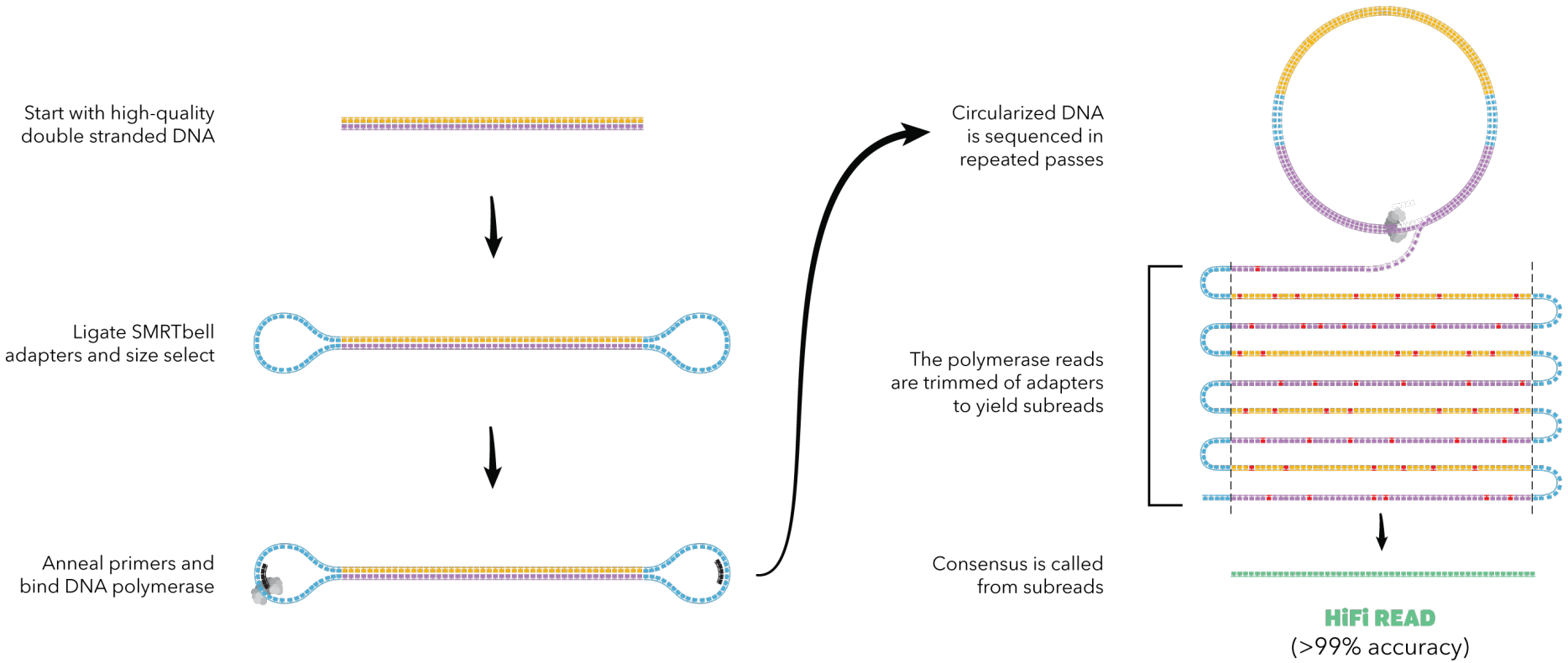
Template... ...Template... (Exit) ...Complement

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
HiFi READ
(>99% accuracy)

Question

What kind of invention led to the ability to sequence (very) long reads?

A. Base incorporation that doesn’t get exhausted
B. Differentiating between bases at a single-molecule scale
C. PCR with unlimited amplicon length

Question

Why are error rates for long read sequencing relatively high?

A. Differentiating between nucleotides at a single-molecule scale is challenging
B. The out-of-phase signal results in low base quality
C. Longer sequences have a larger chance on PCR error