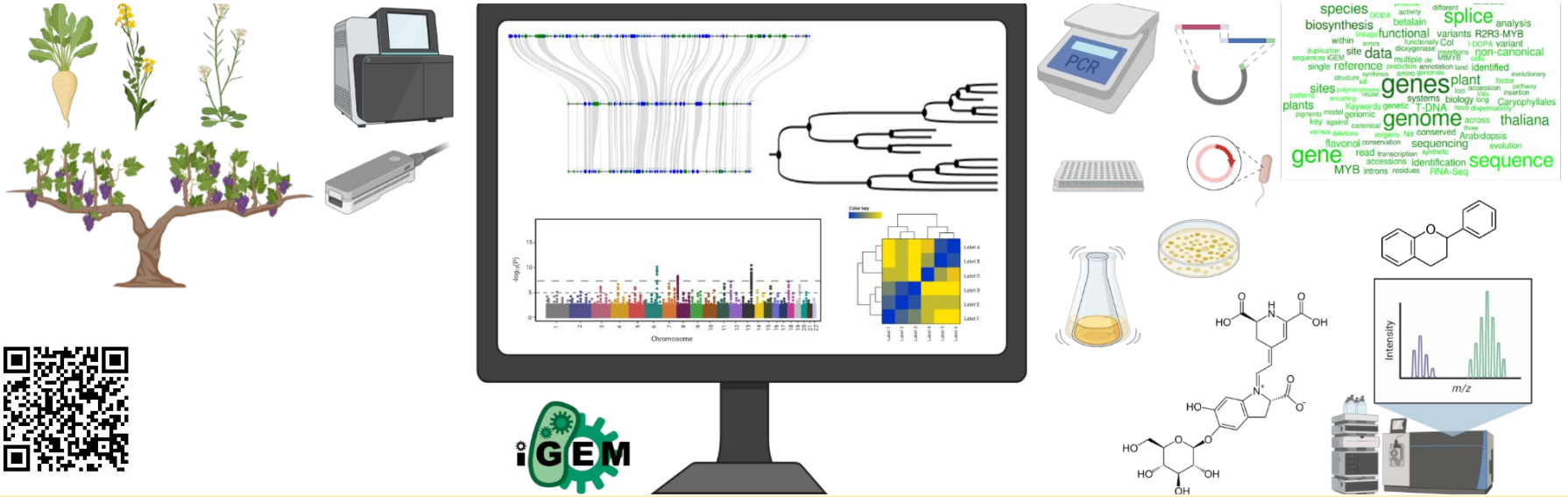
Technische Universität Braunschweig



Sequencing

Prof. Dr. Boas Pucker
(Plant Biotechnology and Bioinformatics)

Availability of slides

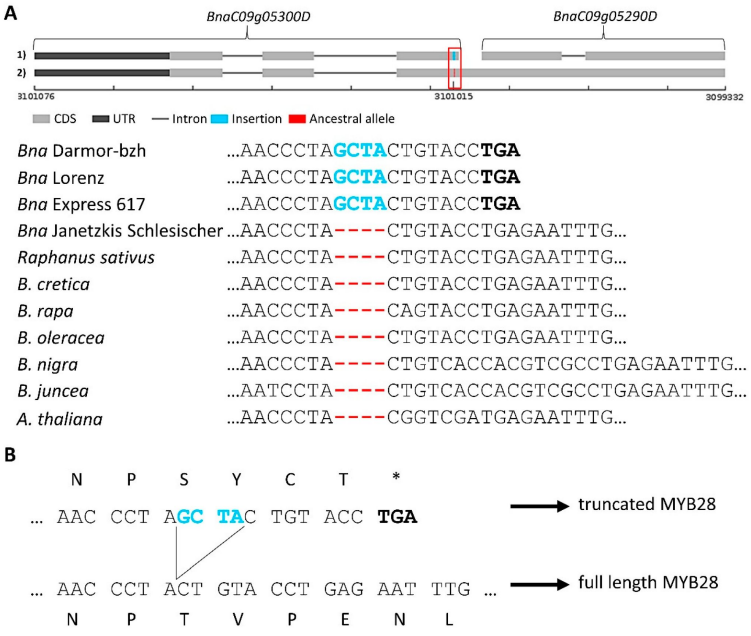
- All materials are freely available (CC BY) - after the lectures:
  - StudIP: LMChemBSc12
  - GitHub: https://github.com/bpucker/teaching

- Questions: Feel free to ask at any time

- Feedback, comments, or questions: b.pucker[a]tu-bs.de

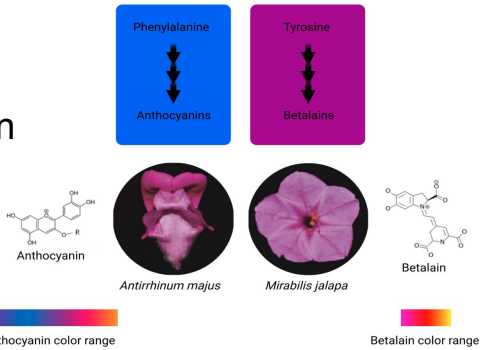
Crop genomics

Glucosinolate content in rapeseed is controlled by the transcription factor MYB28



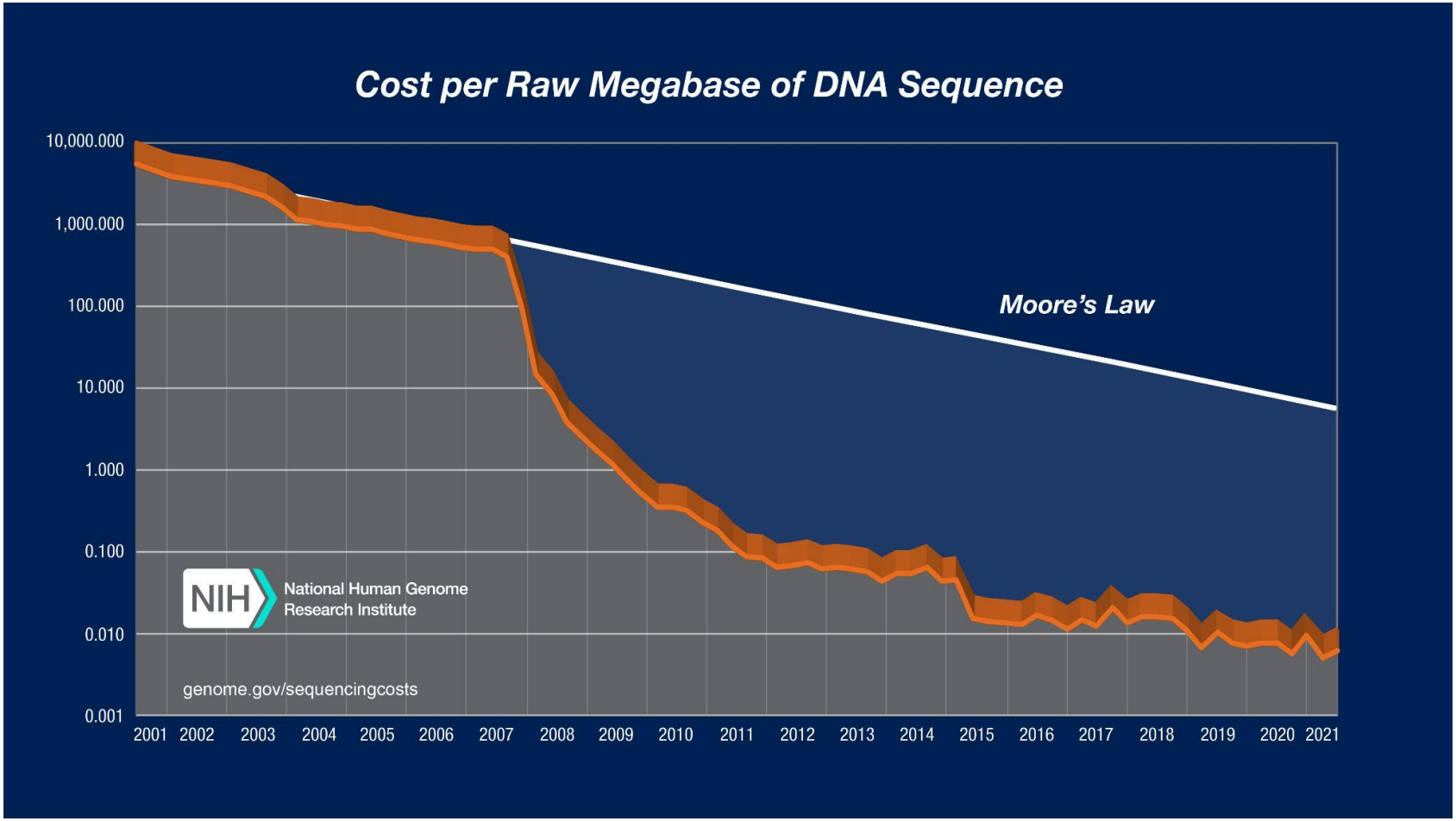
10.3390/genes13071131

Genomics helps to explain mutual exclusion of anthocyanins and betalains



10.1101/2022.10.19.512958

‘Big Data’


Cost per Raw Megabase of DNA Sequence
10,000.000
1,000.000
100.000
10.000
1.000
0.100
0.010
0.001
Moore’s Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021

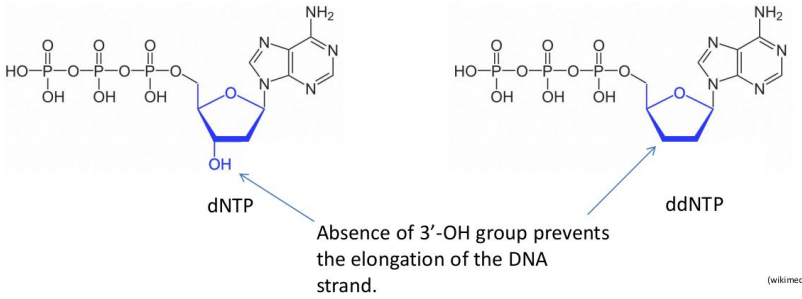
Which sequencing methods do you know?

Overview of sequencing technologies

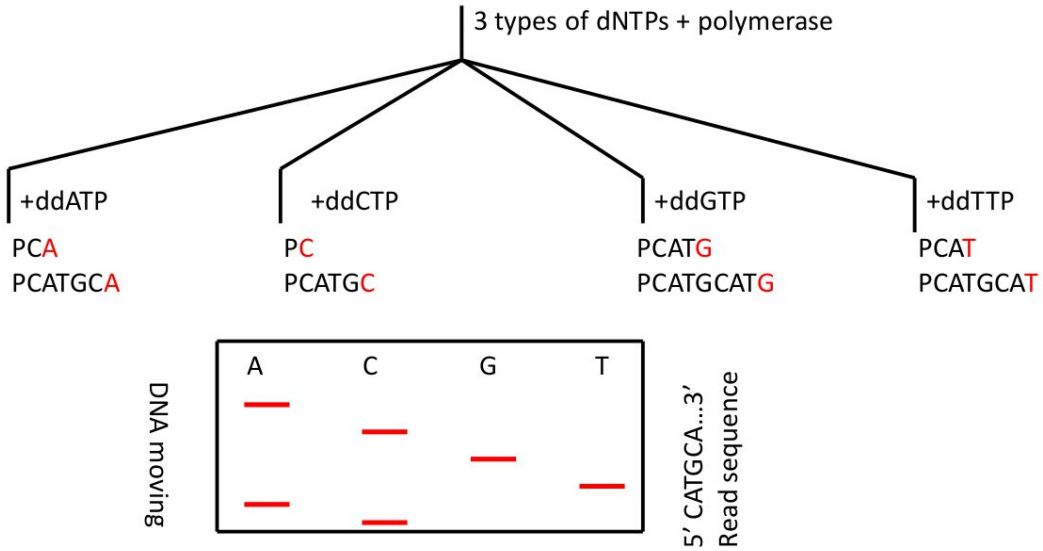
- Generation 1:
  - **Sanger sequencing**
  - Max-Gilbert sequencing
- Generation 2 (massive parallel sequencing):
  - 454 pyrosequencing
  - Solexa/**Illumina sequencing**
  - SOLID
  - Ion Torrent
  - BGI-seq
  - Synthetic long reads
- Generation 3 (long reads):
  - **Pacific Biosciences (PacBio)**
  - **Oxford Nanopore Technologies (ONT)**
- Generation 4:
  - What is next?

Sanger sequencing

Concept of Sanger sequencing

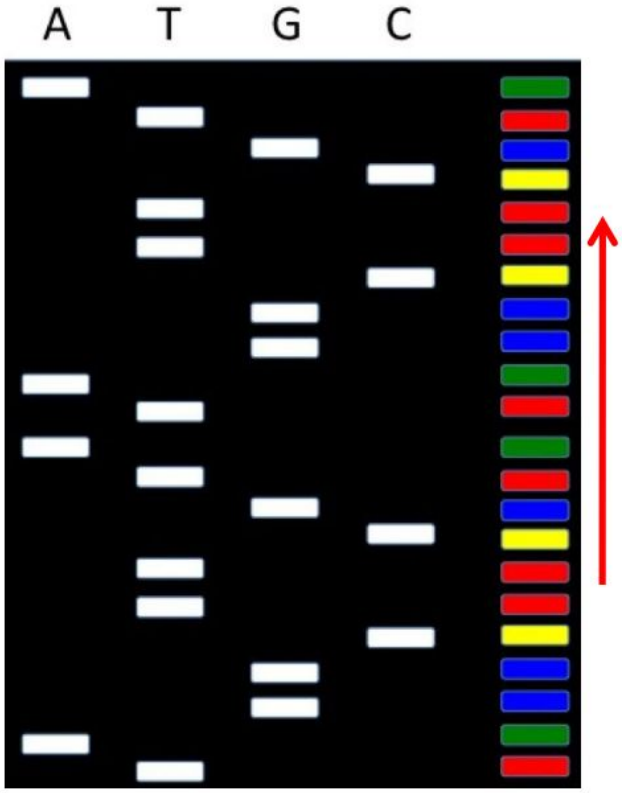


Primer (P): 5' –TGCATGGCATGATGCATG-3'
Template: 3' –ACGTACCGTACTACGTACGTACGTACGTCTAGGT-5'



Sanger sequencing - original version

Two persons analyze the gel: one is calling the base (‘basecaller’) and the other person is writing down the bases



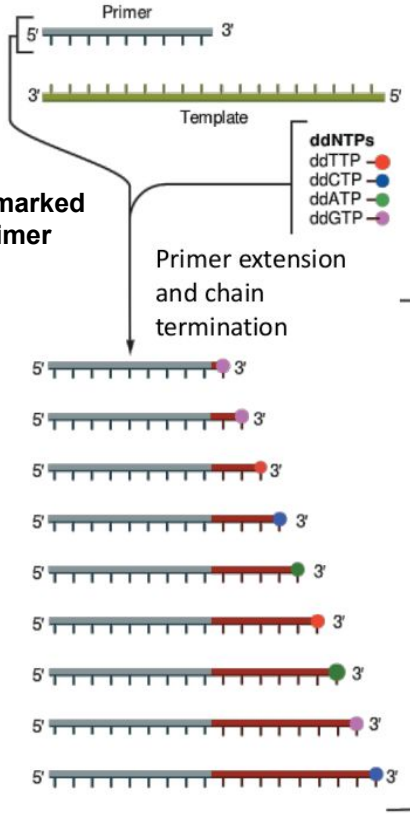
(modified from wikimedia.org)

Sanger sequencing - today

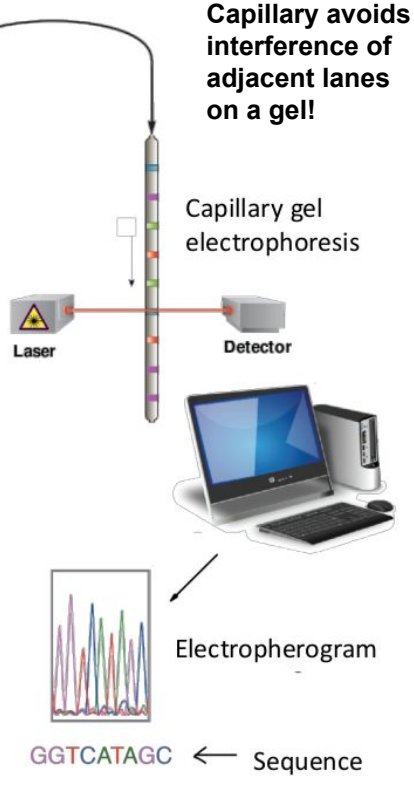
Only one reaction!

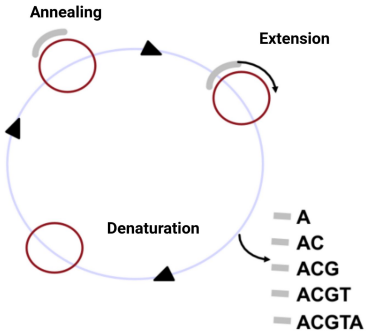
ddNTPs are marked instead of primer

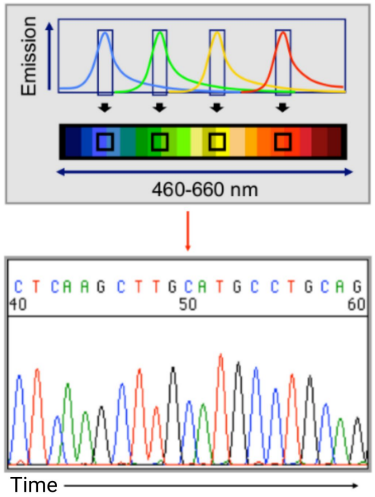
Low input required due to cycle sequencing



Capillary avoids interference of adjacent lanes on a gel!







Figures modified from wikipedia

FASTA

- There are two types of lines: header and sequence
- Header line starts with ‘>’; can contain name and information about sequence
- Example:

```
>seq1 len=5
ACGTA
>seq2 len=10
ACGTA
ACGTA
>seq len=1
A
```

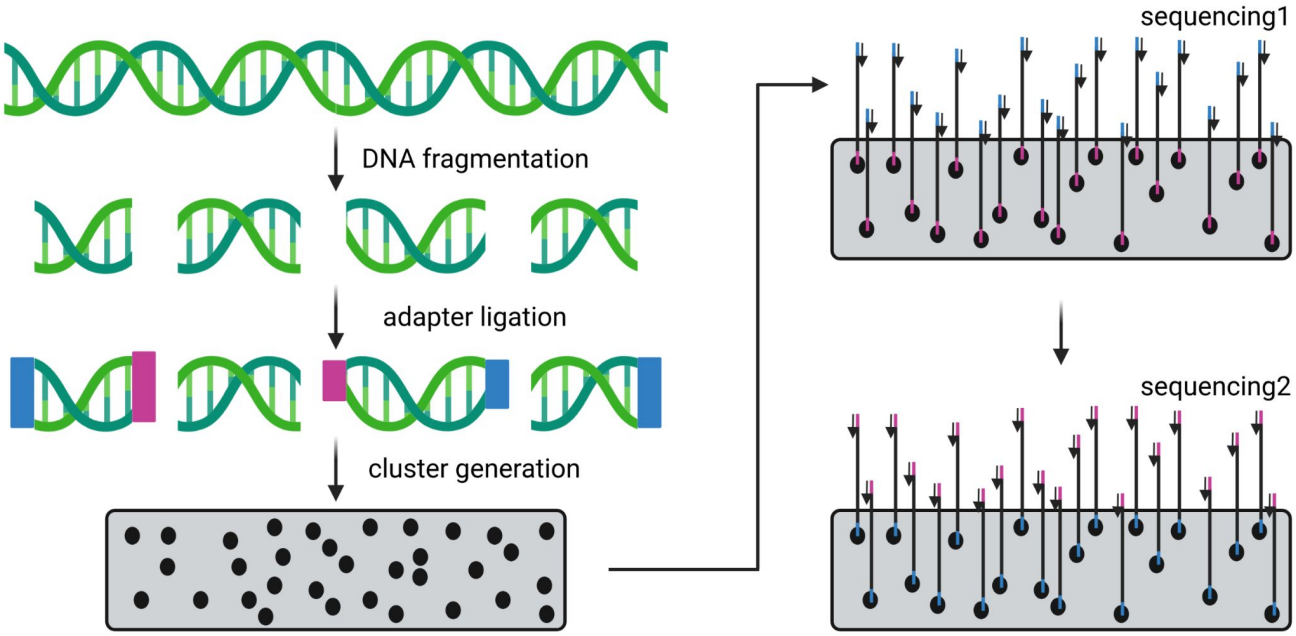
Phred-Score

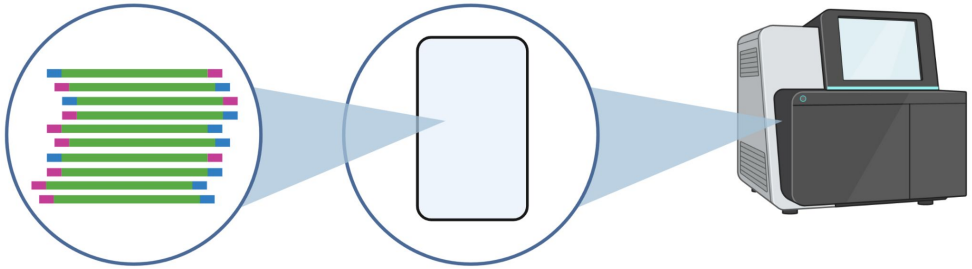
- Negative logarithm of the error probability for given position in read
- Multiplication by 10 to avoid floats

Phred quality score	Error probability	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

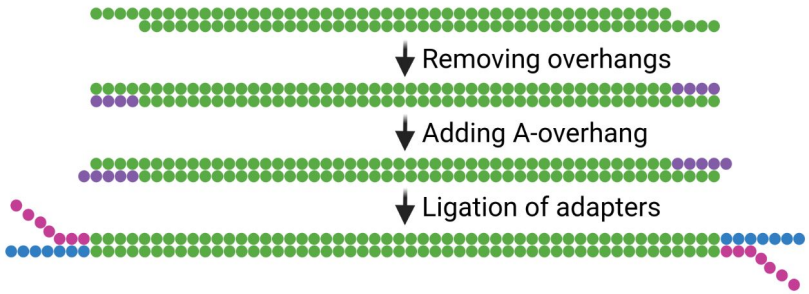
Illumina sequencing

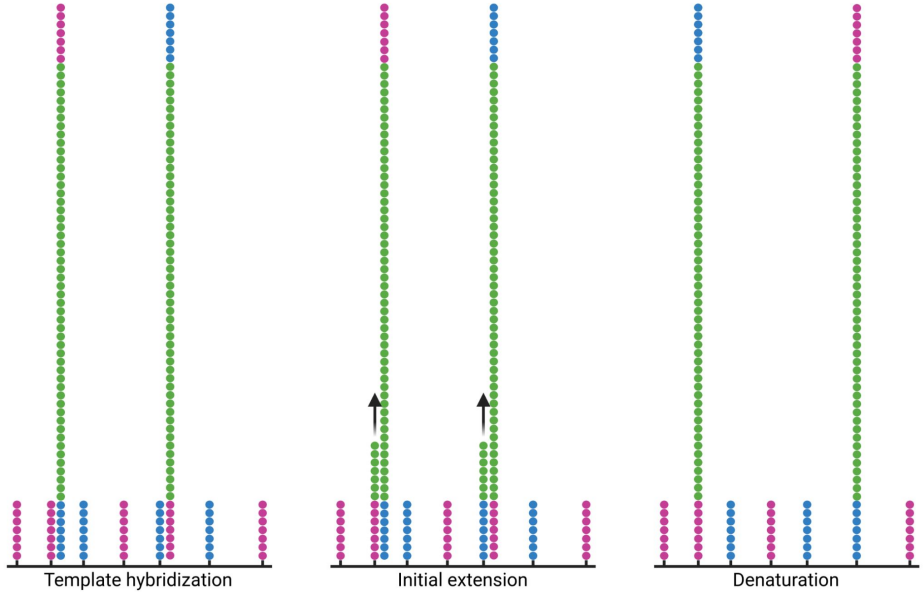
Illumina - sequencing overview



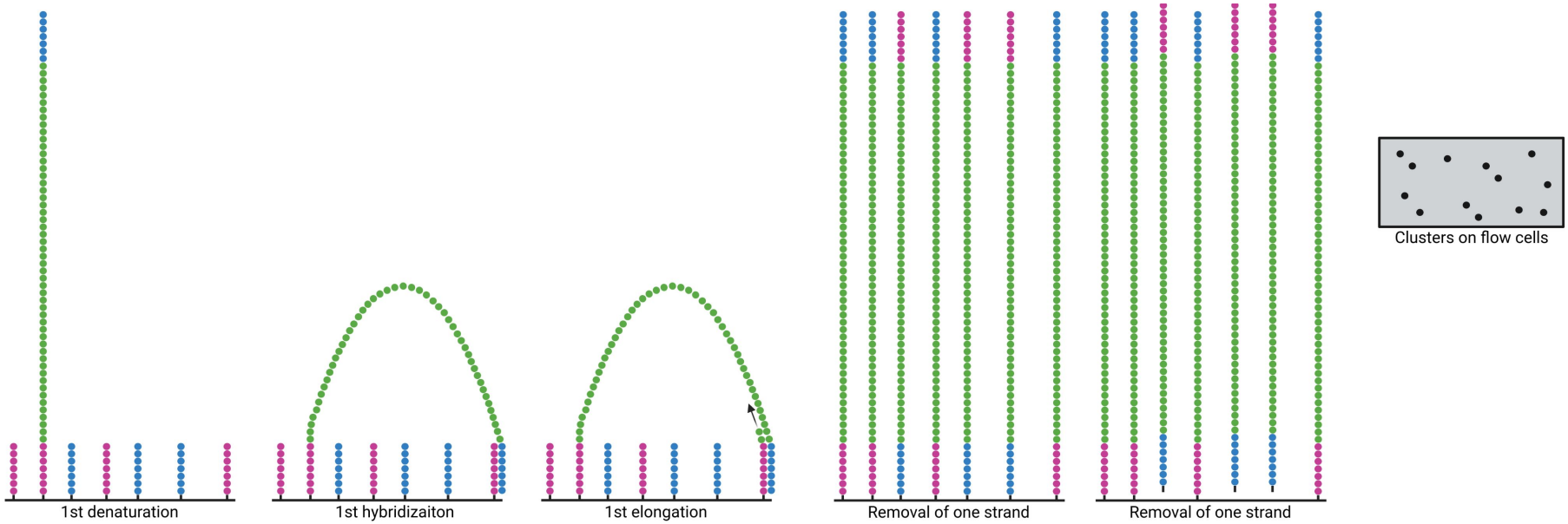


Illumina - sequencing 2

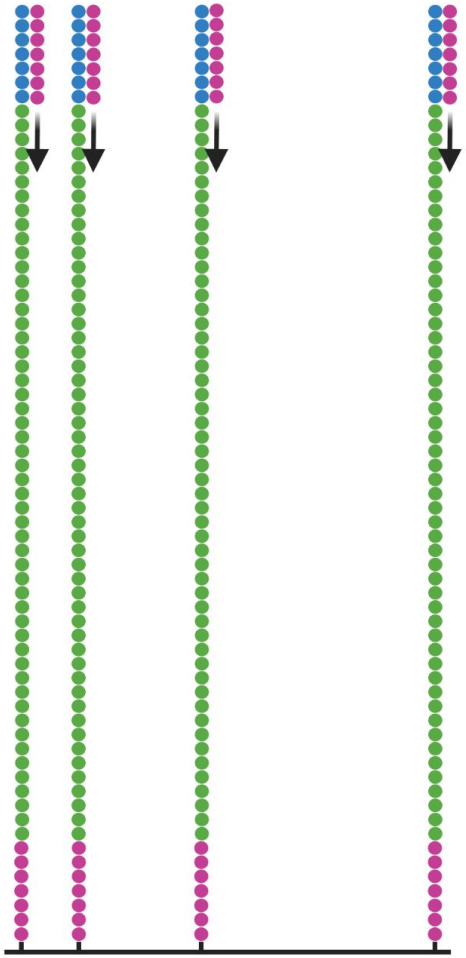




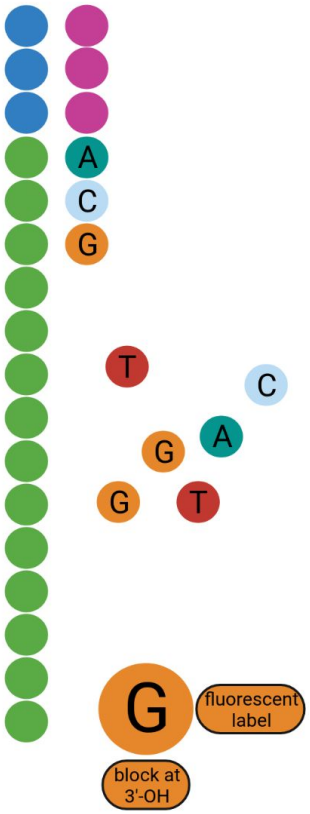
Illumina - sequencing 3



Illumina - sequencing 4

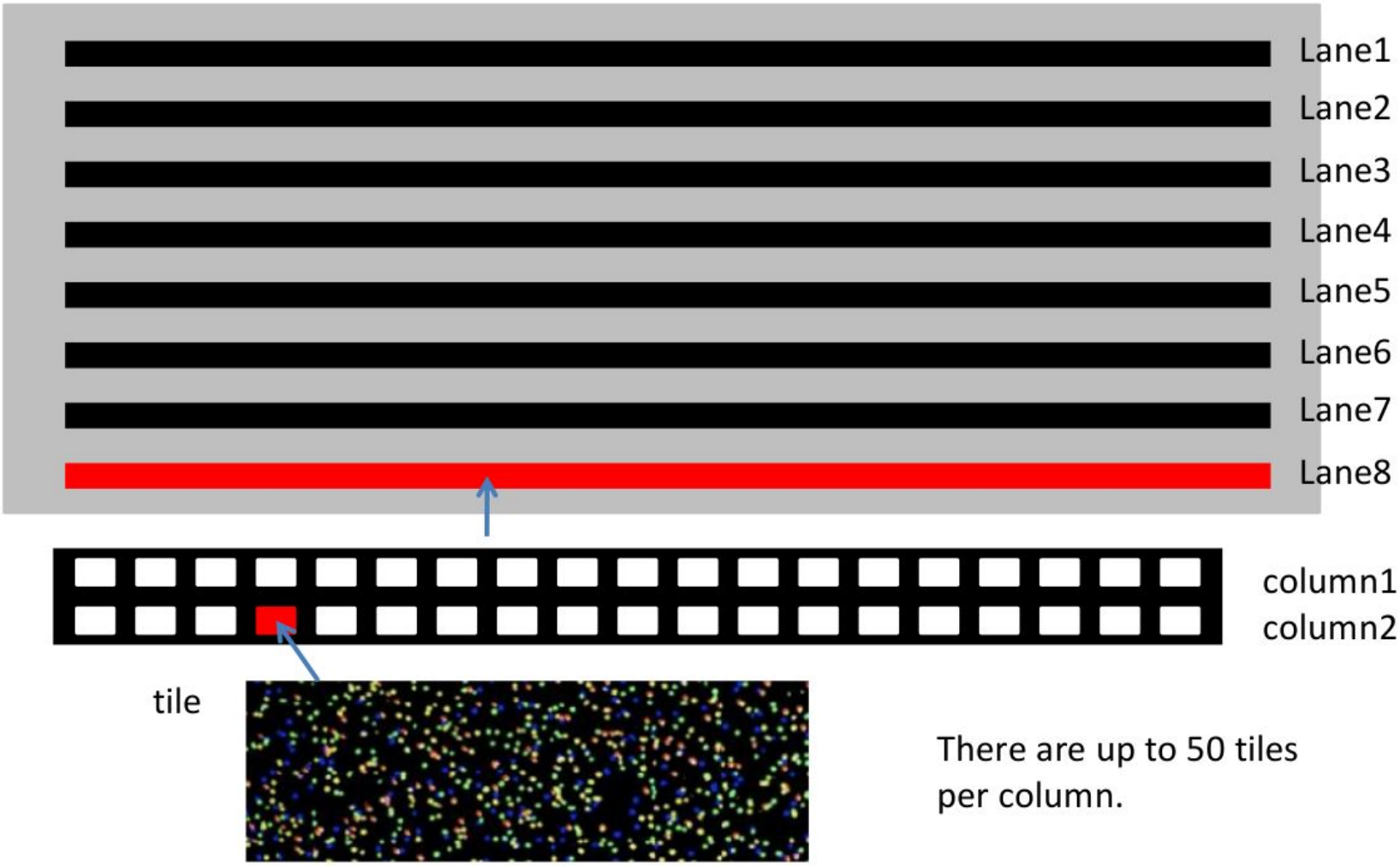


Sequencing

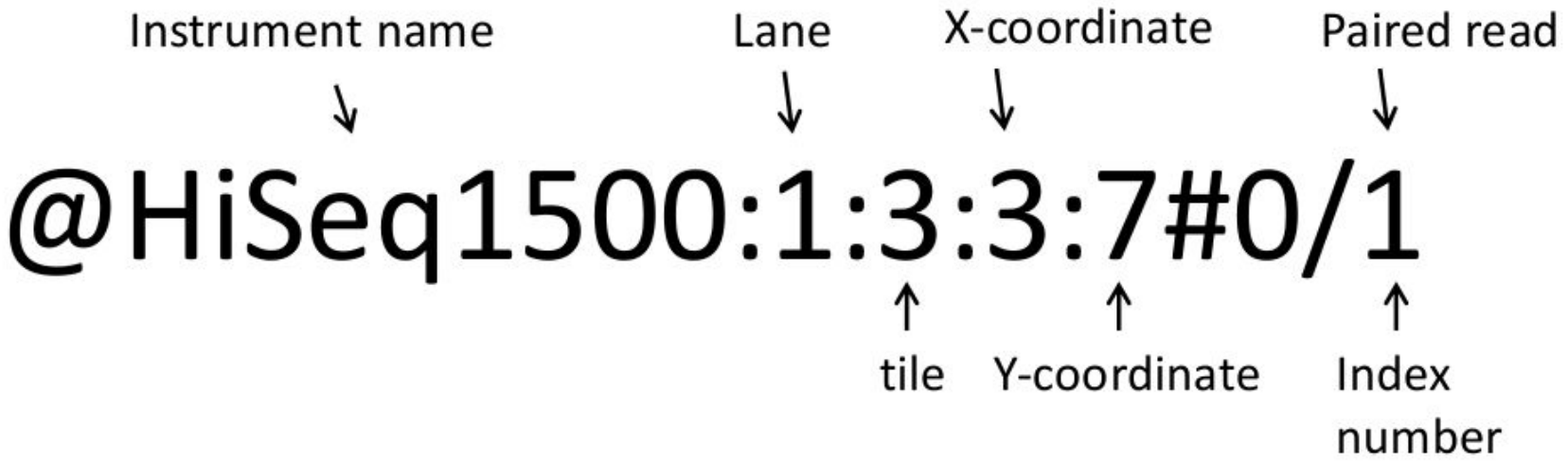


Sequencing chemistry

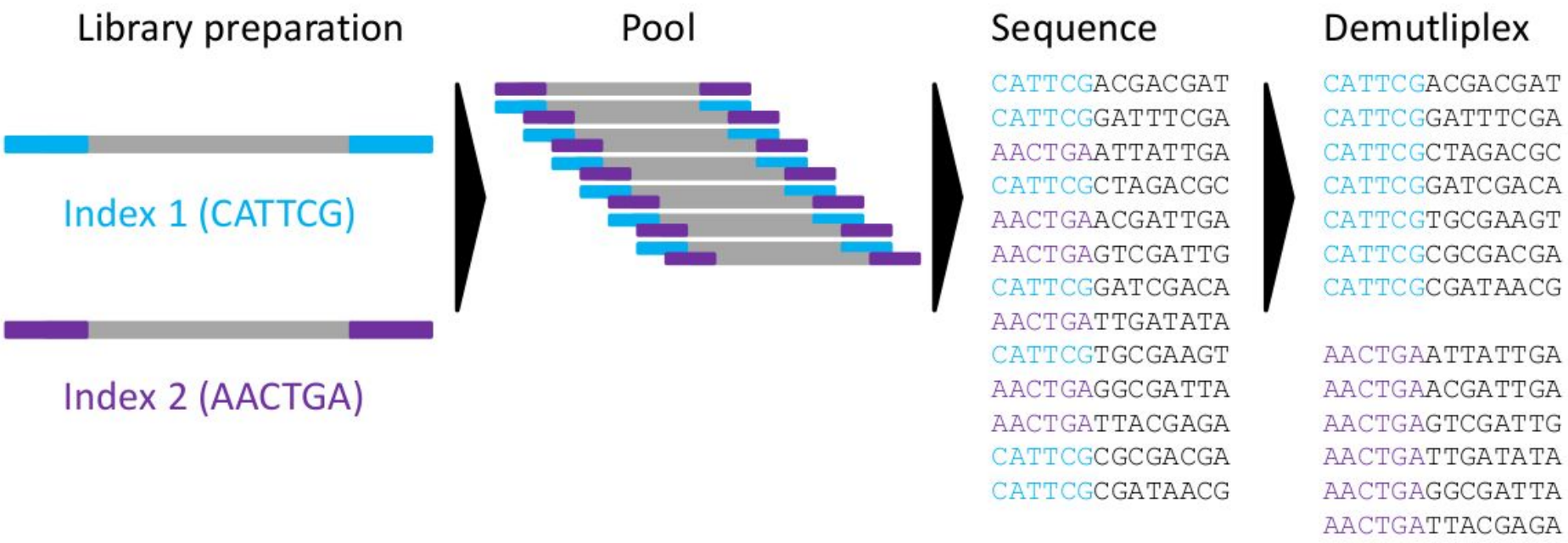
Illumina - flow cell layout



Illumina - Read ID nomenclature



Illumina - multiplexing



Illumina - sequencing modi

- Type:
  - SE = single end
  - PE = paired-end
  - MP = mate pair

- Read length:
  - 32nt, 50nt, 75nt, 100nt, 150nt, 250nt, 300nt

- Examples:
  - 2x250nt PE, 2x100nt MP, 1x100nt SE

Illumina - sequencing modi (single end, paired-end)

- Single end (SE):

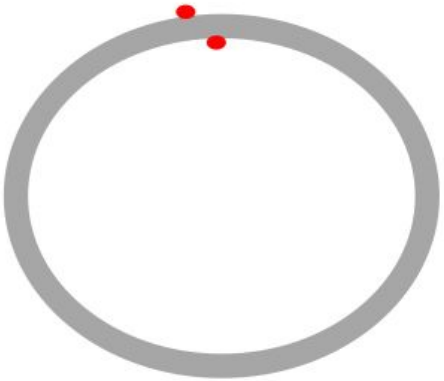
- Paired-end (PE):

Illumina - sequencing modi (mate pair)

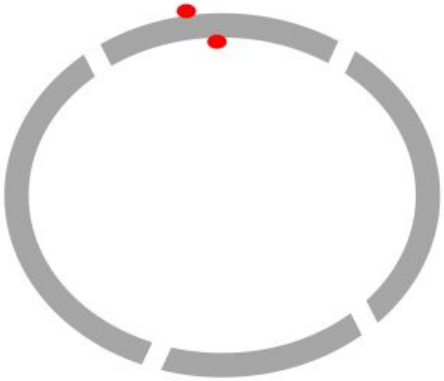
Fragmentation of DNA:

Adding biotin groups:

Circularization:



Fragmentation:



Enrichment of biotinylated fragments:

Sequencing as paired-end:

Result:

FASTQ

- Standard format for sequences with associated quality information

- Four lines per entry:
  - Header starts with @ (title + description)
  - Sequence
  - + (optional repetition of header)
  - Quality (phred encoded in ASCII character)

- Different versions exist that use different quality values offsets

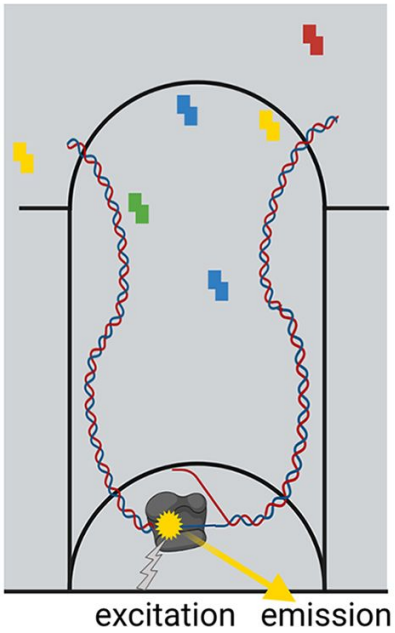
- Example:

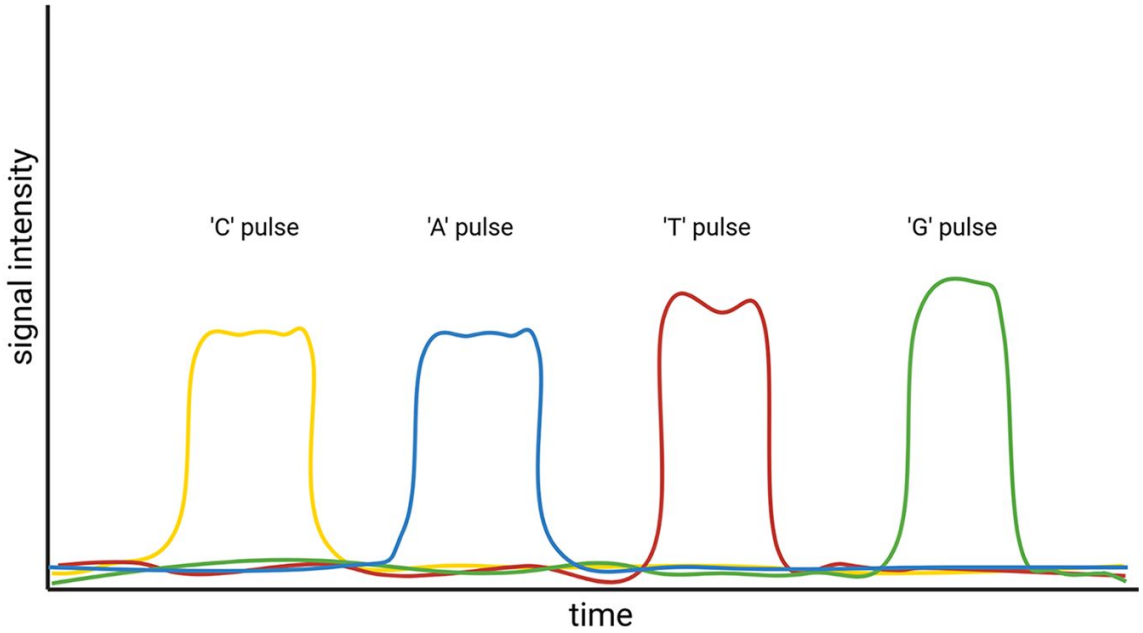
  @seq1
  ACGTACGTACGT
  +
  “”?CB””:DC”

PacBio

Pacific Biosciences (PacBio)

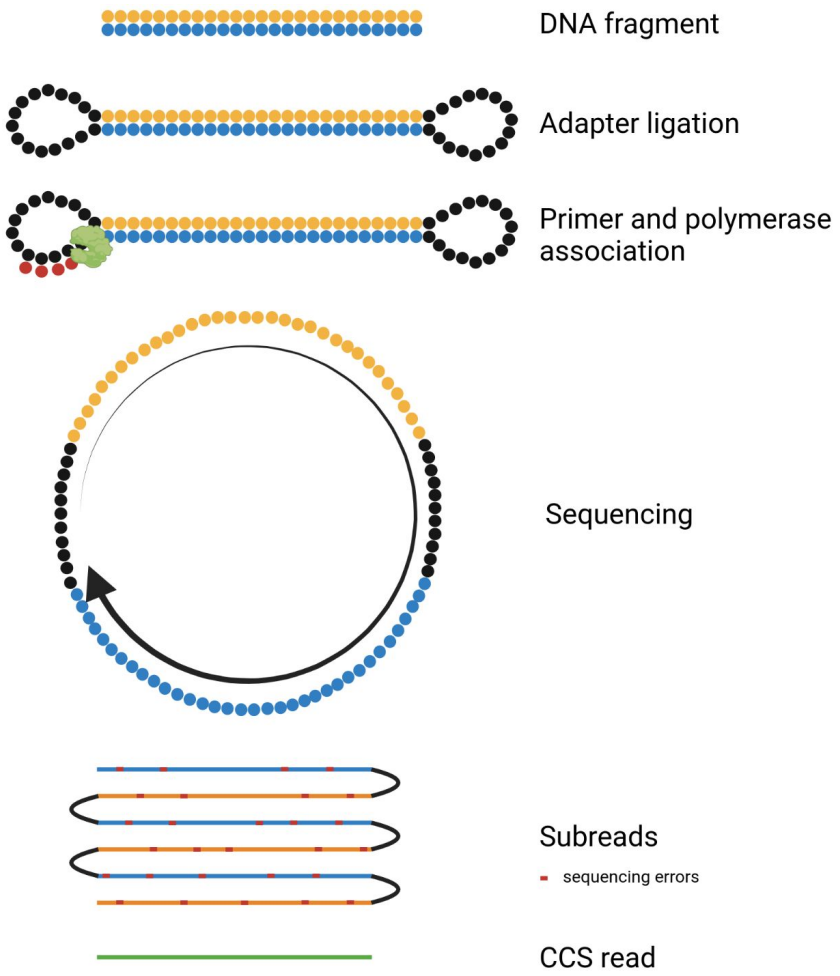
- Polymerase located at bottom of well
- ZMWs = Zero Mode Waveguide





Pucker et al., 2022: 10.1017/qpb.2021.18

PacBio - HiFi

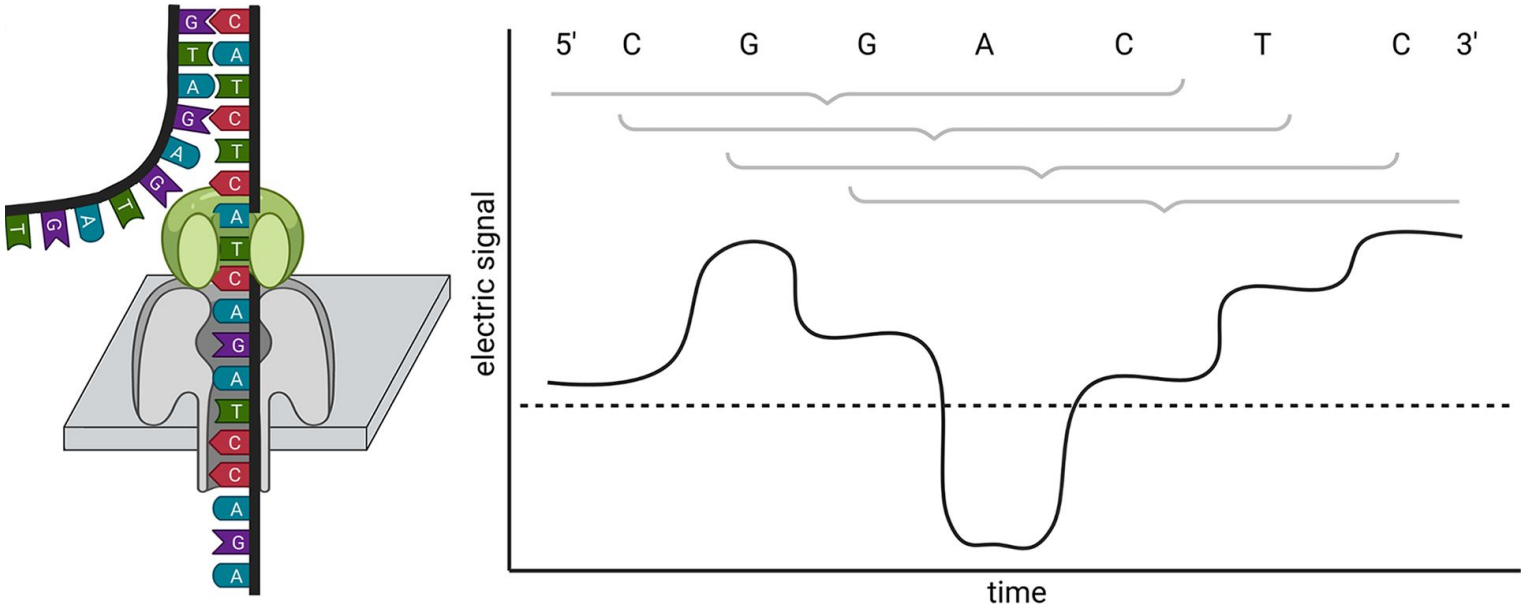

DNA fragment
Adapter ligation
Primer and polymerase association
Sequencing
Subreads
sequencing errors
CCS read

ONT

Oxford Nanopore Technologies (ONT)

Under development since the 1980s

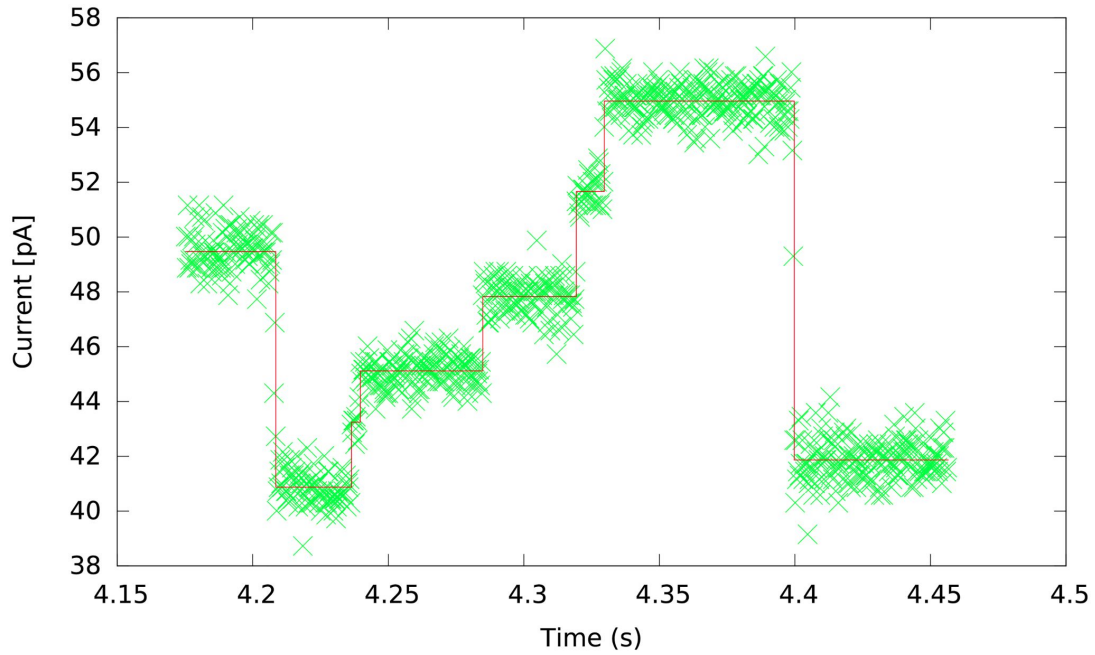
“Analysis of anything, anywhere by anyone” … not restricted to DNA sequencing



Pucker et al., 2022: 10.1017/qpb.2021.18

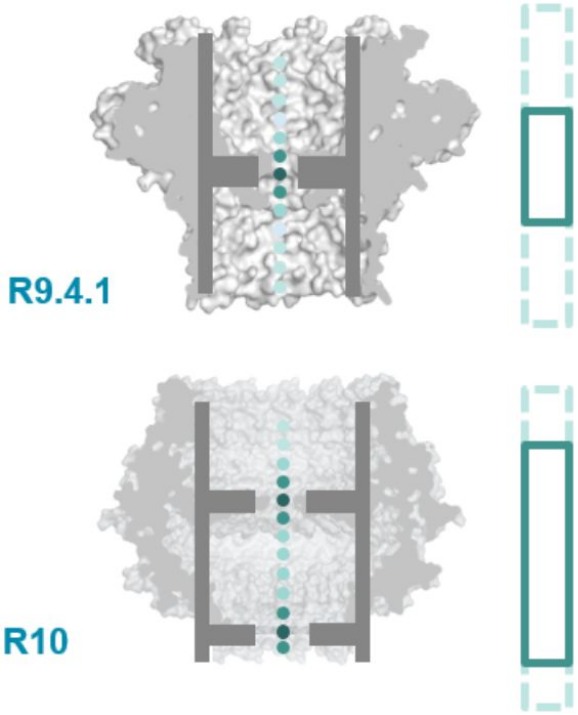
Basecalling

- Electric signal is converted into sequence information (bascalling)
- Algorithmic improvement lead to higher read accuracy
- Raw sequencing data (FAST5) need to be stored



Boza et al., 2017: 10.1371/journal.pone.0178751

Nanopore comparison



https://nanoporetech.com/about-us/news/r103-newest-nanopore-high-accuracy-nanopore-sequencing-now-available-store

ONT vs. PacBio

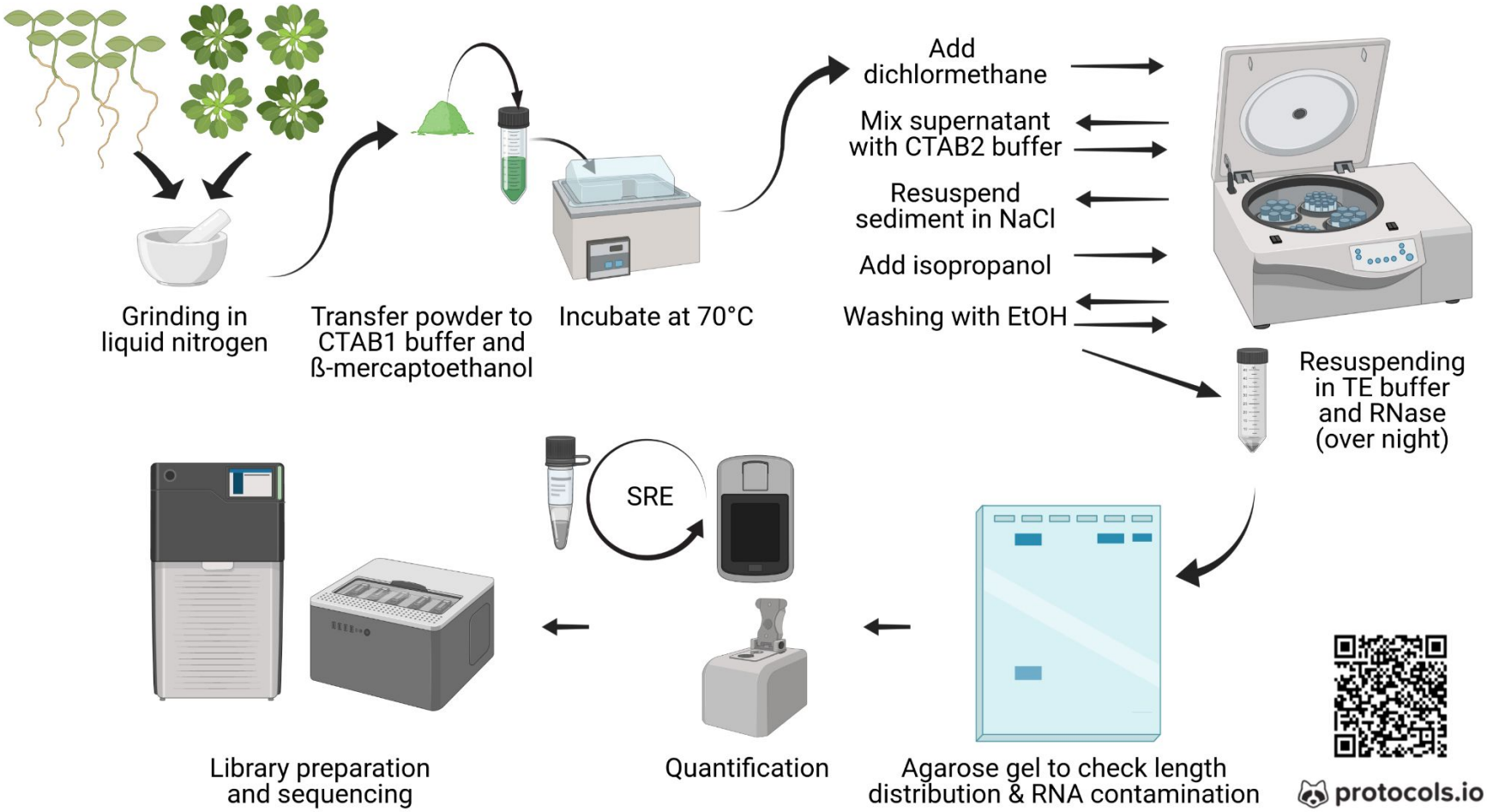
	ONT	PacBio (HiFi)
Maximal read length	DNA molecule size	25kb
Raw read accurracy	99% with Q20+	99.5%
DNA input	1 µg	3 µg
Instrument costs	$1000 (MinION)	High
Costs per genome	$3000 per Gbp	

ONT sequencing workflow

		task	consumed time	hands-on time	equipment	estimated costs of consumables	estimated costs of lab equipment
A		plant incubation in darkness	2-3d	1h			
B		non-destructive sampling	-	1h			
C		DNA extraction	1d	8h	waterbath, centrifuge	$50	$1000 $8000
D		quality control	1h	1h	NanoDrop, Qubit	$20	
E		short fragment depletion	2h	1h	centrifuge	$50	
F		quality control	1h	1h	NanoDrop, Qubit	$20	$5000 $5000
G		library preparation & sequencing	1-5d	4-16h	centrifuge, magnetic rack, sequencer	$3000	$250 $1000
H		basecalling	1d	1h	computer with GPU		$3000

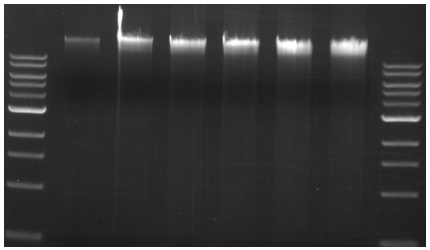
Pucker et al., 2022: 10.1017/qpb.2021.18

DNA extraction workflow



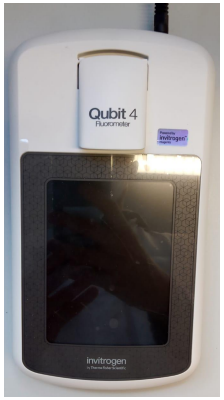
Quality control

- Agarose gel electrophoresis



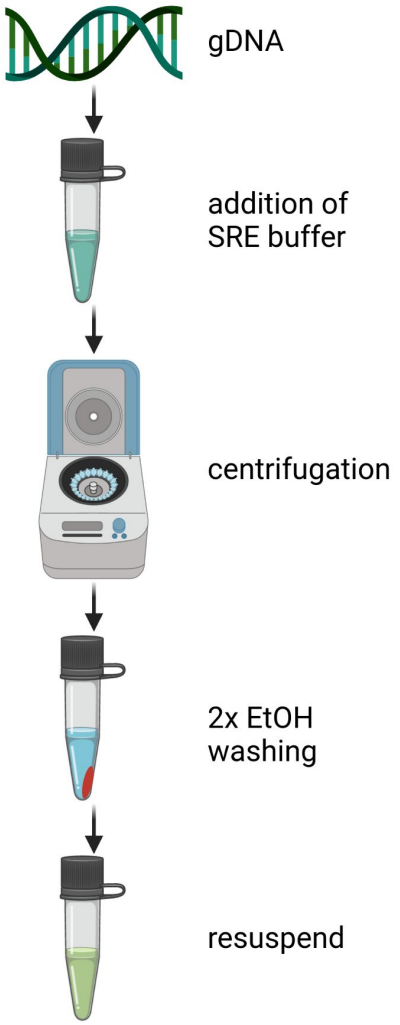
- Photometric measurement via NanoDrop

- Quantification with Qubit



Short Read Eliminator (SRE)

- Proprietary salt mix for DNA precipitation
- Removal of <10kb DNA fragments
- Depletion of <20kb DNA fragments
- ONT read length distribution can be substantially improved



DNA repair

- Repairing single strand DNA breaks
- Repairing DNA ends (3’-A overhang required for adapter ligation)

Library preparations

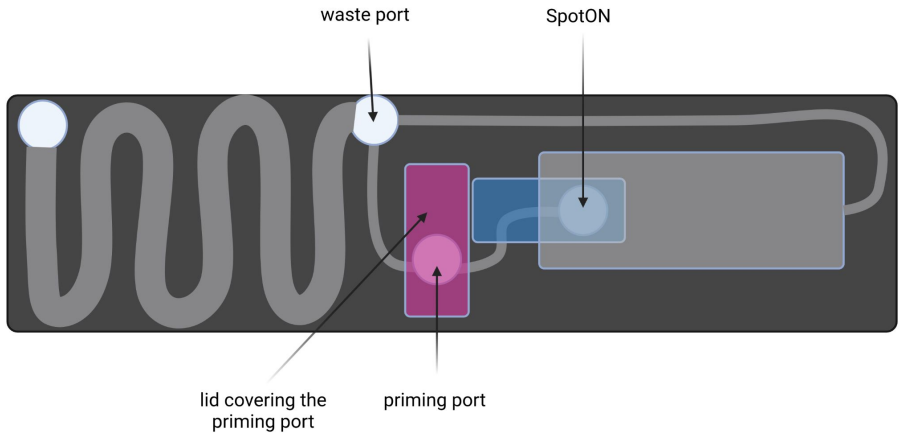
- Repaired DNA is subjected to library preparation
- Addition of adapters to DNA fragments
- Concentration of DNA can be quantified via Qubit measurement (optional)
  - Control step to ensure that library construction is working
- Purification of DNA with magnetic beads

Flow cell check

- Flow cells are delivered with storage buffer (green)
- Buffer allows technical check of flow cell (number of active pores is determined)
- Number of nanopores must be >800
- Replacement flow cells are provided if number of pores is lower

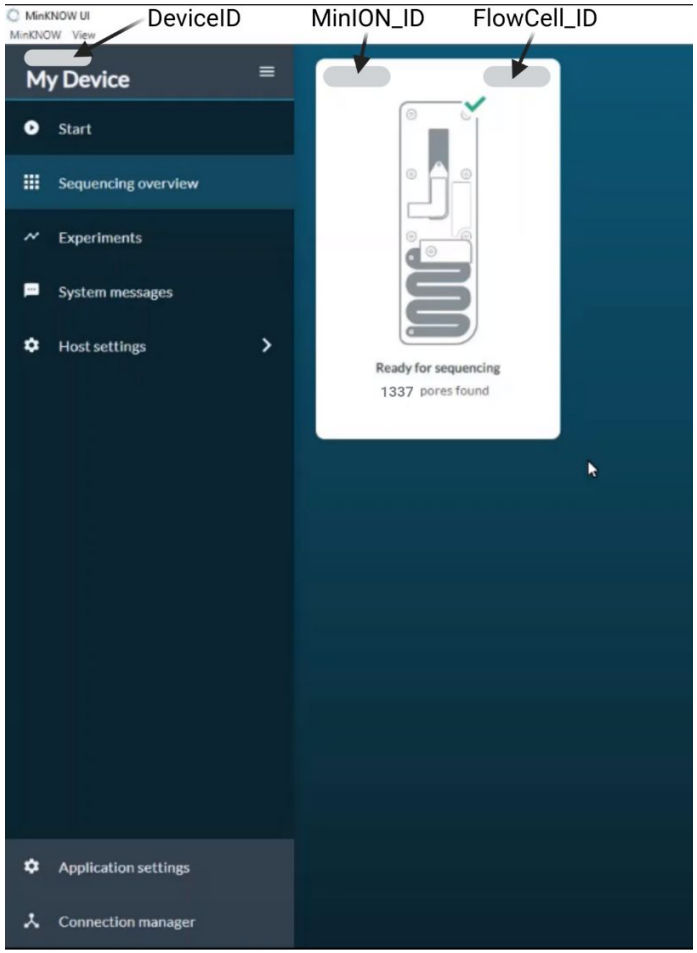
Loading flow cell

- Removal of storage buffer
- Priming of flow cell
- Introduction of air bubbles must be avoided!!!
- Fully open ports are crucial to inject solutions (avoid force)
- Video tutorial: https://www.youtube.com/watch?v=Pt-iaemrM88



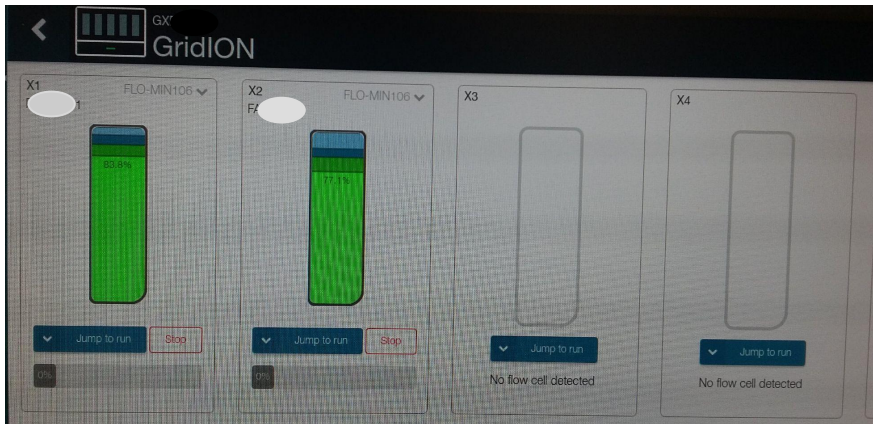
Starting sequencing

- Define the flow cell type (R9.4.1)
- Select parameters (default)
- Set the output location, run name, …
- Start the sequencing about 30 minutes after loading the flow cell
  - This allows the DNA to get into contact with the nanopores



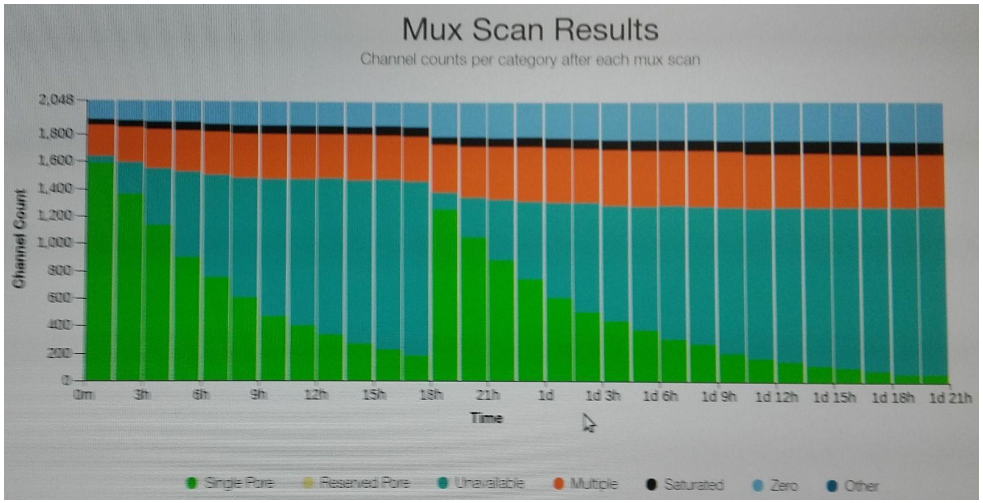
Monitor sequencing

- Number of active nanopores can be monitored in real time
- Output is estimated in real time
- Read length distribution can be assessed
- Speed of the sequencing can be monitored
- Quality of the reads is displayed



Stop & wash

- Sequencing is stopped once the number of active pores gets low
- Washing with DNase to free blocked nanopores
- Flow cells are regenerated and can be re-used
- Process can be repeated multiple times (3-5x)



Summary ONT sequencing workflow

		task	consumed time	hands-on time	equipment	estimated costs of consumables	estimated costs of lab equipment
A		plant incubation in darkness	2-3d	1h			
B		non-destructive sampling	-	1h			
C		DNA extraction	1d	8h	waterbath, centrifuge	$50	$1000 $8000
D		quality control	1h	1h	NanoDrop, Qubit	$20	
E		short fragment depletion	2h	1h	centrifuge	$50	
F		quality control	1h	1h	NanoDrop, Qubit	$20	$5000 $5000
G		library preparation & sequencing	1-5d	4-16h	centrifuge, magnetic rack, sequencer	$3000	$250 $1000
H		basecalling	1d	1h	computer with GPU		$3000

Summary sequencing technologies

- Generation 1:
  - Sanger sequencing
- Generation 2 (massive parallel sequencing):
  - Illumina sequencing
- Generation 3 (long reads):
  - Pacific Biosciences (PacBio)
  - Oxford Nanopore Technologies (ONT)

Time for questions!

Questions

1. Which sequencing technologies do you know?
2. What is the structure of a FASTA file?
3. What is a Phred score?
4. What are differences between Sanger sequencing and Illumina sequencing?
5. How does mate pair sequencing work?
6. What is the structure of a FASTQ file?
7. How does PacBio sequencing work?
8. How does ONT sequencing work?
9. What are the important steps of an ONT sequencing workflow?
10. What are important differences of ONT vs. PacBio sequencing?
11. Which parameters can be monitored during ONT sequencing?