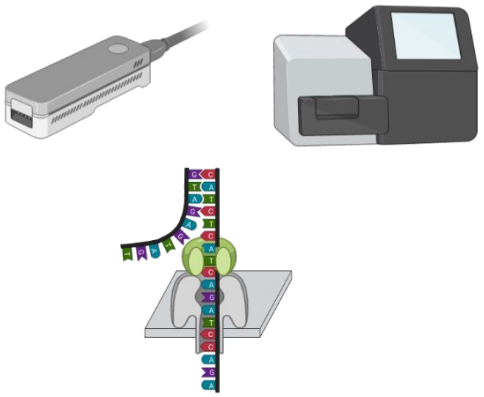
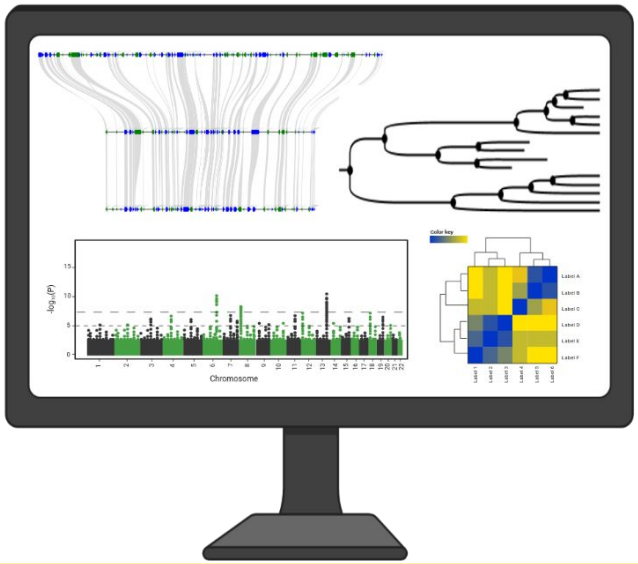
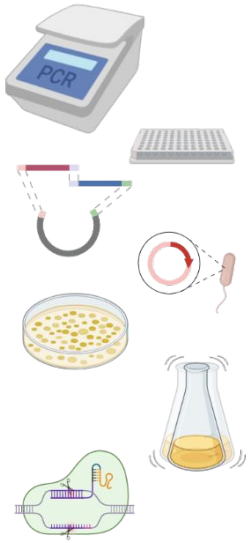
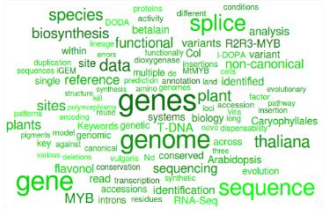
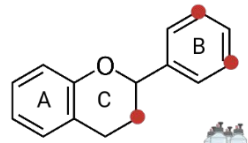
Technische Universität Braunschweig

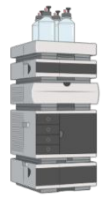












Plant Biotechnology and Bioinformatics

Sequencing

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

Availability of slides

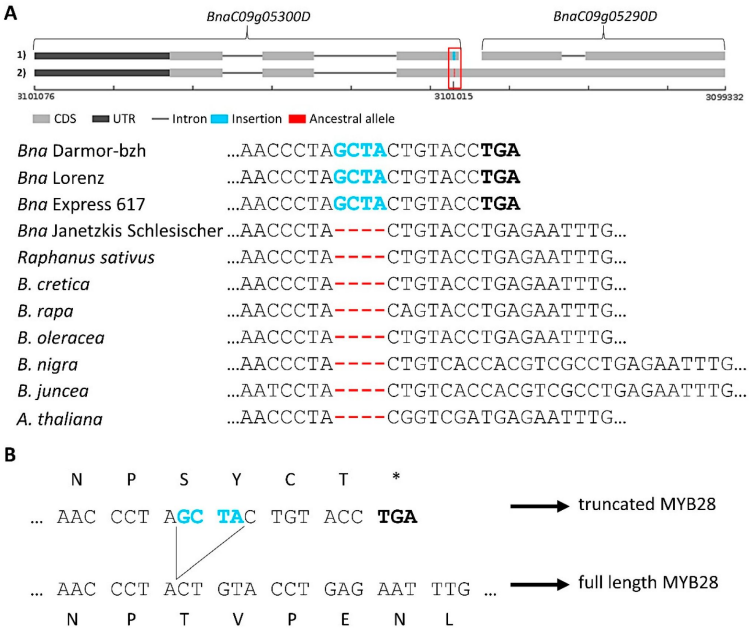
- All materials are freely available (CC BY) - after the lectures:
  - StudIP: LMChemBSc12
  - GitHub: https://github.com/bpucker/teaching

- Questions: Feel free to ask at any time

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

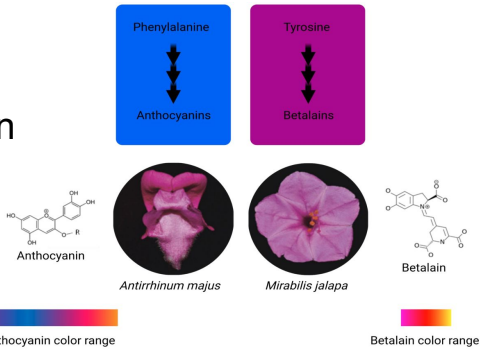
Crop genomics

Glucosinolate content in rapeseed is controlled by the transcription factor MYB28



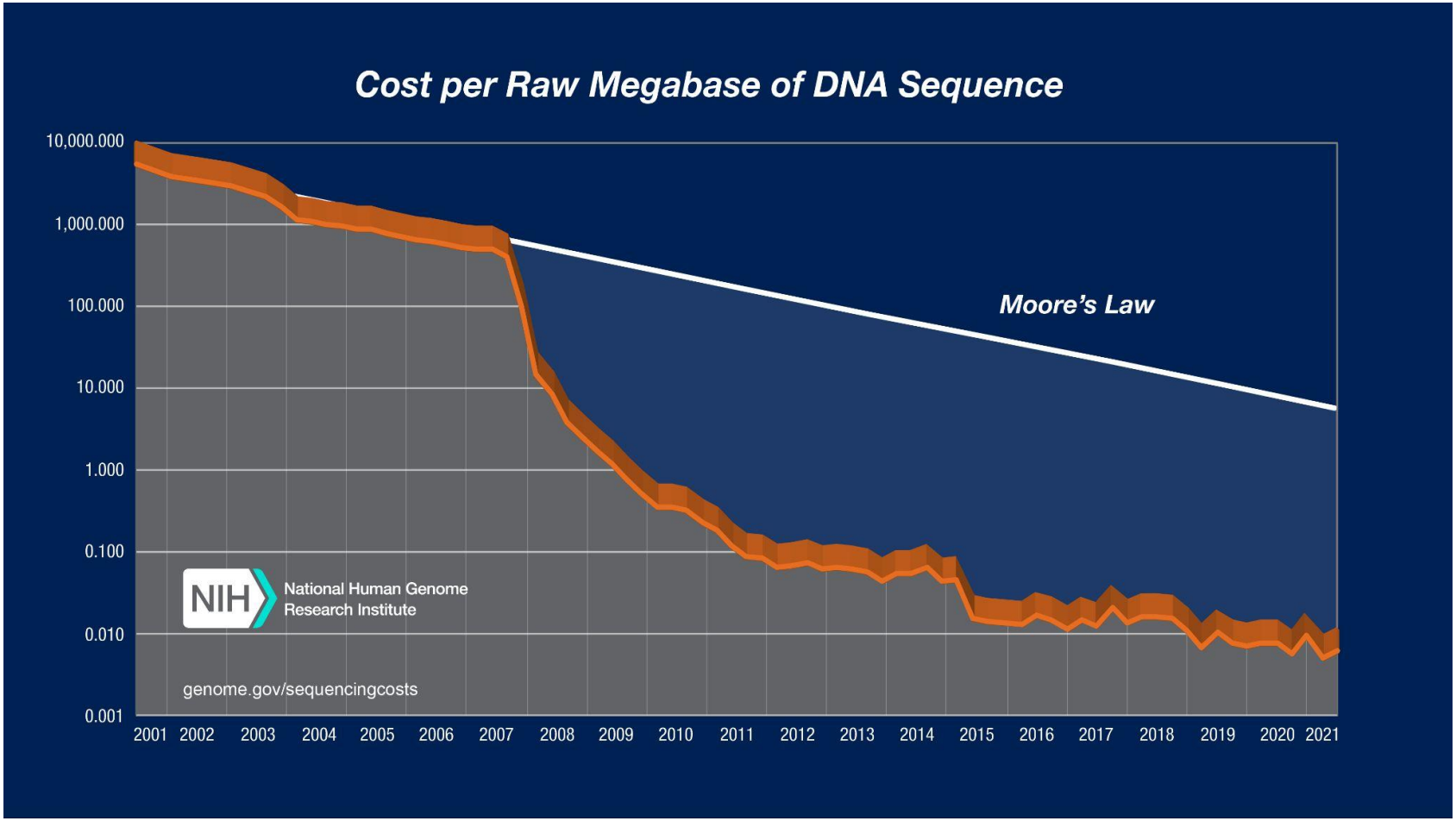
10.3390/genes13071131

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

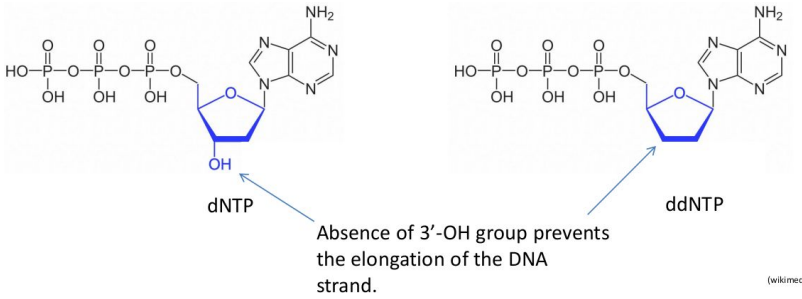
Which sequencing methods do you know?

Overview of sequencing technologies

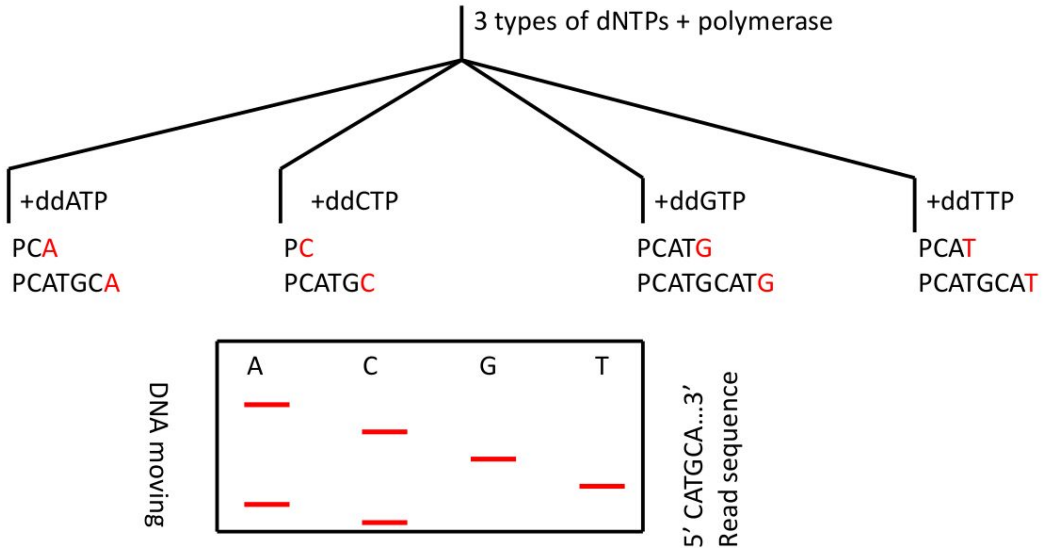
- Generation 1:
  - **Sanger sequencing**
  - Max-Gilbert sequencing
- Generation 2 (massive parallel sequencing):
  - 454 pyrosequencing
  - Solexa/**Illumina sequencing**
  - SOLID
  - Ion Torrent
  - BGI-seq
  - Synthetic long reads
- Generation 3 (long reads):
  - **Pacific Biosciences (PacBio)**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

Sanger sequencing

Concept of Sanger sequencing

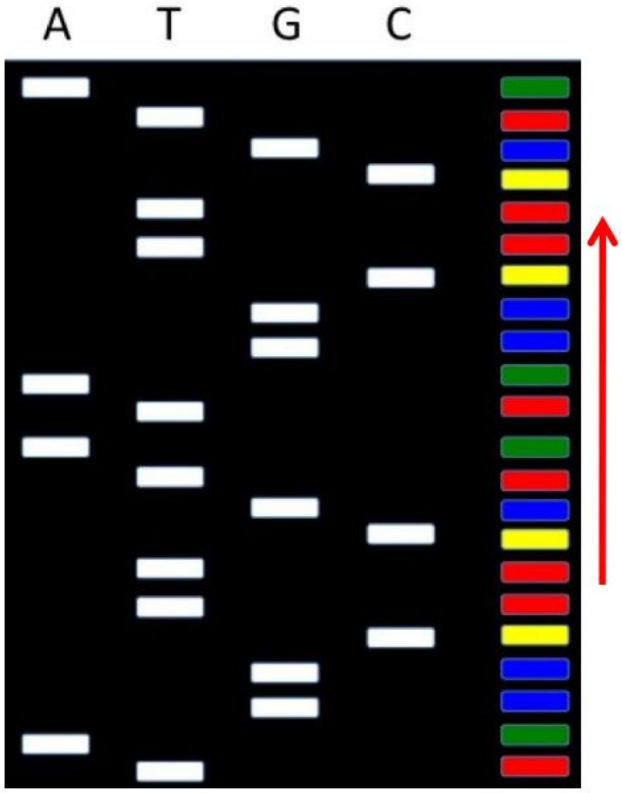


Primer (P): 5' –TGCATGGCATGATGCATG-3'
Template: 3' –ACGTACCGTACTACGTACGTACGTACGTCTAGGT-5'



Sanger sequencing - original version

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



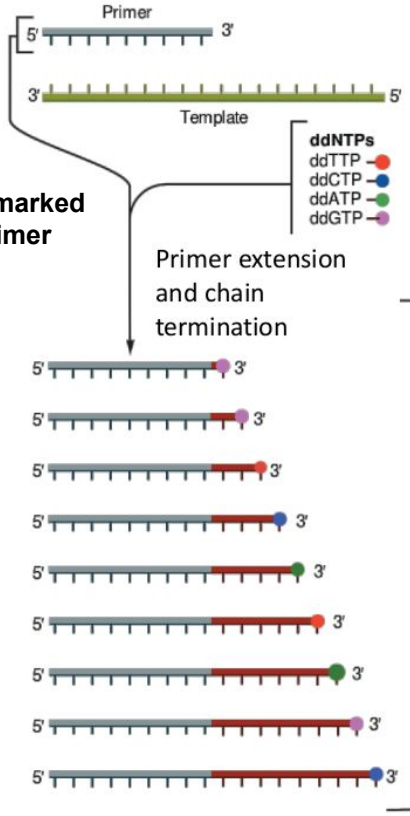
(modified from wikimedia.org)

Sanger sequencing - today

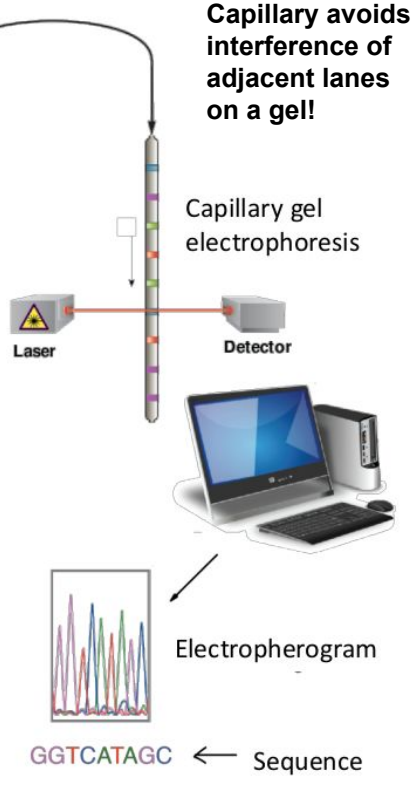
Only one reaction!

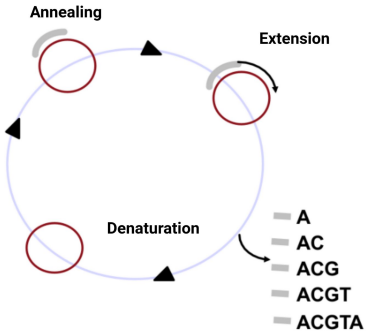
ddNTPs are marked instead of primer

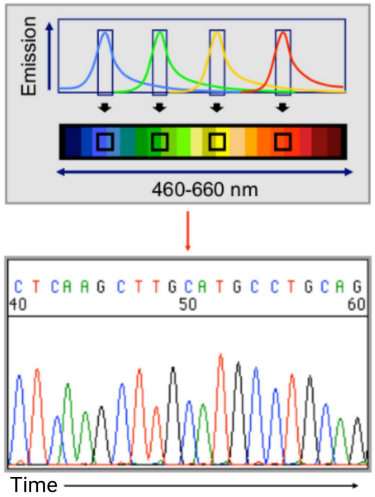
Low input required due to cycle sequencing



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

- Phred score indicates the per base quality in an efficient way
- Single character is used to show the quality
- Formula: $Q = -10 \log_{10} P$ $P = 10^{-Q/10}$

Phred quality score	Probability of incorrect base call	Base call 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%

ASCII code

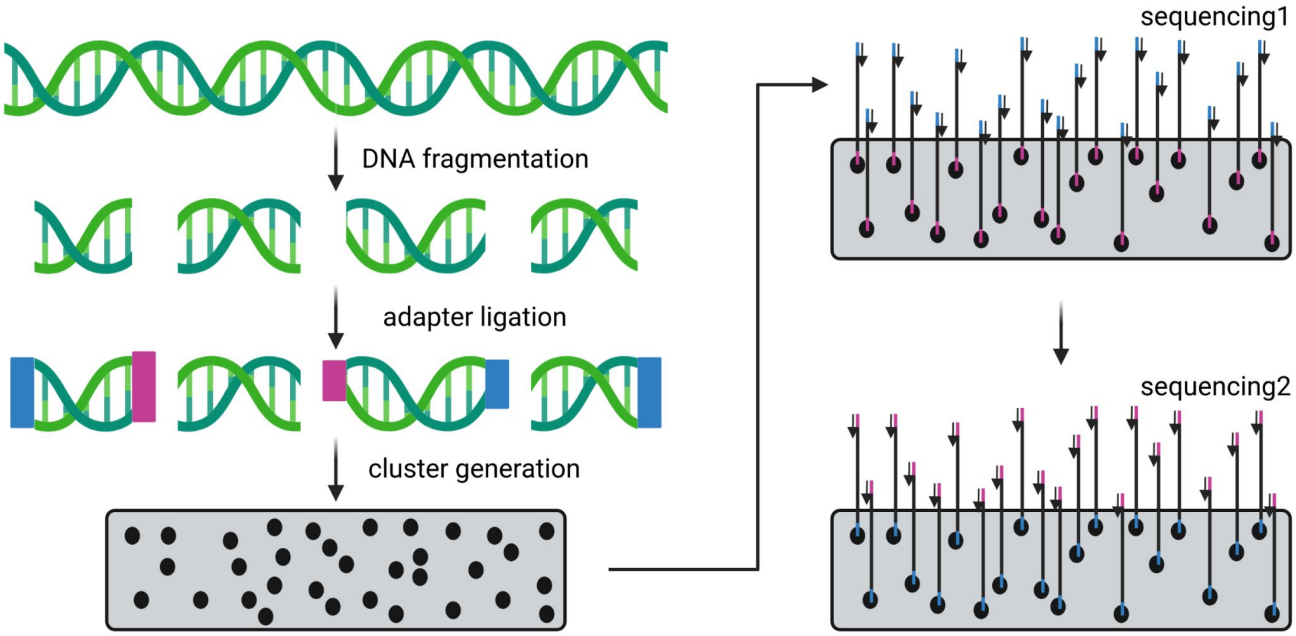
- Phred score encoded with ASCII
- Phred32, phred64 (different offset values to avoid special characters)

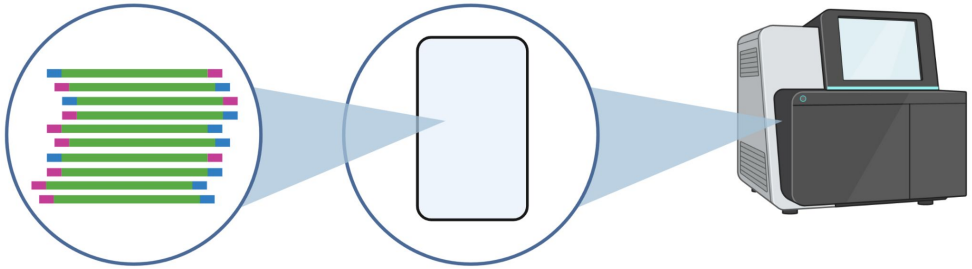
Binary	Oct	Dec	Hex	Glyph		
				1963	1965	1967
010 0000	040	32	20	space		
010 0001	041	33	21	!		
010 0010	042	34	22	"		
010 0011	043	35	23	#		
010 0100	044	36	24	$		
010 0101	045	37	25	%		
010 0110	046	38	26	&		
010 0111	047	39	27	'		
010 1000	050	40	28	(		
010 1001	051	41	29	)		
010 1010	052	42	2A	*		
010 1011	053	43	2B	+		
010 1100	054	44	2C	,		
010 1101	055	45	2D	-		
010 1110	056	46	2E	.		
010 1111	057	47	2F	/		
011 0000	060	48	30	0		
011 0001	061	49	31	1		
011 0010	062	50	32	2		
011 0011	063	51	33	3		
011 0100	064	52	34	4		
011 0101	065	53	35	5		
011 0110	066	54	36	6		
011 0111	067	55	37	7		
011 1000	070	56	38	8		
011 1001	071	57	39	9		
011 1010	072	58	3A	:		
011 1011	073	59	3B	;		
011 1100	074	60	3C	<		
011 1101	075	61	3D	=		
011 1110	076	62	3E	>		
011 1111	077	63	3F	?		
100 0000	100	64	40	@	`	@

100 0000	100	64	40	@	`	@
100 0001	101	65	41	A		
100 0010	102	66	42	B		
100 0011	103	67	43	C		
100 0100	104	68	44	D		
100 0101	105	69	45	E		
100 0110	106	70	46	F		
100 0111	107	71	47	G		
100 1000	110	72	48	H		
100 1001	111	73	49	I		
100 1010	112	74	4A	J		
100 1011	113	75	4B	K		
100 1100	114	76	4C	L		
100 1101	115	77	4D	M		
100 1110	116	78	4E	N		
100 1111	117	79	4F	O		
101 0000	120	80	50	P		
101 0001	121	81	51	Q		
101 0010	122	82	52	R		
101 0011	123	83	53	S		
101 0100	124	84	54	T		
101 0101	125	85	55	U		
101 0110	126	86	56	V		
101 0111	127	87	57	W		
101 1000	130	88	58	X		
101 1001	131	89	59	Y		
101 1010	132	90	5A	Z		
101 1011	133	91	5B	[		
101 1100	134	92	5C	\	~	\
101 1101	135	93	5D	]		
101 1110	136	94	5E	↑	^	
101 1111	137	95	5F	←	_	
110 0000	140	96	60		@	`
110 0001	141	97	61		a	
110 0010	142	98	62		b	
110 0011	143	99	63		c	
110 0100	144	100	64		d	
110 0101	145	101	65		e	
110 0110	146	102	66		f	
110 0111	147	103	67		g	
110 1000	150	104	68		h	

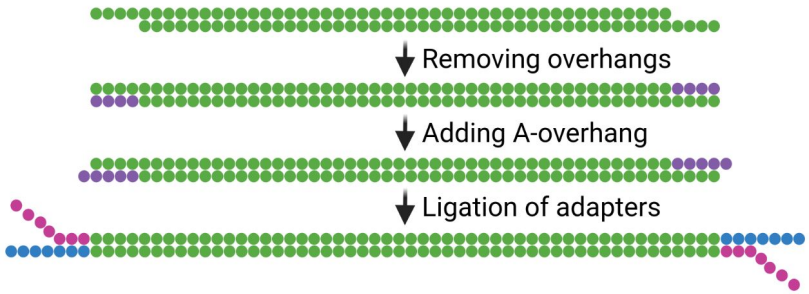
Illumina sequencing

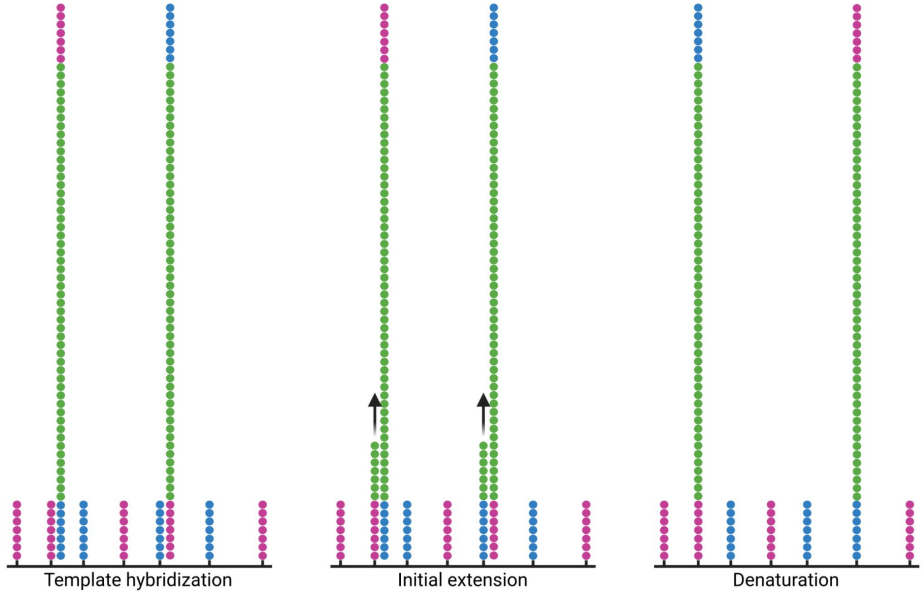
Illumina - sequencing overview



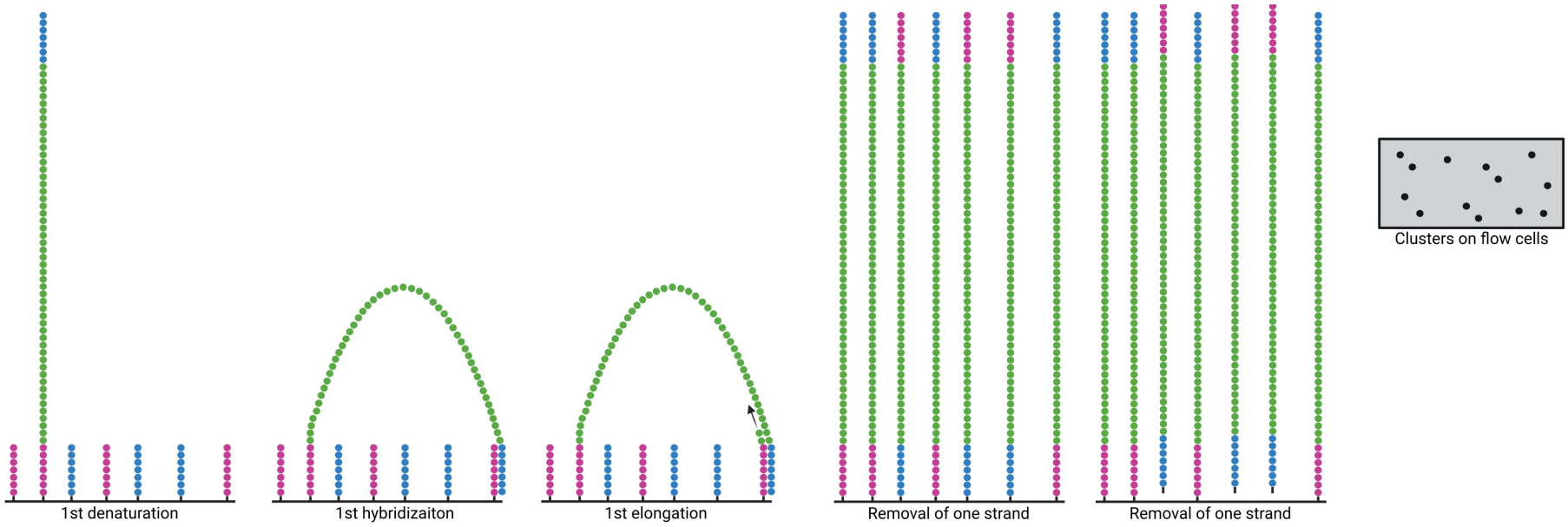


Illumina - sequencing 2

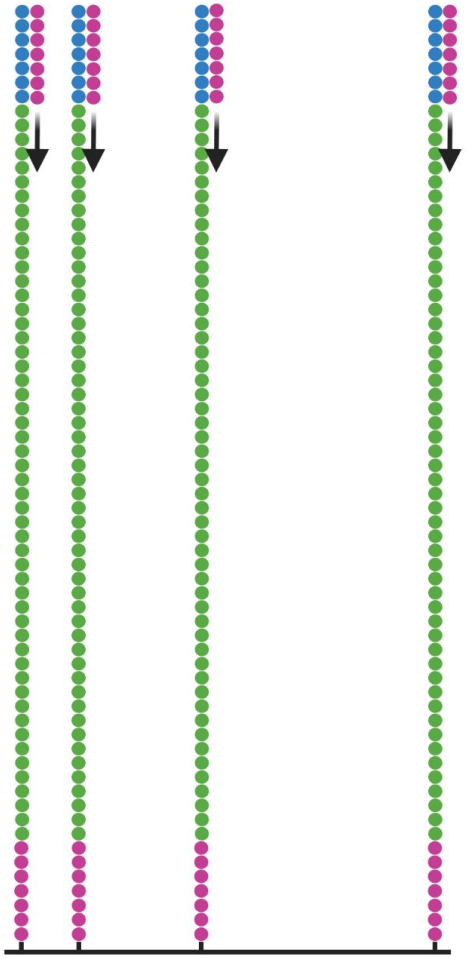




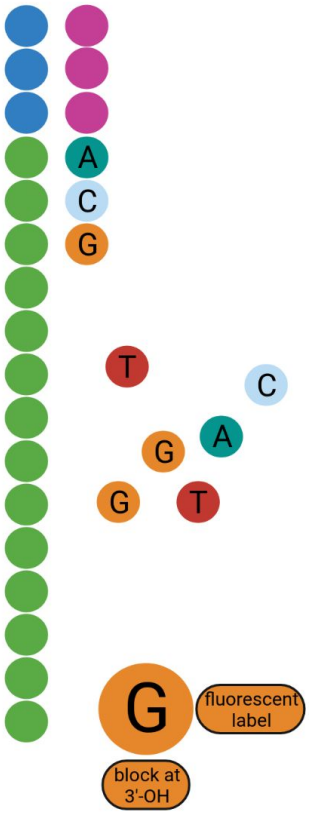
Illumina - sequencing 3



Illumina - sequencing 4

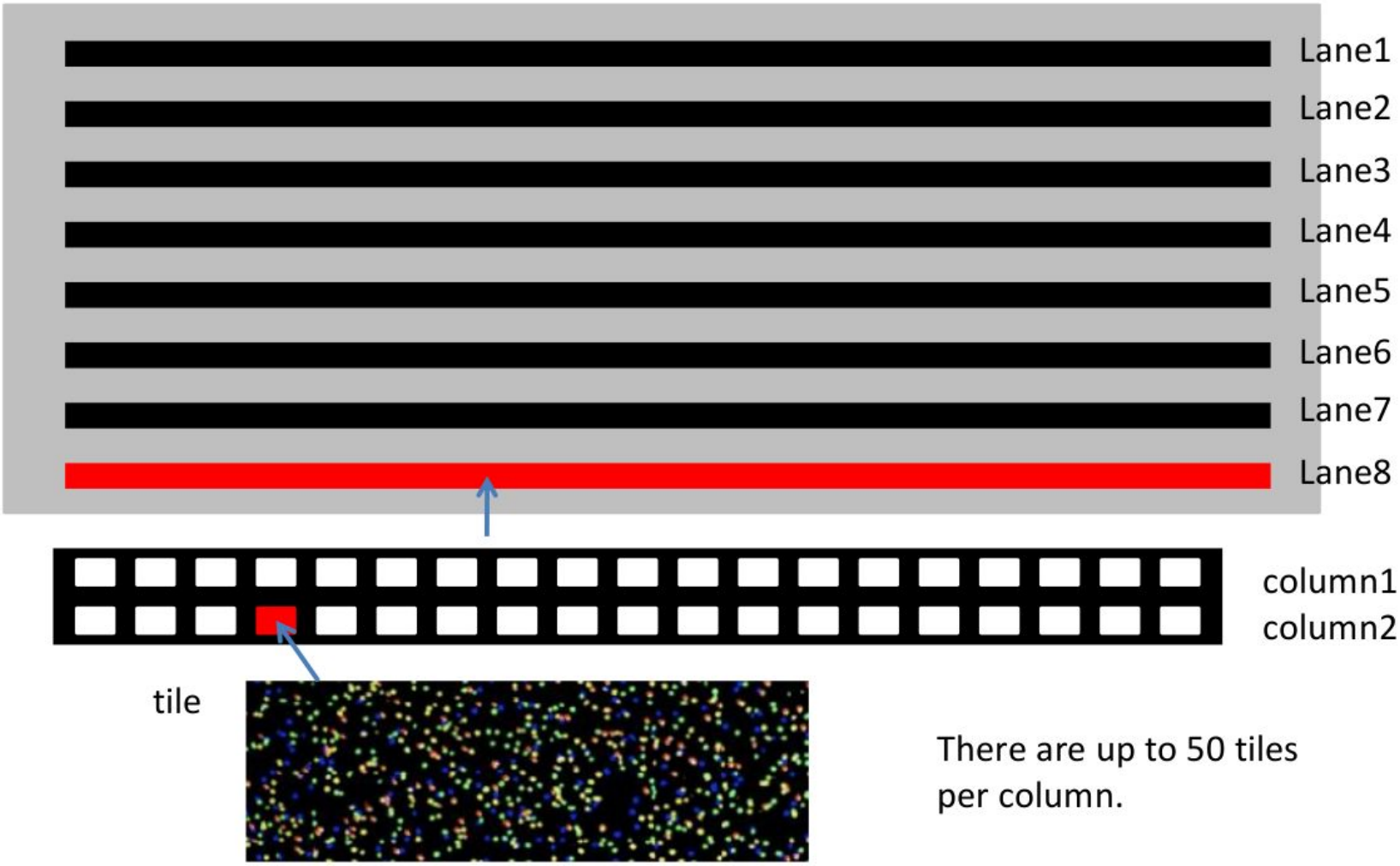


Sequencing

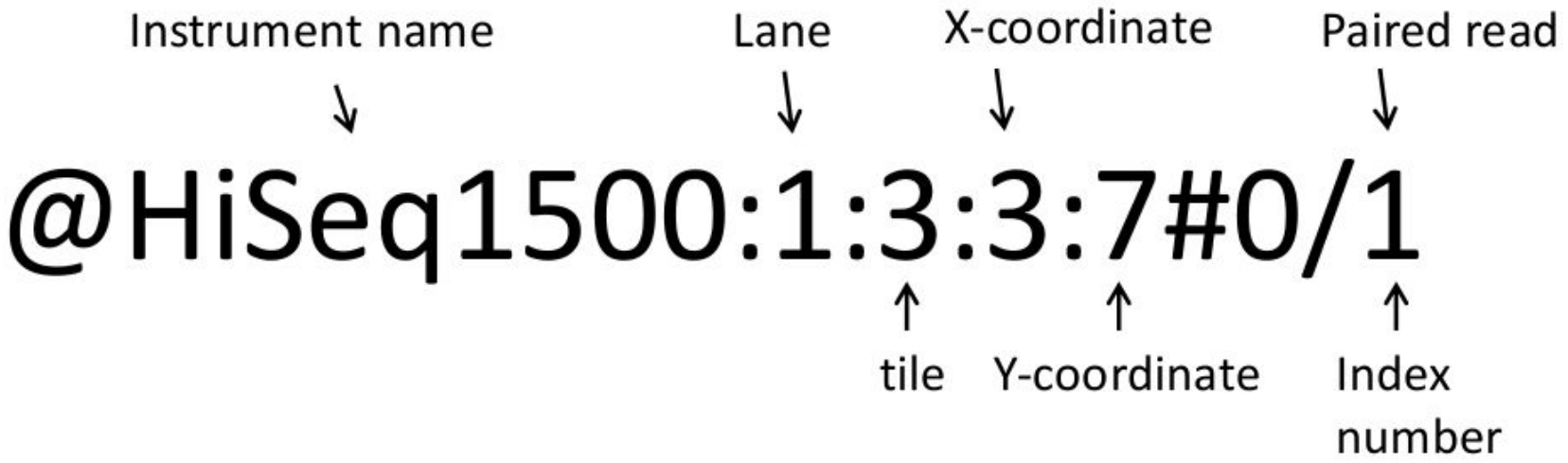


Sequencing chemistry

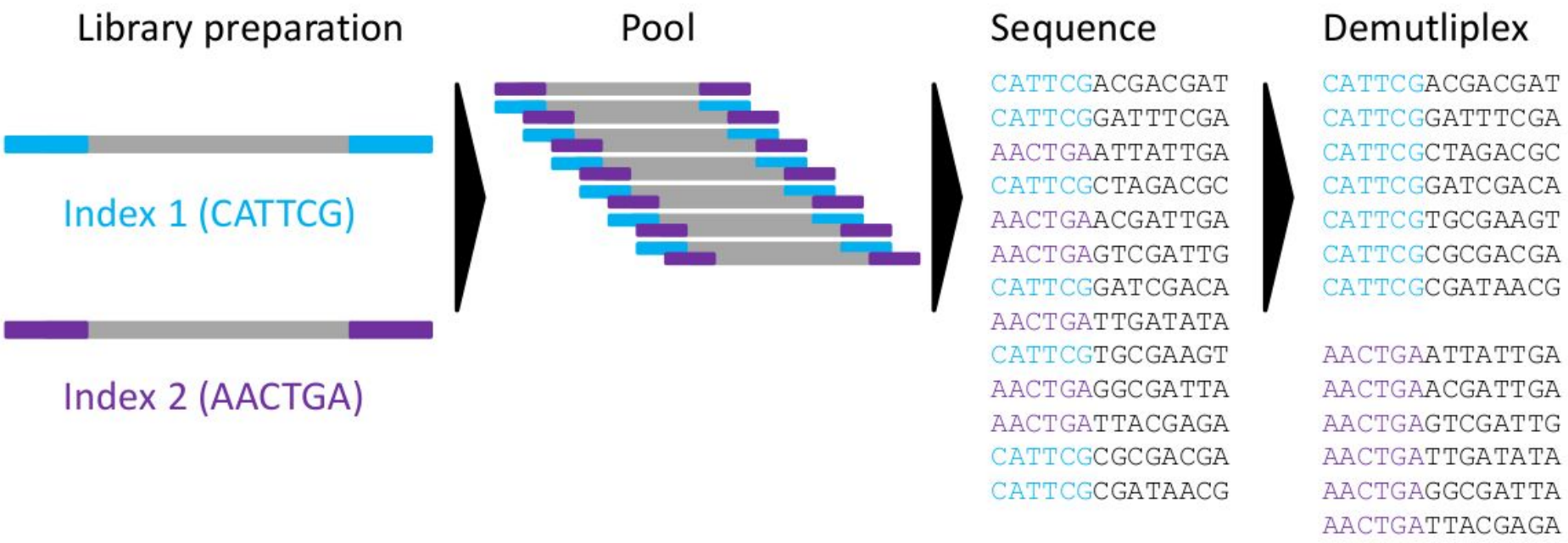
Illumina - flow cell layout



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):

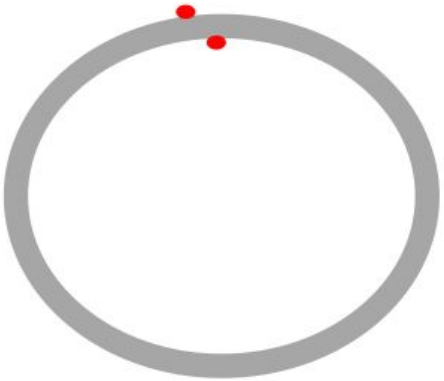
- Paired-end (PE):

Illumina - sequencing modi (mate pair)

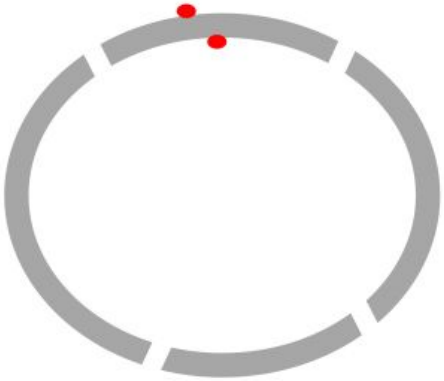
Fragmentation of DNA:

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

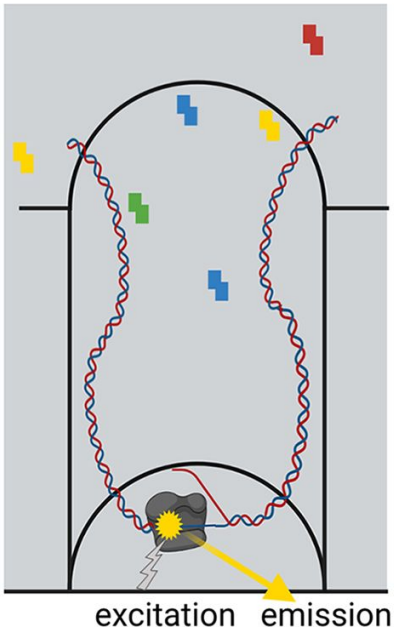
- Example:

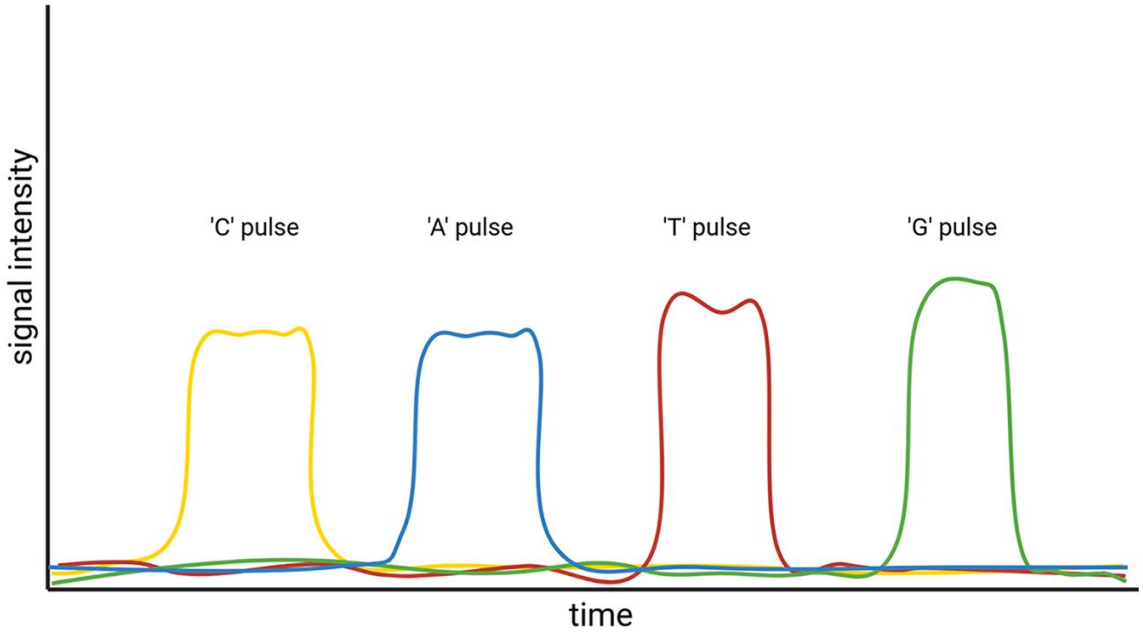
```
@seq1
ACGTACGTACGT
+
“”?CB””:DC”:
```

PacBio

Pacific Biosciences (PacBio)

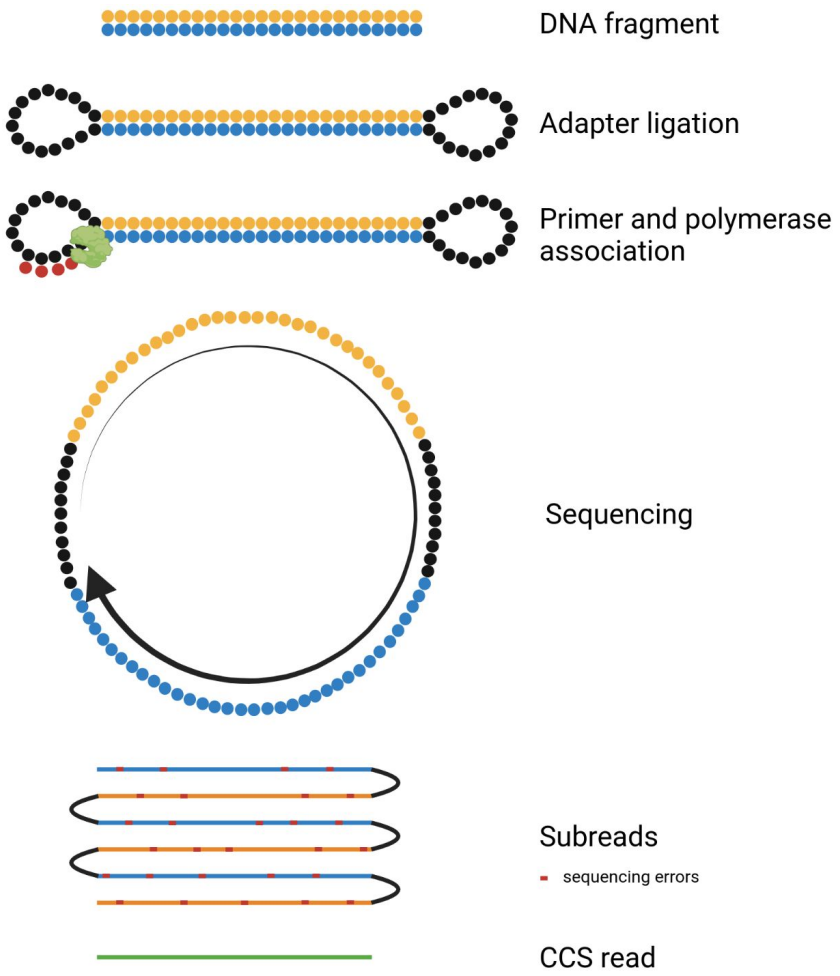
- Polymerase located at bottom of well
- ZMWs = Zero Mode Waveguide





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

PacBio - HiFi

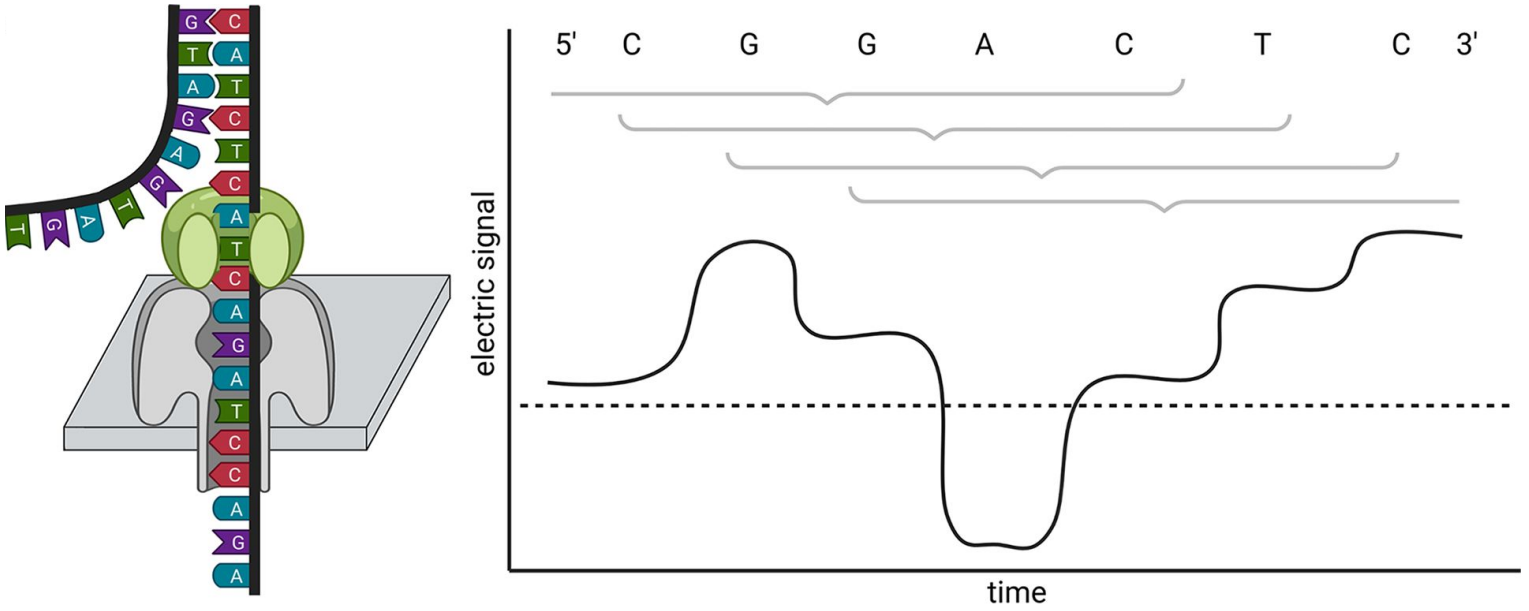

DNA fragment
Adapter ligation
Primer and polymerase association
Sequencing
Subreads
sequencing errors
CCS read

ONT

Oxford Nanopore Technologies (ONT)

Under development since the 1980s

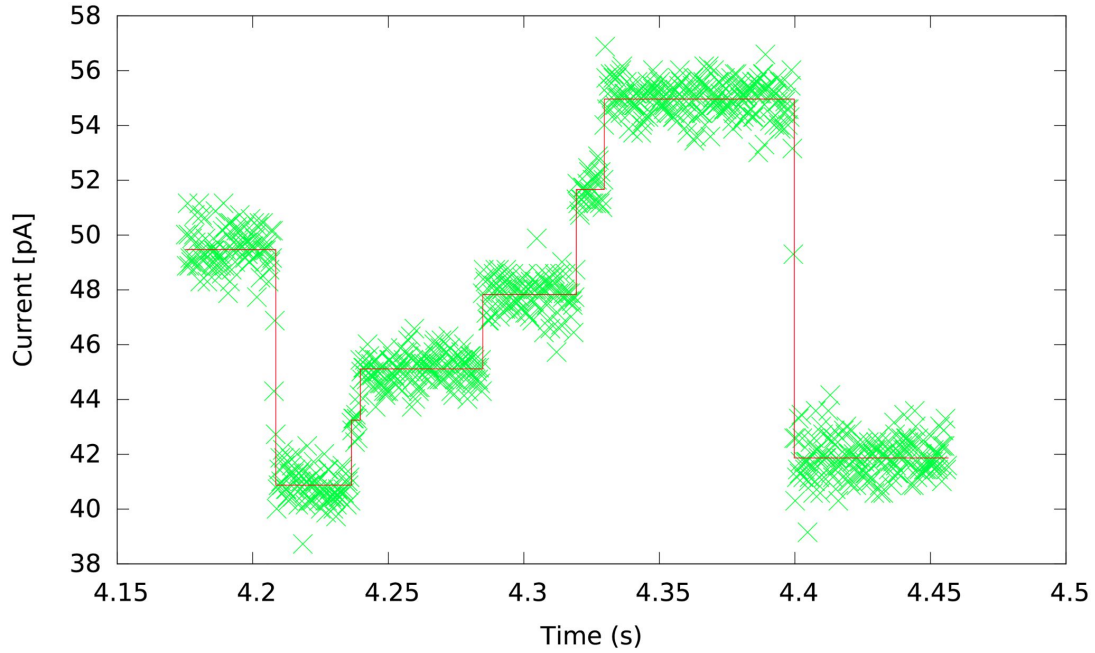
“Analysis of anything, anywhere by anyone” … not restricted to DNA sequencing



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

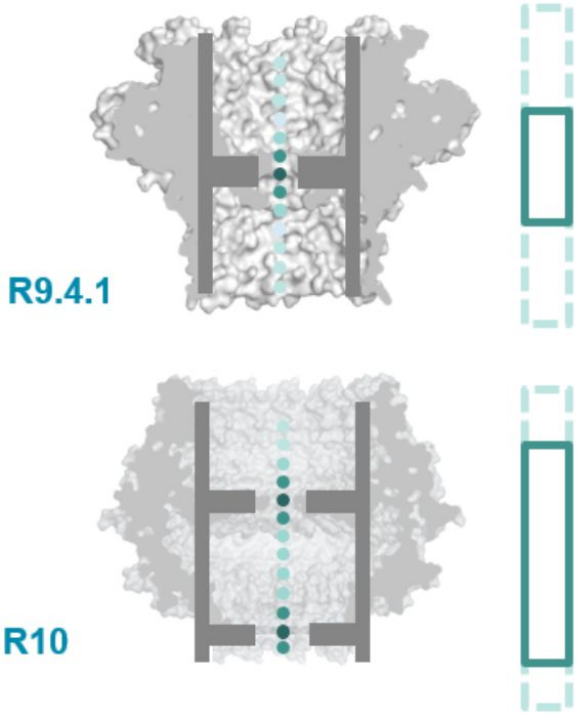
Basecalling

- 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

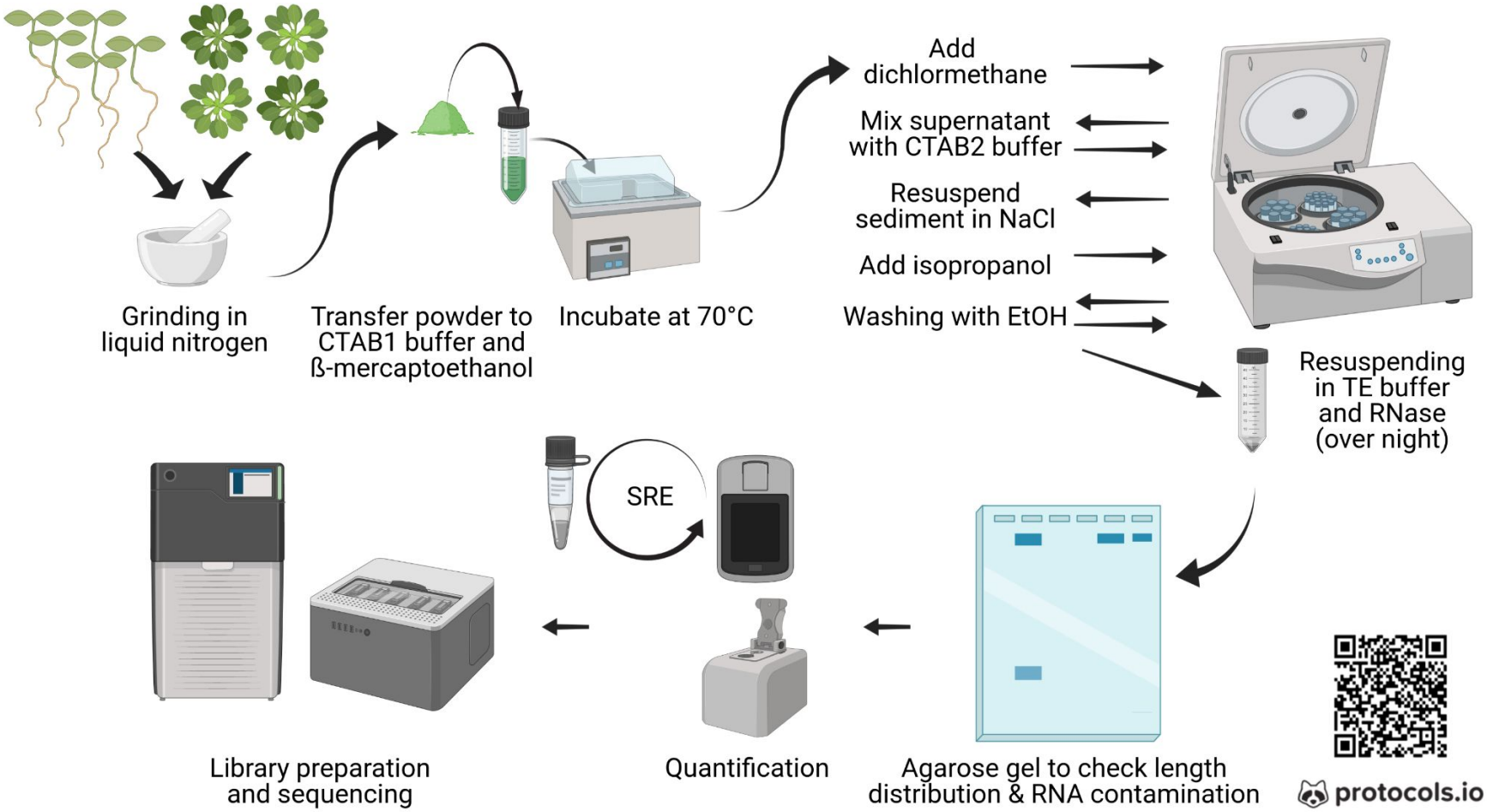
	ONT	PacBio (HiFi)
Maximal read length	DNA molecule size	25kb
Raw read accurracy	99% with Q20+	99.5%
DNA input	1 µg	3 µg
Instrument costs	$1000 (MinION)	High
Costs per genome	$3000 per Gbp	

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

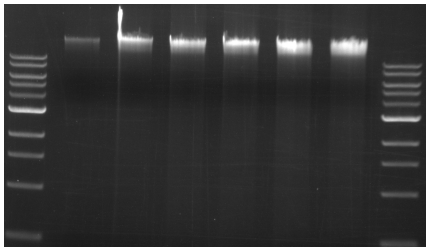
DNA extraction workflow



10.17504/protocols.io.bcvyiw7w

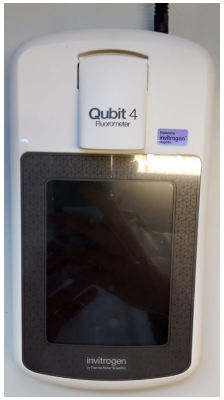
Quality control

- Agarose gel electrophoresis



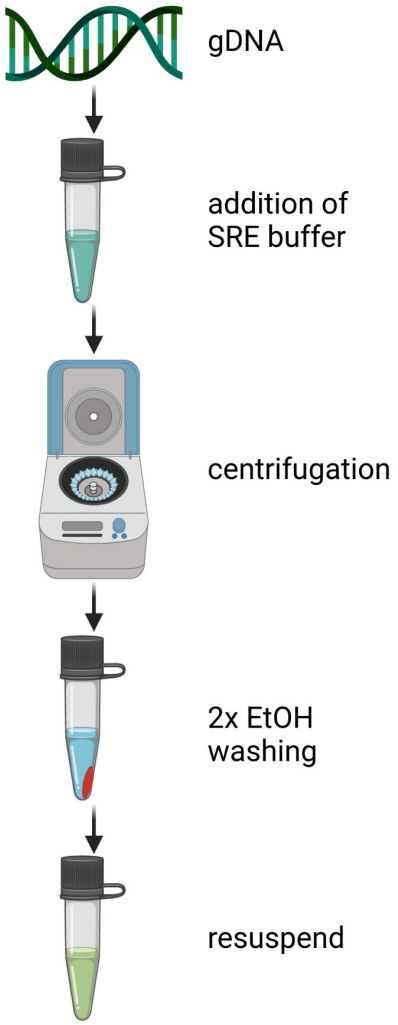
- Photometric measurement via NanoDrop

- 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

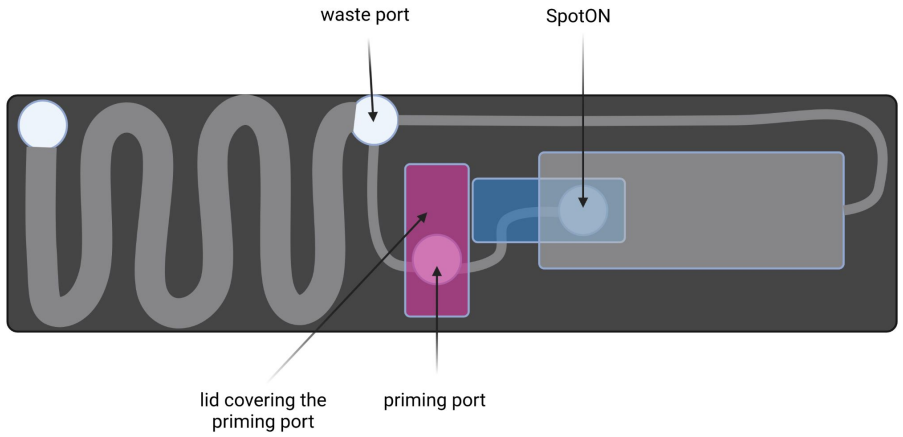
- Repaired DNA is subjected to library preparation
- Addition of adapters to DNA fragments
- Concentration of DNA can be quantified via Qubit measurement (optional)
  - Control step to ensure that library construction is working
- Purification of DNA with magnetic beads

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

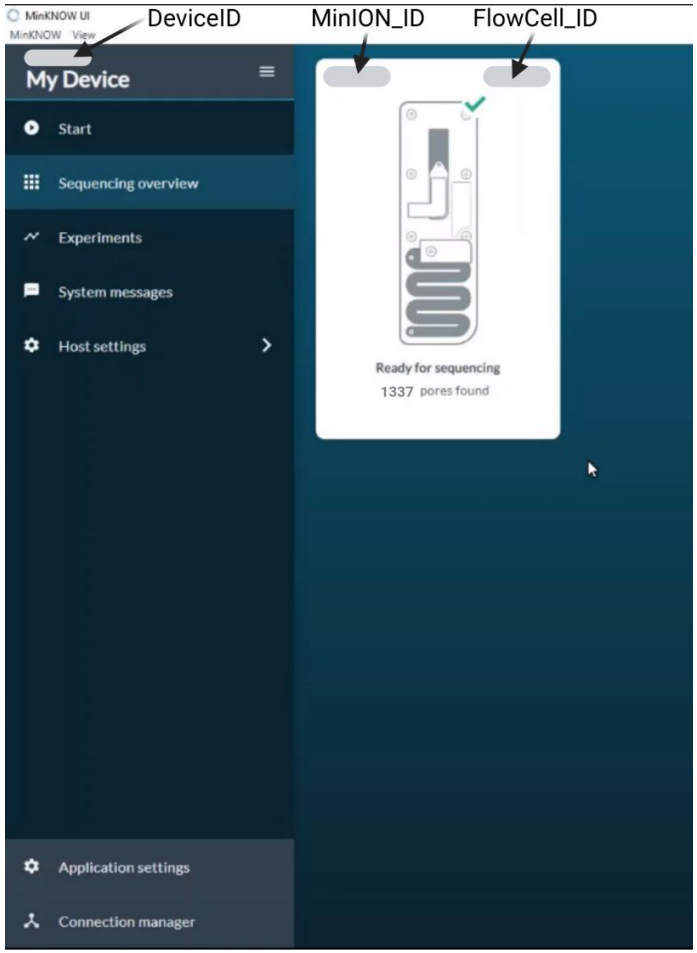
Loading flow cell

- 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



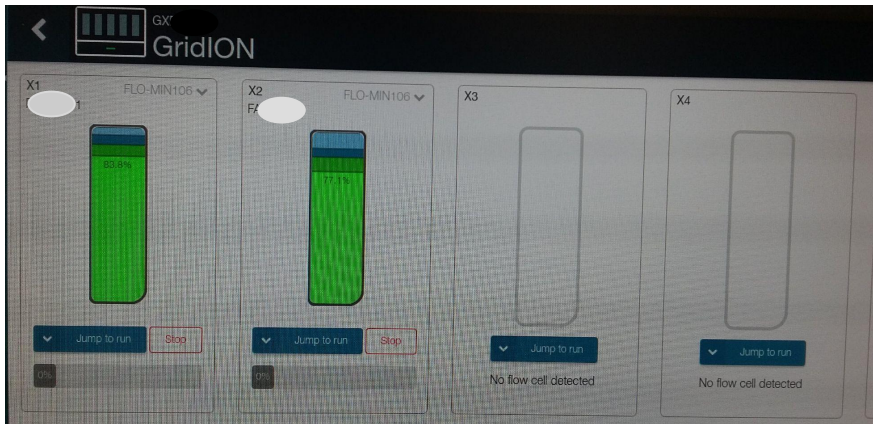
Starting sequencing

- 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



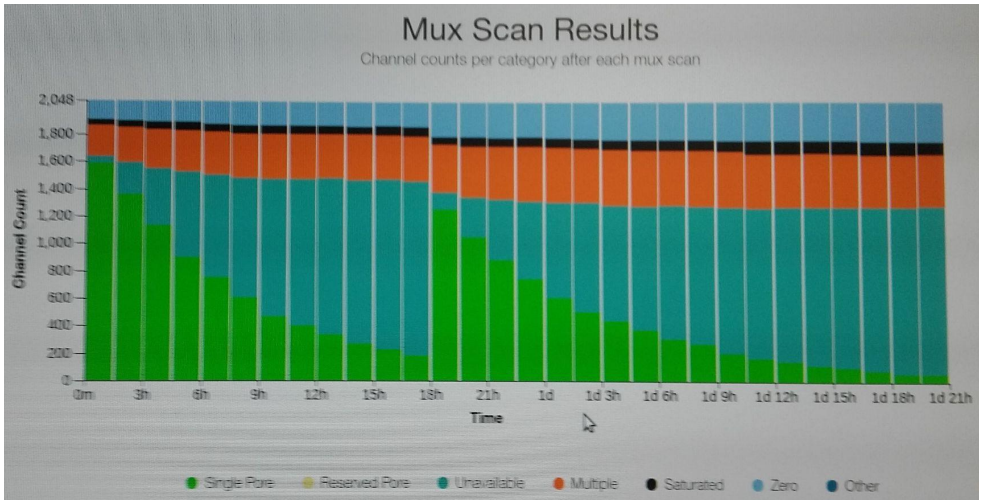
Monitor sequencing

- 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 of 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

Summary sequencing technologies

- Generation 1 (chain termination):
  - 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?