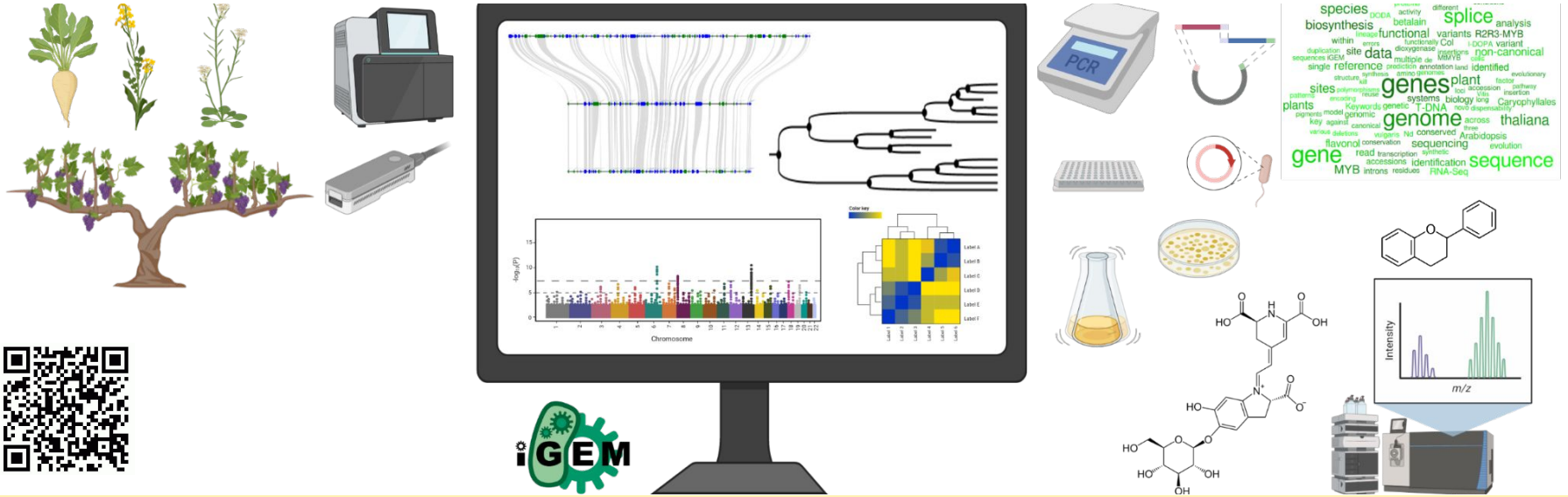
Technische Universität Braunschweig



Sequencing

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

Availability of slides

- All materials are freely available (CC BY) - after the lectures:
  - StudIP: **Data Literacy in Genomics**
  - GitHub: https://github.com/bpucker/teaching

- Questions: Feel free to ask at any time

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

Who discovered DNA and when?

Discovery of DNA

- Isolation of DNA by Friedrich Miescher in 1869
  - Working in the castle in Tübingen

Friedrich Miescher
(1844-1895)

- WATSON, J., CRICK, F. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171,** 737–738 (1953). https://doi.org/10.1038/171737a0

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

What is the genome of a plant?

Major types of DNA in plants

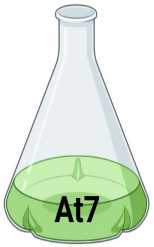
- gDNA from the nucleus
- mtDNA from the mitochondria (chondrome)
- cpDNA from the chloroplast (plastome)
- pDNA (plasmids, only in biotechnological applications)

Why is it important to have a genome sequence?

Advantages of a genome sequence

- Primer design (banana)
- Assessment of genetic diversity (population genomics)
- Trait discovery (yam)
- Breeding (sugar beet)
- Understanding the evolution (At7)

Markus Hagenlocher; https://de.wikipedia.org/wiki/Zucker#/media/Datei:Zucker%C3%BCbe.jpg



DNA extraction challenges

- High amount of cpDNA is a big issue in sequencing projects
  - 50-100 chloroplasts per cell with 50-100 plastomes per chloroplast

- Sequencing capacity is wasted on the cpDNA molecules
  - Very high coverage of the plastome; reduced coverage of nucleome

- Reducing amount of chloroplasts by incubating plants in the dark for some days prior to DNA isolation
  - Reduced amount of chloroplasts
  - Reduced concentration of starch/sugar

Golczyk et al., 2014: 10.1105/tpc.113.117465

Other nucleic acids

macromolecule	Percentage of total dry weight	Number of molecules per cell
protein	55	3,000,000
RNA	20	-
DNA	3	-
lipid	9	20,000,000

DNA extraction methods

- Plant genomic DNA
  - Edwards preparation: low quality but quick
  - cetyl trimethylammonium bromide (CTAB): high quality but slow
  - Nuclei isolation

- Plasmid DNA
  - TELT: cheap and good quality for small plasmids
  - Alkaline lysis: cheap and good quality
  - Standard plasmid isolation kit: high quality but expensive

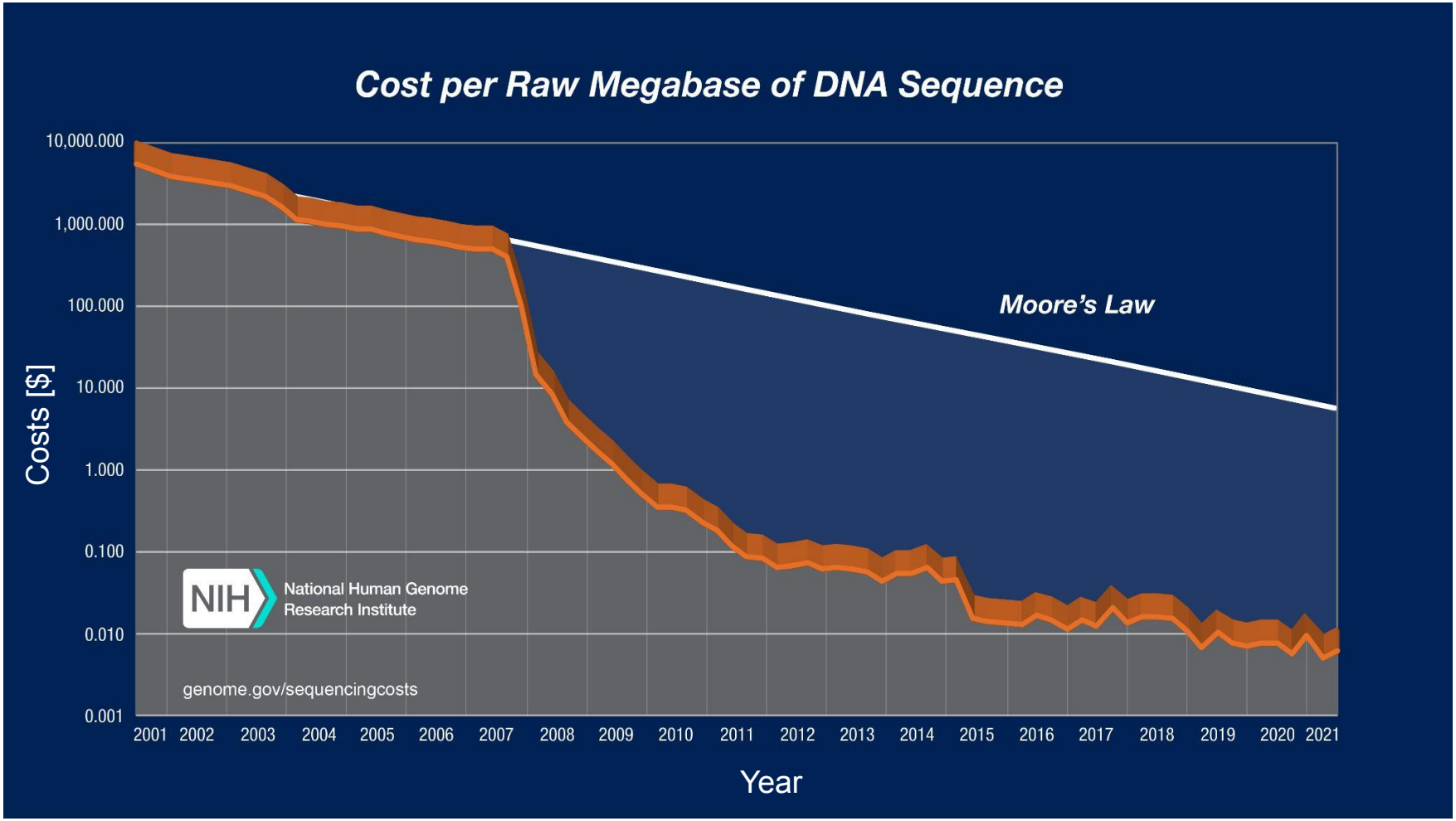
Edwards et al., 1991: 10.1093/nar/19.6.1349
Rosso et al., 2003: 10.1023/B:PLAN.0000009297.37235.4a
Mariac, 2021: 10.17504/protocols.io.83shyne
Medina-Acosta & Cross, 1993: 10.1016/0166-6851(93)90231-L
Birnboim & Doly, 1979: 10.1093/nar/7.6.1513

Do you know sequencing methods?

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?

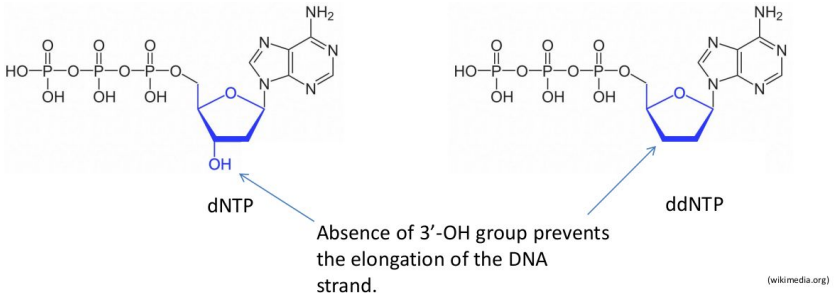
Development of sequencing capacity



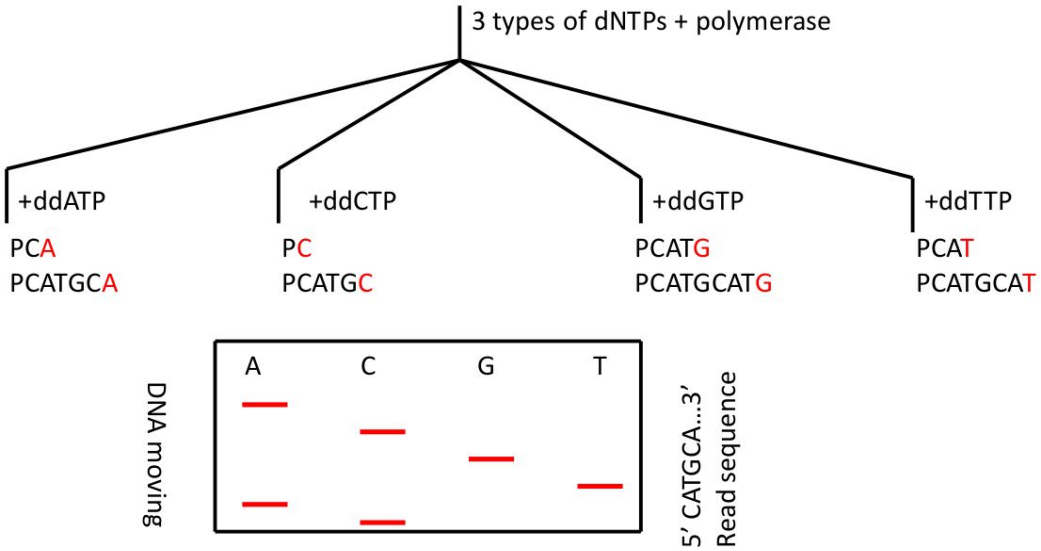
https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Sanger sequencing

Concept of Sanger sequencing

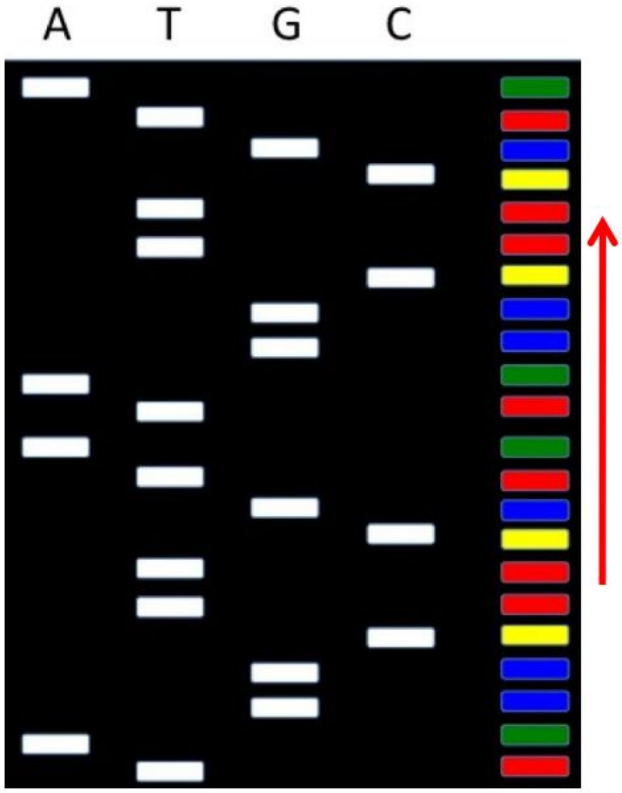


Primer (P): 5' –TGCATGGCATGATGCATG-3'
Template: 3' –ACGTACCGTACTACGTACGTACGTACGTCTAGGT-5'



Sanger sequencing - original version

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



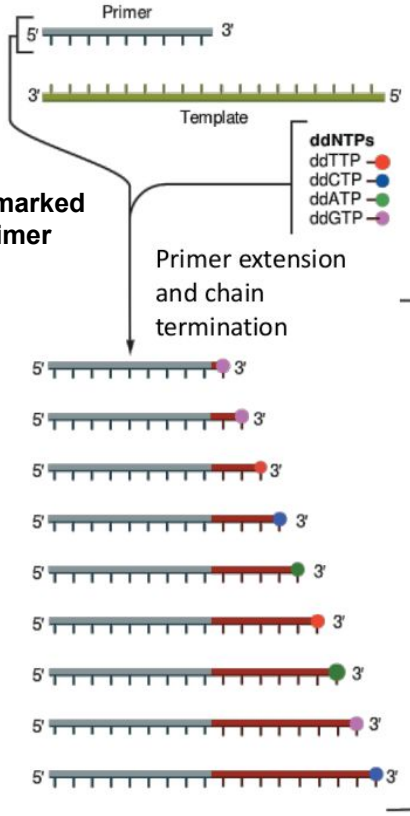
(modified from wikimedia.org)

Sanger sequencing - today

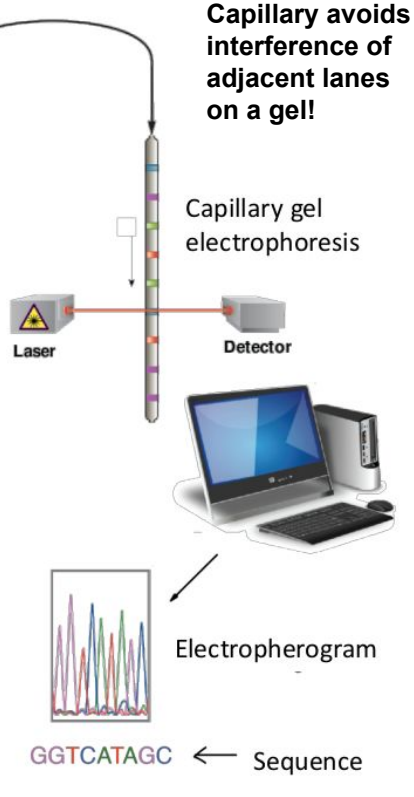
Only one reaction!

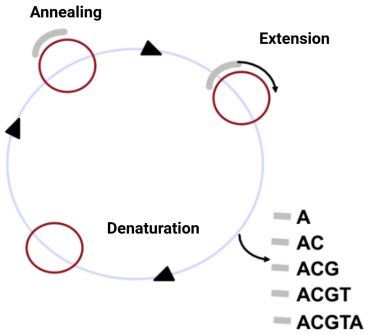
ddNTPs are marked instead of primer

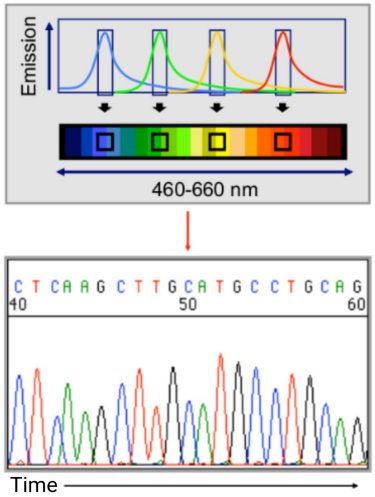
Low input required due to cycle sequencing



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
```

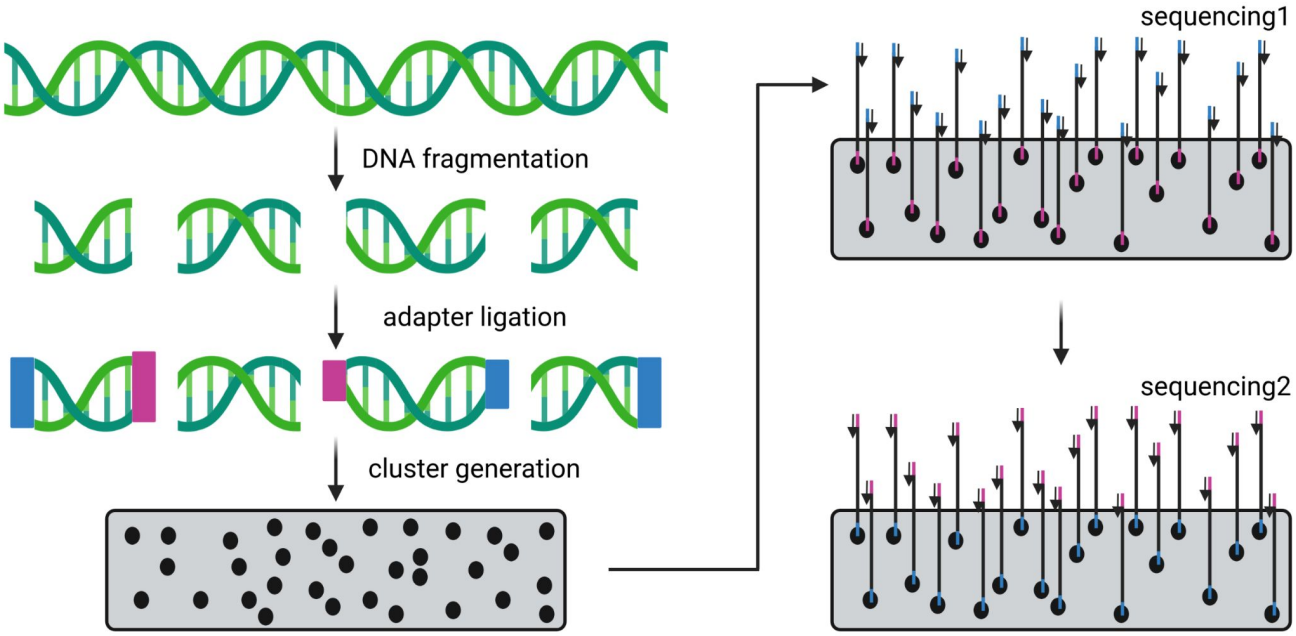
Phred-Score

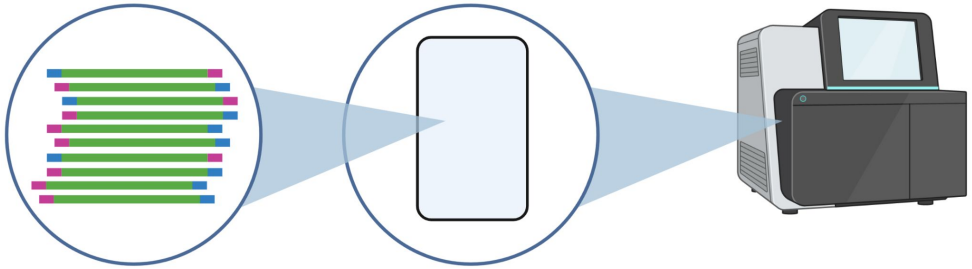
- Negative logarithm of the error probability for given position in read
- Multiplication by 10 to avoid floats

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%

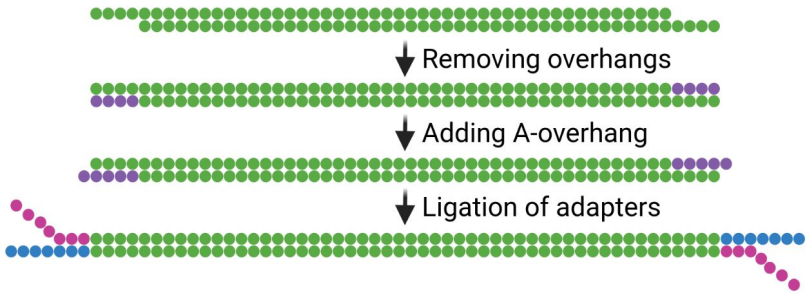
Illumina sequencing

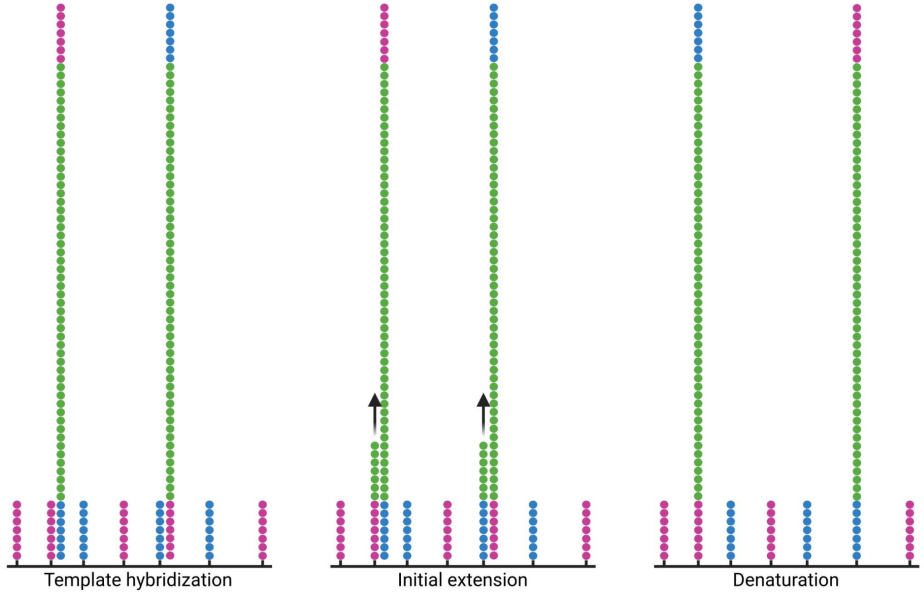
Illumina - sequencing overview



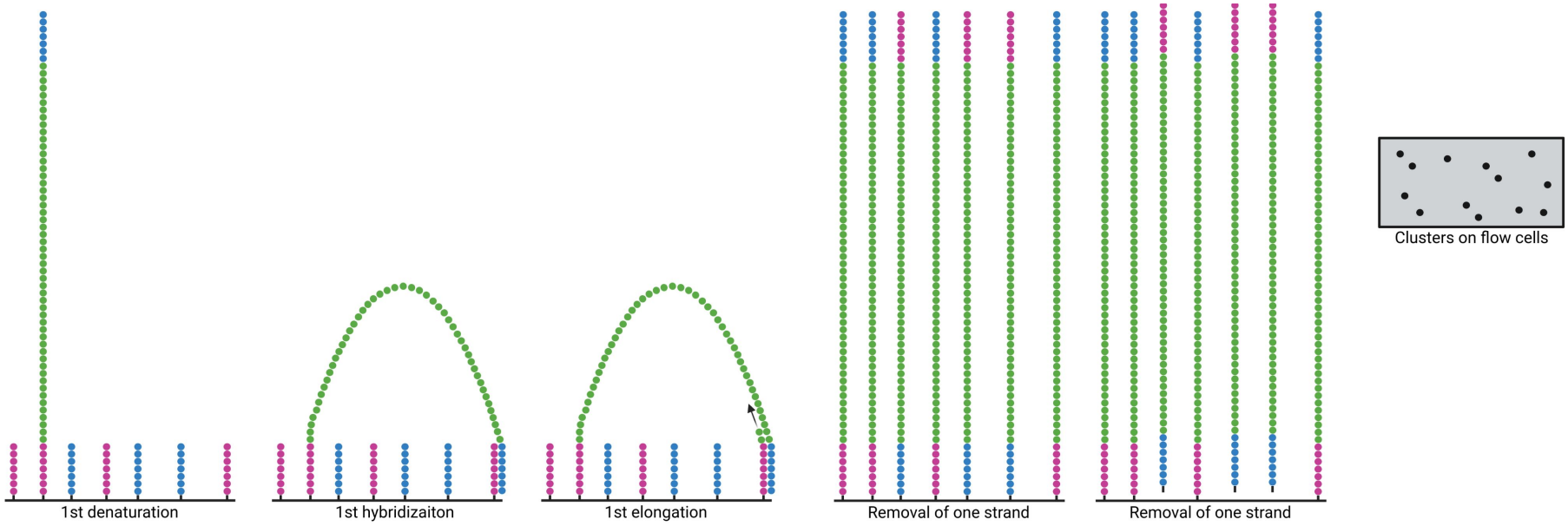


Illumina - sequencing 2

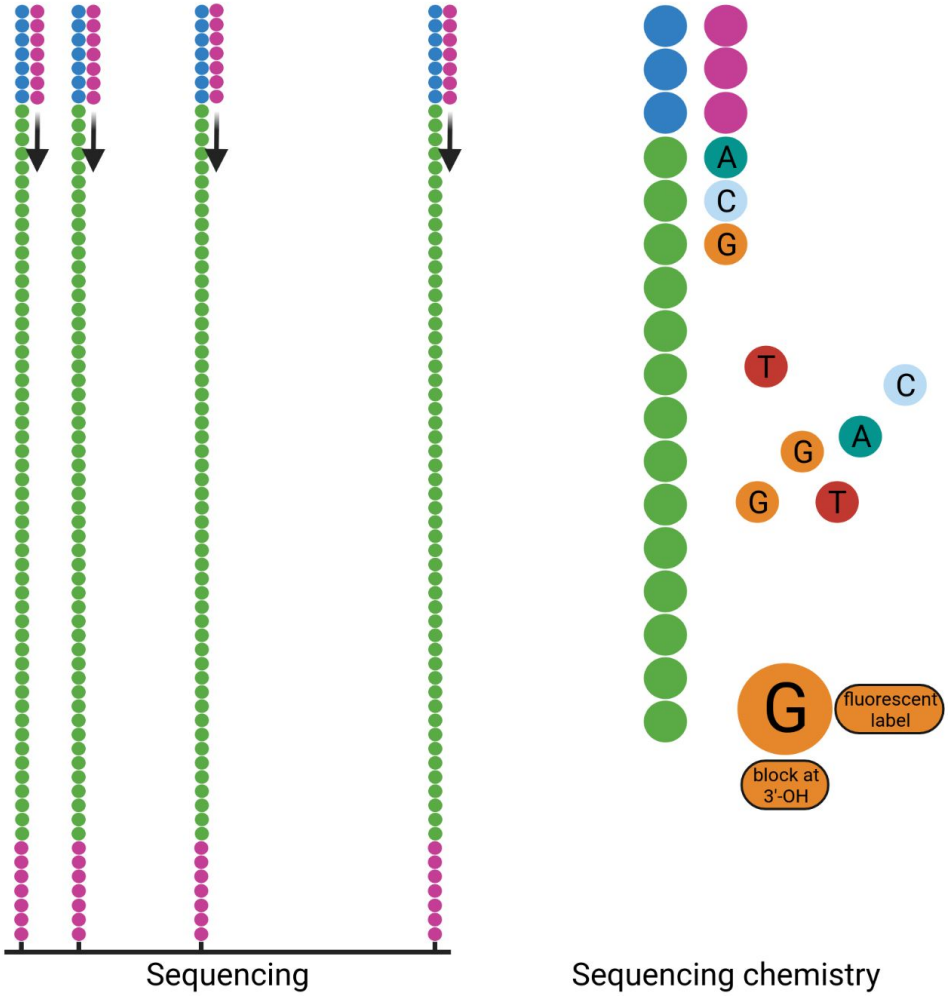




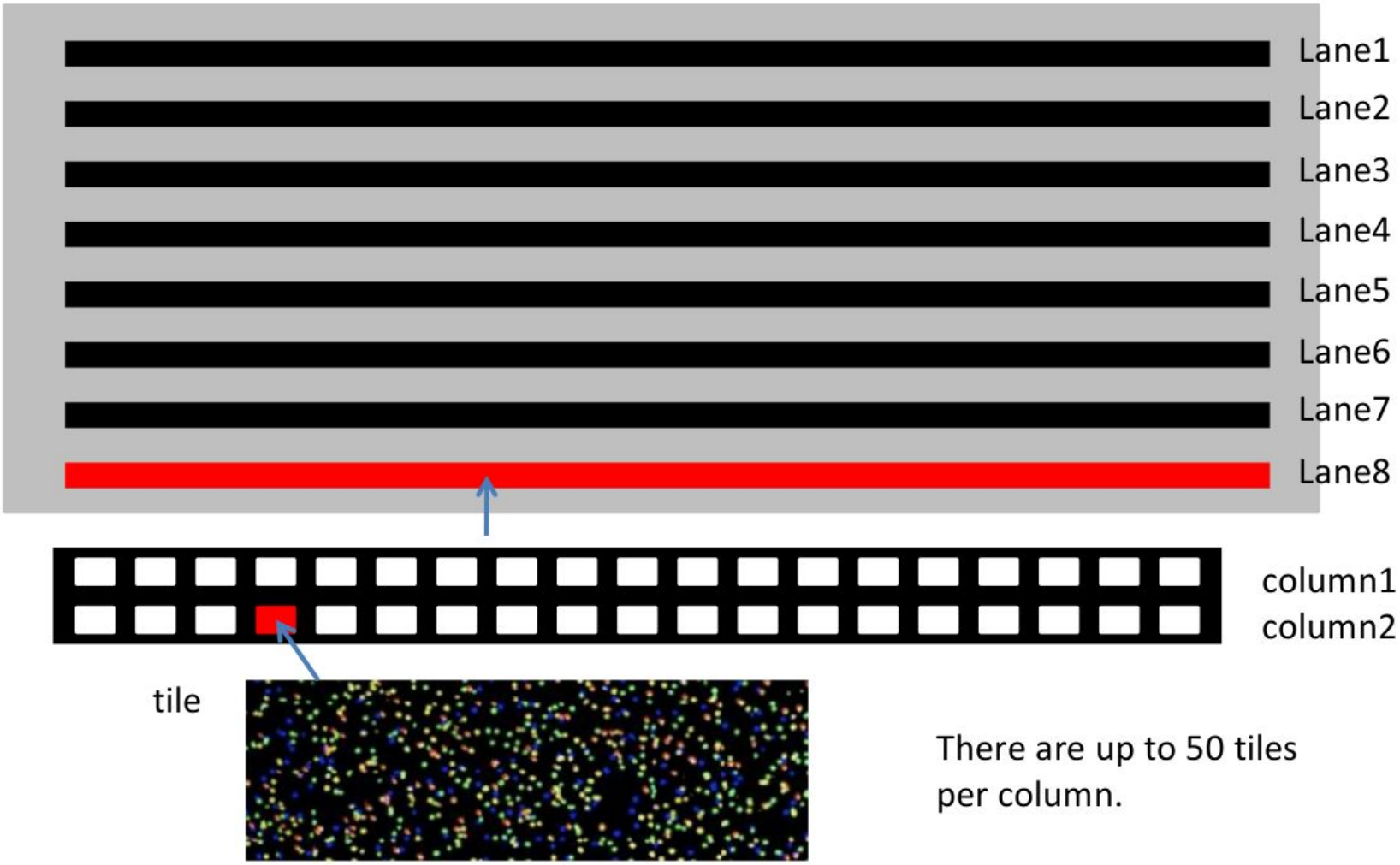
Illumina - sequencing 3



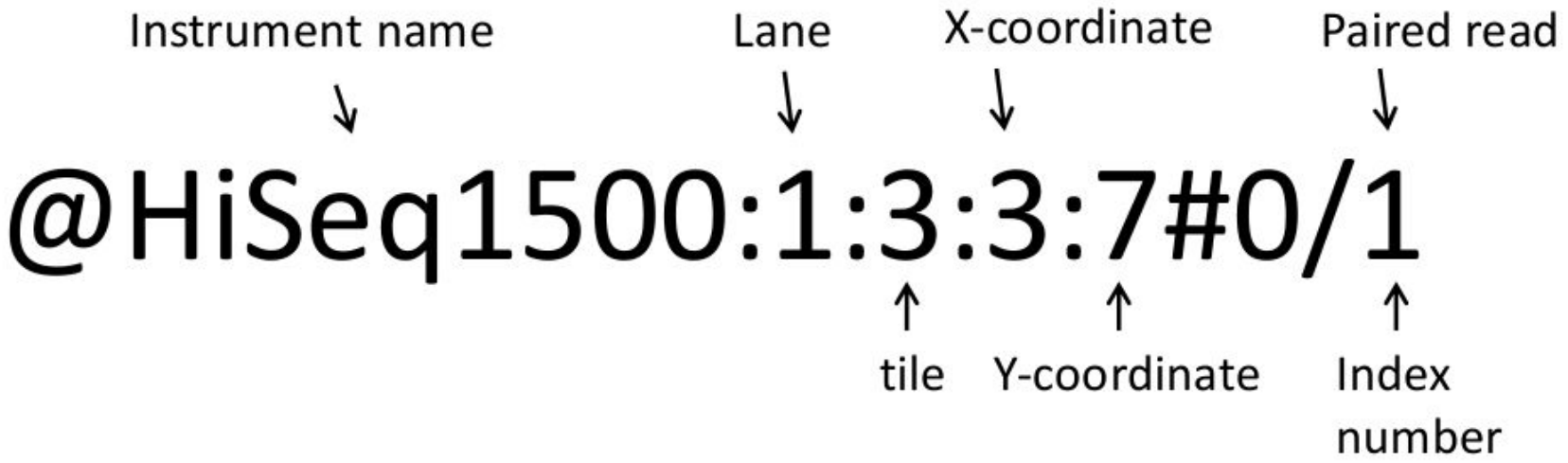
Illumina - sequencing 4



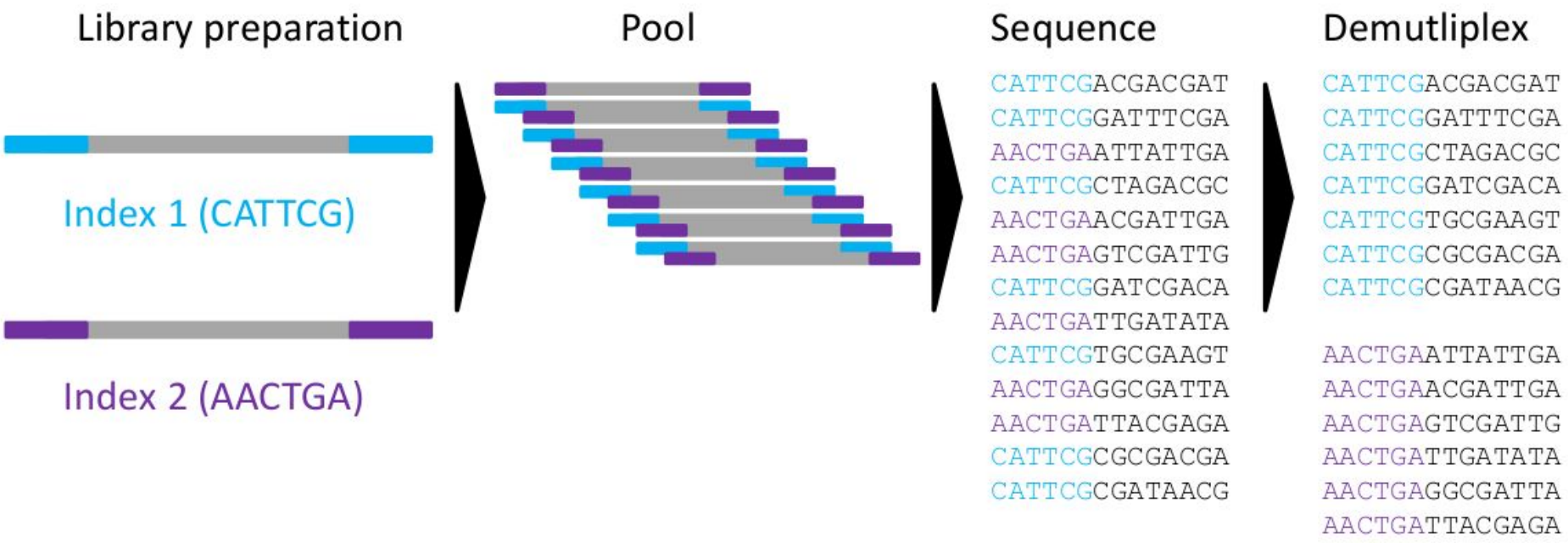
Illumina - flow cell layout



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):

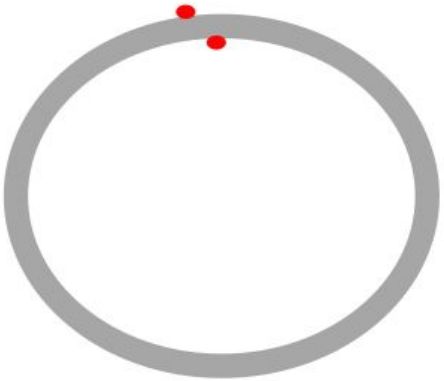
- Paired-end (PE):

Illumina - sequencing modi (mate pair)

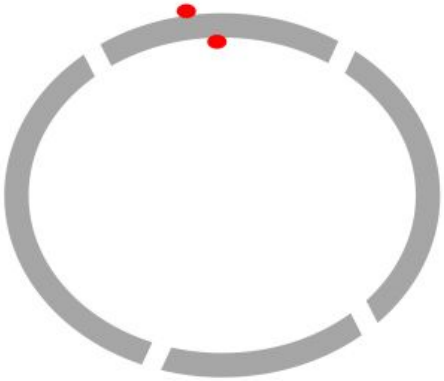
Fragmentation of DNA:

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

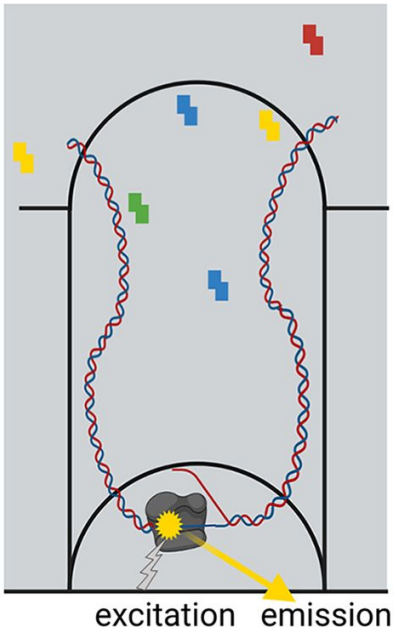
- Example:

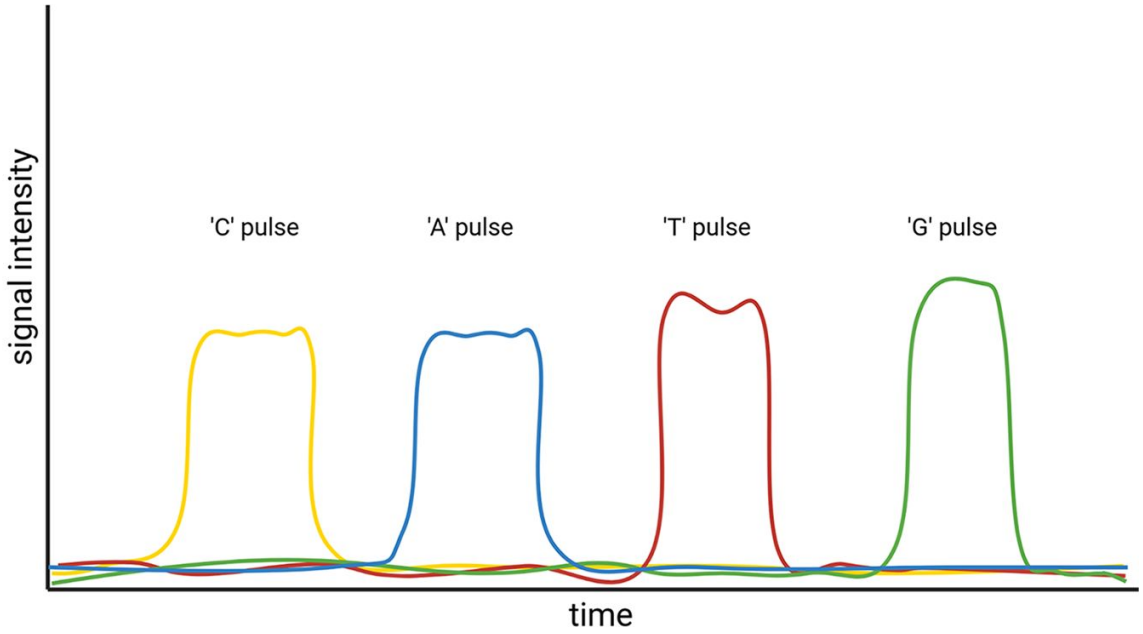
  @seq1
  ACGTACGTACGT
  +
  “”?CB””:DC”

PacBio

Pacific Biosciences (PacBio)

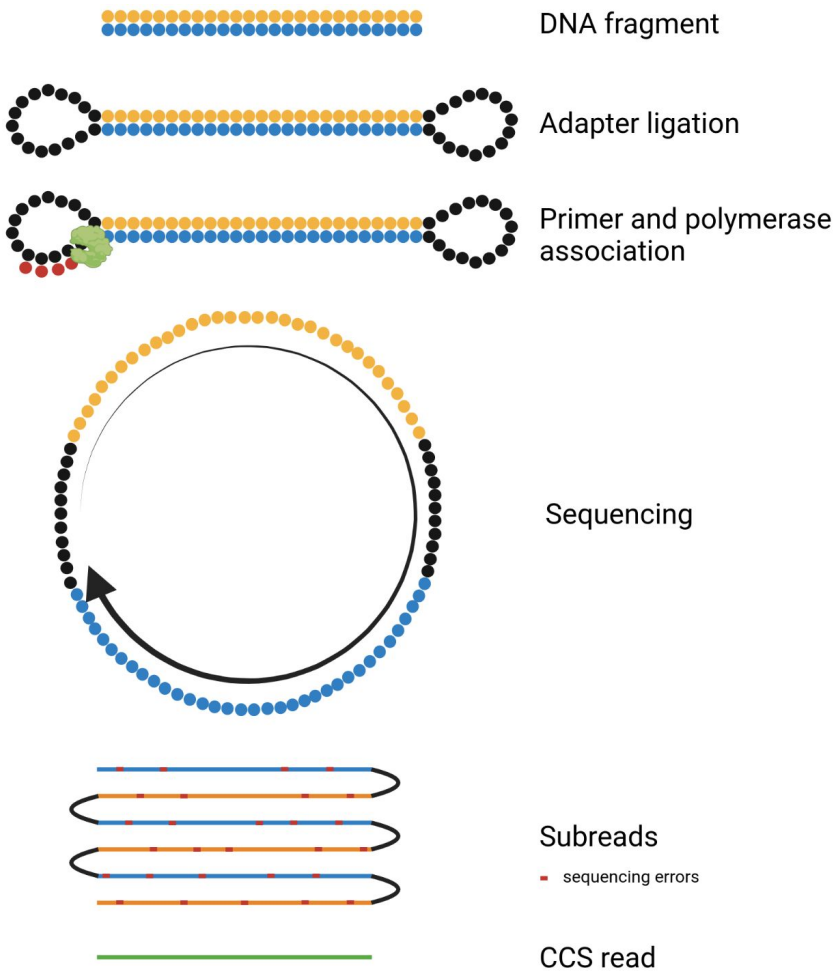
- Polymerase located at bottom of well
- ZMWs = Zero Mode Waveguide





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

PacBio - HiFi

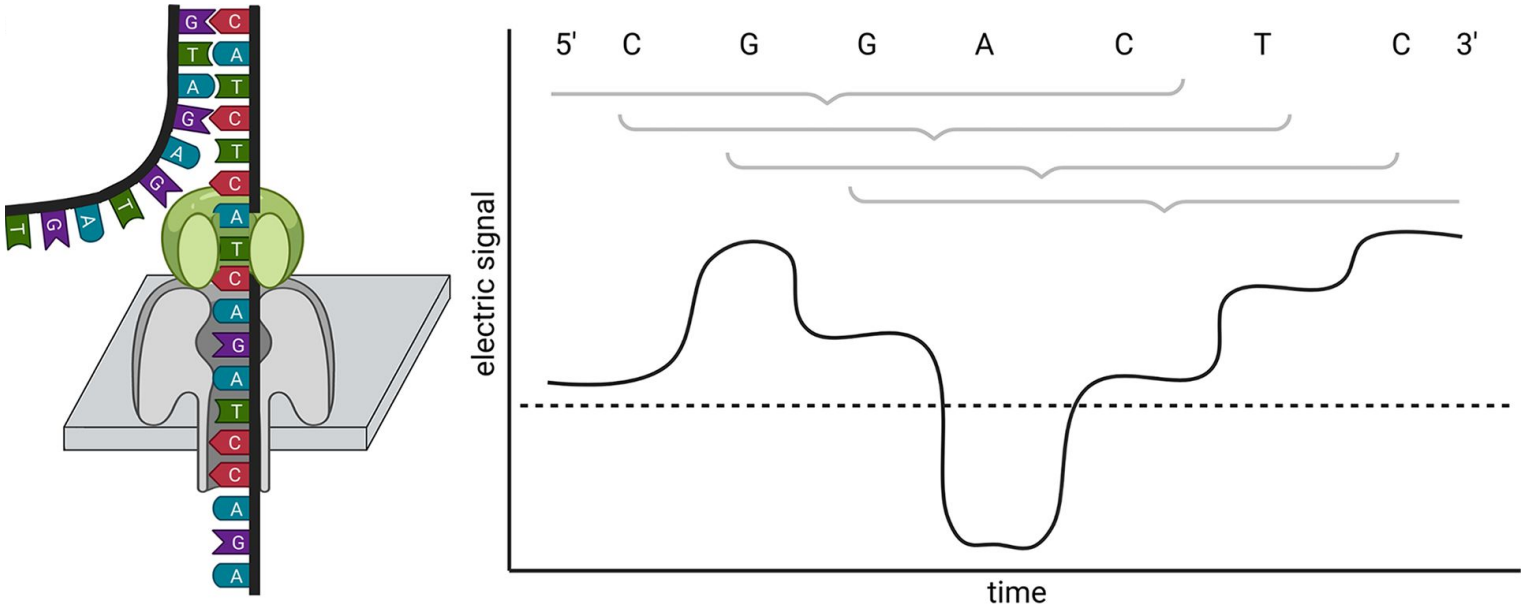

DNA fragment
Adapter ligation
Primer and polymerase association
Sequencing
Subreads
sequencing errors
CCS read

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

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
I		assembly	1-15d	1h			
J		polishing	1-5d	1h	compute cluster / cloud		
K		annotation	1-5d	1h			
L		data submission	2h	2h	fast internet connection		

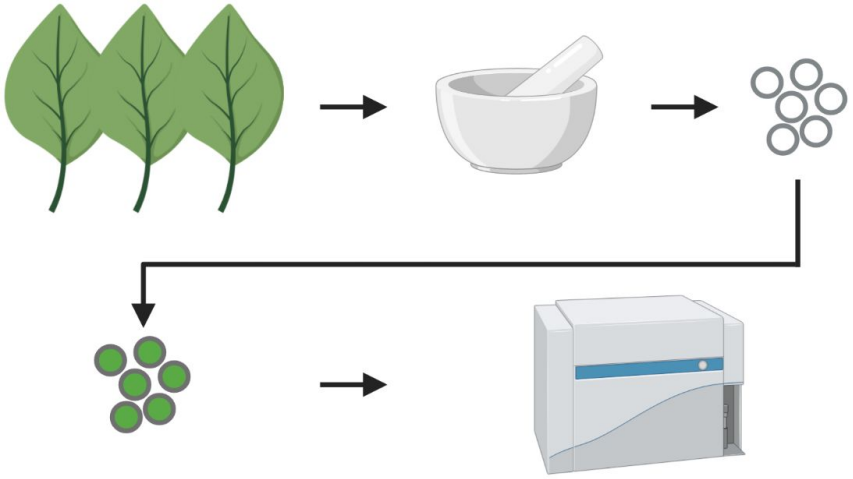
Pucker et al., 2022: 10.1017/qpb.2021.18

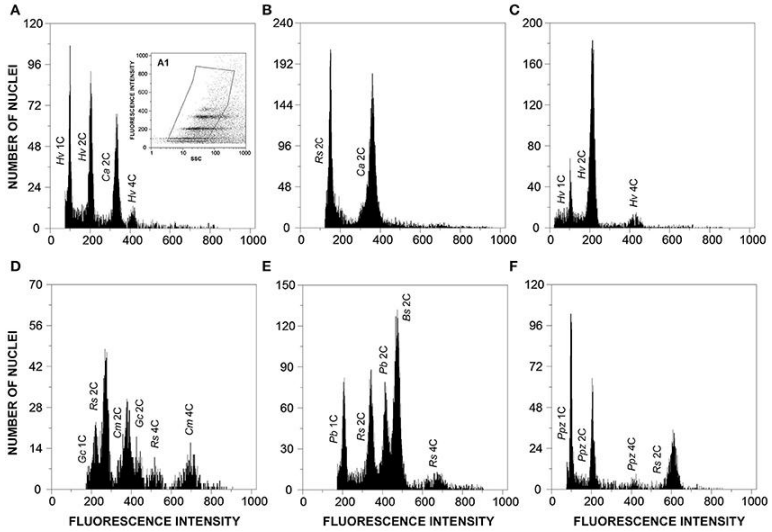
Considerations & preparations

- What is the expected/estimated genome size?
- What is the ploidy of the species?
- Which individual should be sequenced?
- Which plant parts are suitable for DNA extraction?
- What materials are needed for DNA extraction and sequencing?

Genome size

- Genome sequences of closely related species
- Flow cytometry is used to measure genome size biochemically
  - C-value
- Databases for plant genome sizes: https://cvalues.science.kew.org/

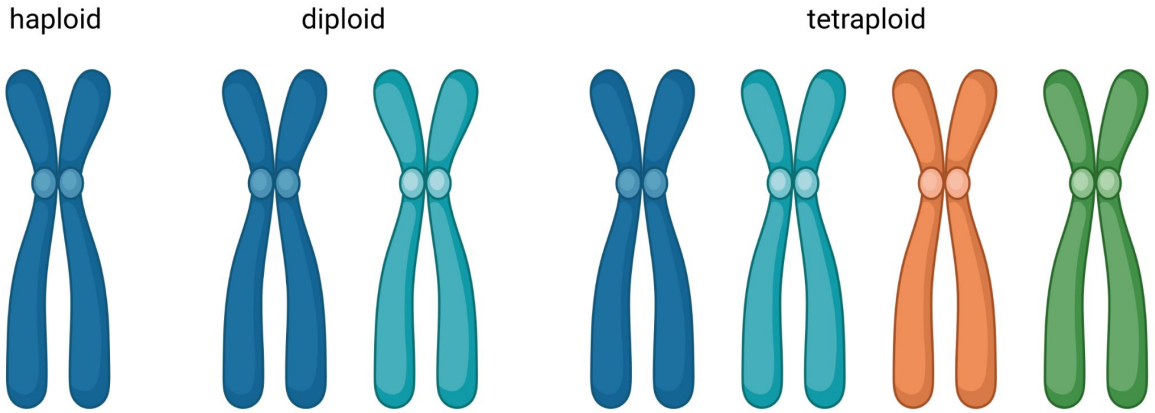




https://doi.org/10.3389/fpls.2014.00422

Ploidy

- Ploidy = copy number of same chromosome
- Many plants are polyploid
- Some polyploid plants have close diploid relatives



Picking an individual

- Plant should survive the sampling for DNA extraction
- Plant should be a good representation of the species
- Source of the plant is important:
  - Restrictions through material transfer agreements (MTAs)?
  - Restrictions through Nagoya protocol / Access and Benefit Sharing (ABS) law

Plant part for DNA extraction

- Young leaf is often a good choice
- Small cells result in higher density of nuclei per weight
- Concentration of specialized metabolites should be low
- Amount of sugar should be low
- Amount of chloroplast should be low
- Sample should not be contaminated with bacteria/fungi

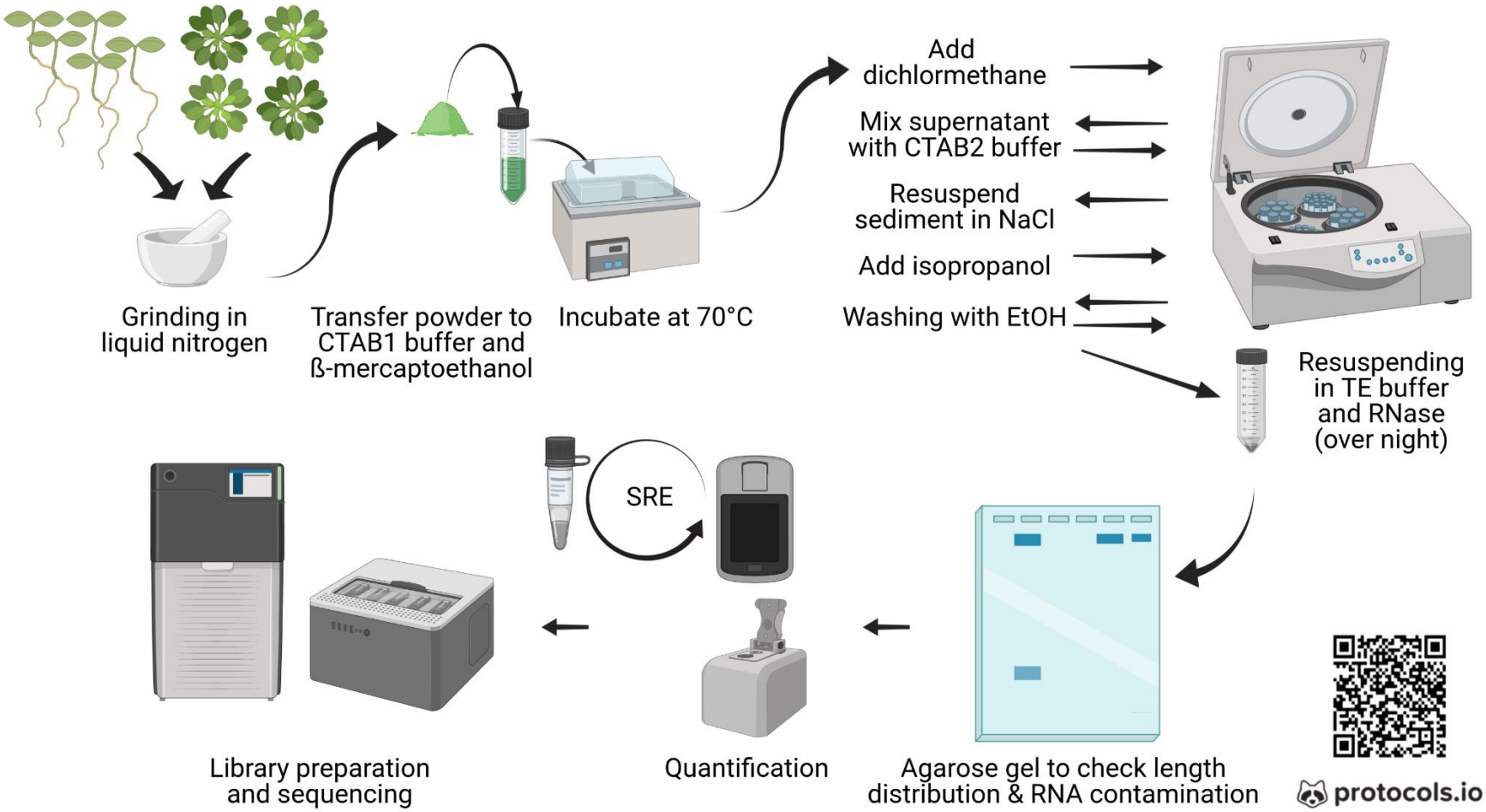
Material for DNA extraction

- Cetyltrimethylammonium-bromide (CTAB) buffer
- beta-mercaptoehtanol (ß-ME)
- Dichlormethane (replaces chloroform)
- TrisEDTA (TE)
- RNase A
- Short Read Eliminator (SRE) kit

Material required for sequencing

- DNA repair (companion module)
- Magnetic beads for purification
- Library preparation kit
- Sequencing kit
- Wash kit with DNase

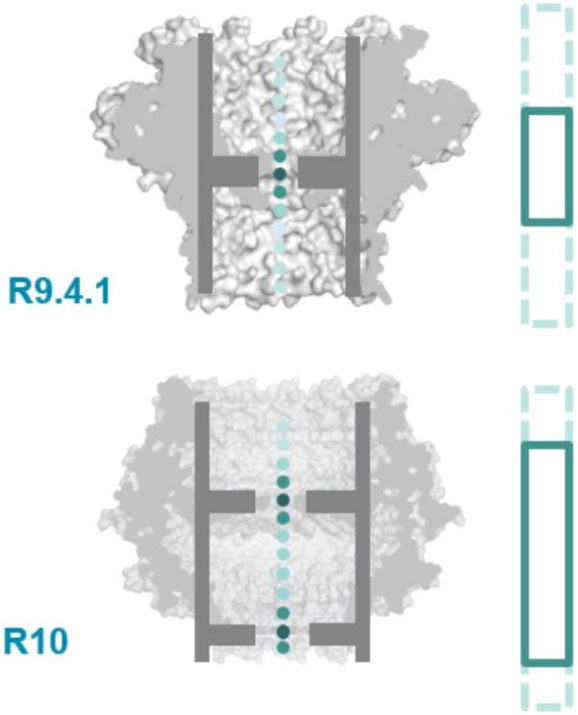
DNA extraction workflow



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	

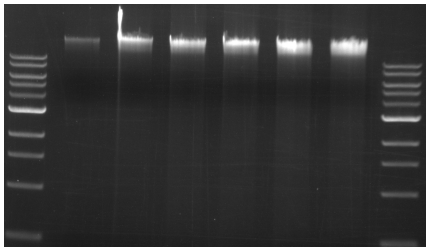
Nanopore comparison



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

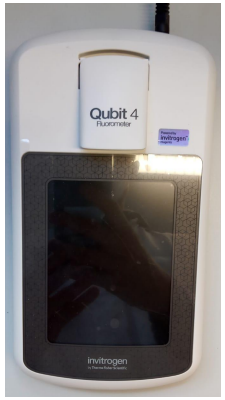
Quality control

- Agarose gel electrophoresis



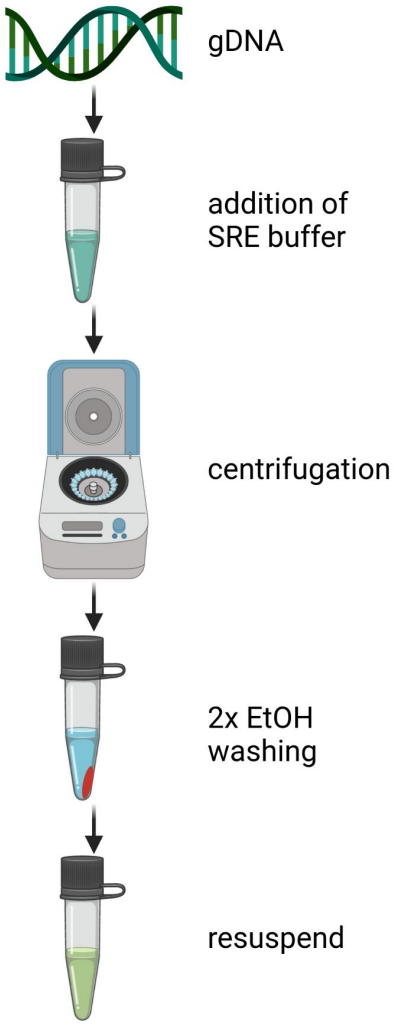
- Photometric measurement via NanoDrop

- 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

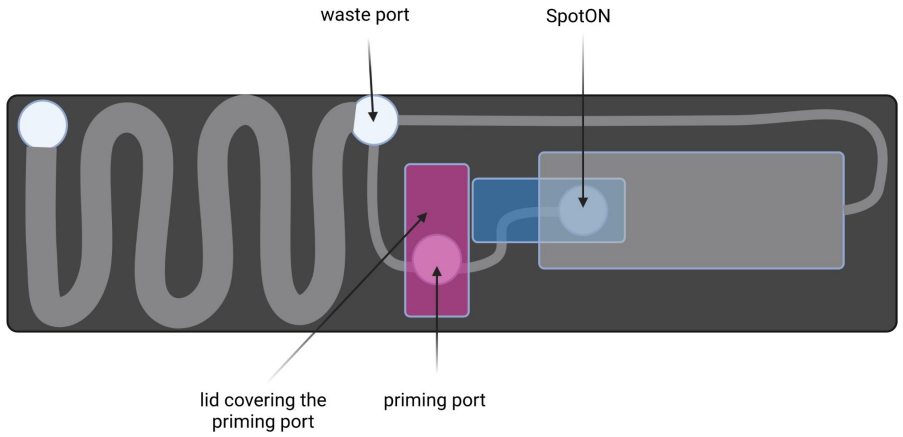
- Repaired DNA is subjected to library preparation
- Addition of adapters to DNA fragments
- Concentration of DNA can be quantified via Qubit measurement (optional)
  - Control step to ensure that library construction is working
- Purification of DNA with magnetic beads

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

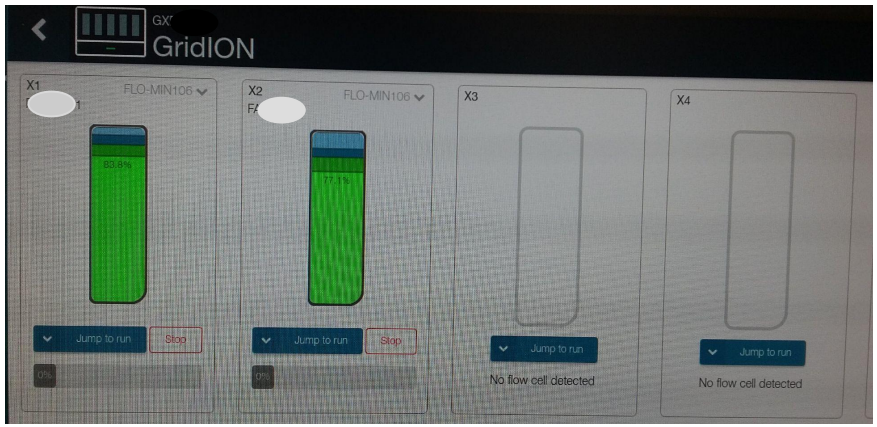


Starting sequencing

- 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

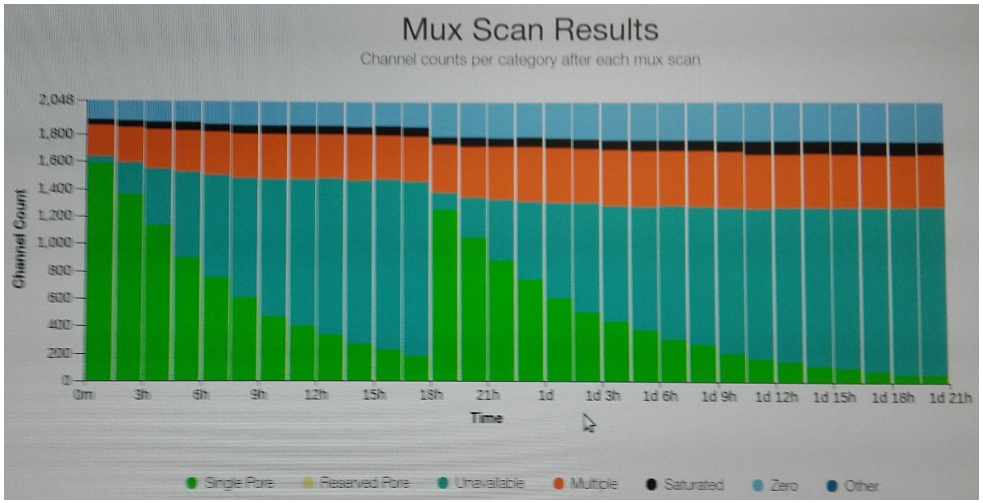
Monitor sequencing

- 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



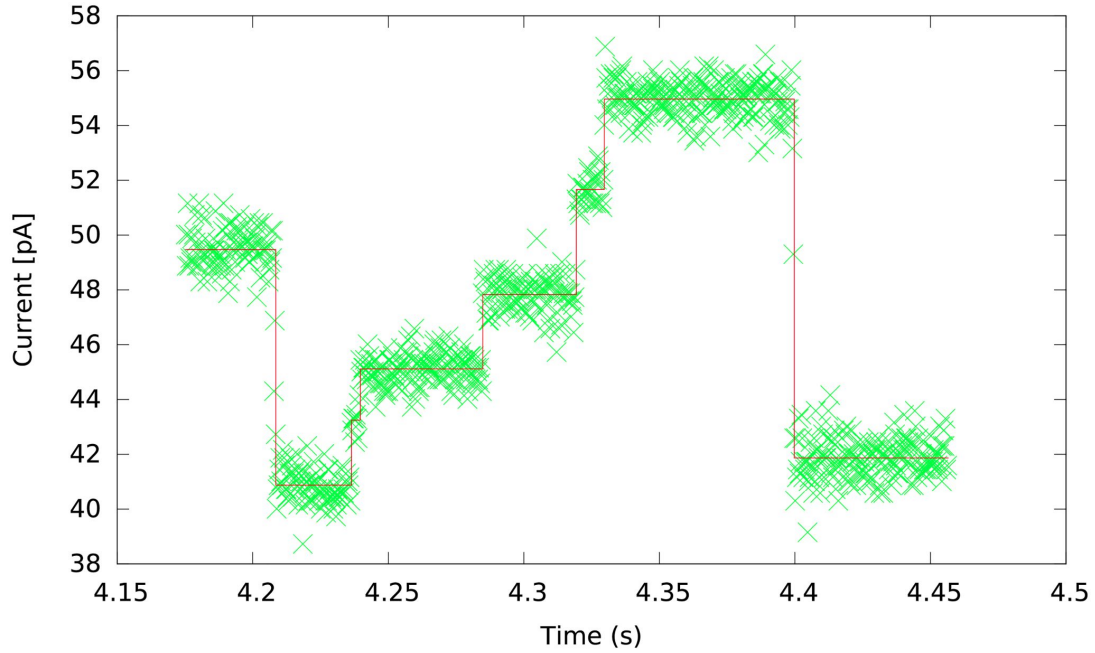
Stop & wash

- 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)



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

Barcoding

- Multiplexing of different samples in one sequencing run requires individual tags (=barcodes)
- PCR-free barcoding of 12 samples available
- Reads can be separated in real time based on barcodes

‘Read Until’ adaptive sampling / Targeted sequencing

- Check DNA strand based on first sequenced bases against reference:
  - strand is of interest: continue
  - strand is not of interest: reject

- Examples: UNCALLED (Kovaka et al., 2020); COSMIC (Payne et al., 2020)

- Cas9 Sequencing Kit:
  - Dephosphorylation of 5’-DNA ends
  - Cas9 binds to target sequences (based on guide RNA)
  - Cleavage results in blunt ends and 5’-phosphorylation
  - 3’ dA-tailing to prepare for adapter ligation
  - Adapters are preferentially ligated to Ca9 cut sites

Typical results of ONT sequencing projects

- Cost-effective and quick generation of descend genome sequence
- Some chromosome arms represented by single contigs
- N50 lengths depend on genome size and repetitiveness
- Centromeric and other large repetitive regions remain a challenge

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. What do you know about the discovery of DNA?
2. What is a plant genome?
3. What are the advantages of having a genome sequence?
4. Which sequencing technologies do you know?
5. What is the structure of a FASTA file?
6. What is a Phred score?
7. What are differences between Sanger sequencing and Illumina sequencing?
8. How does mate pair sequencing work?
9. What is the structure of a FASTQ file?
10. How does PacBio sequencing work?
11. How does ONT sequencing work?
12. What are the important steps of an ONT sequencing workflow?
13. How can you estimate a genome size?
14. What are important differences of ONT vs. PacBio sequencing?
15. What are important QC steps in high molecular weight DNA extraction?
16. Which parameters can be monitored during ONT sequencing?