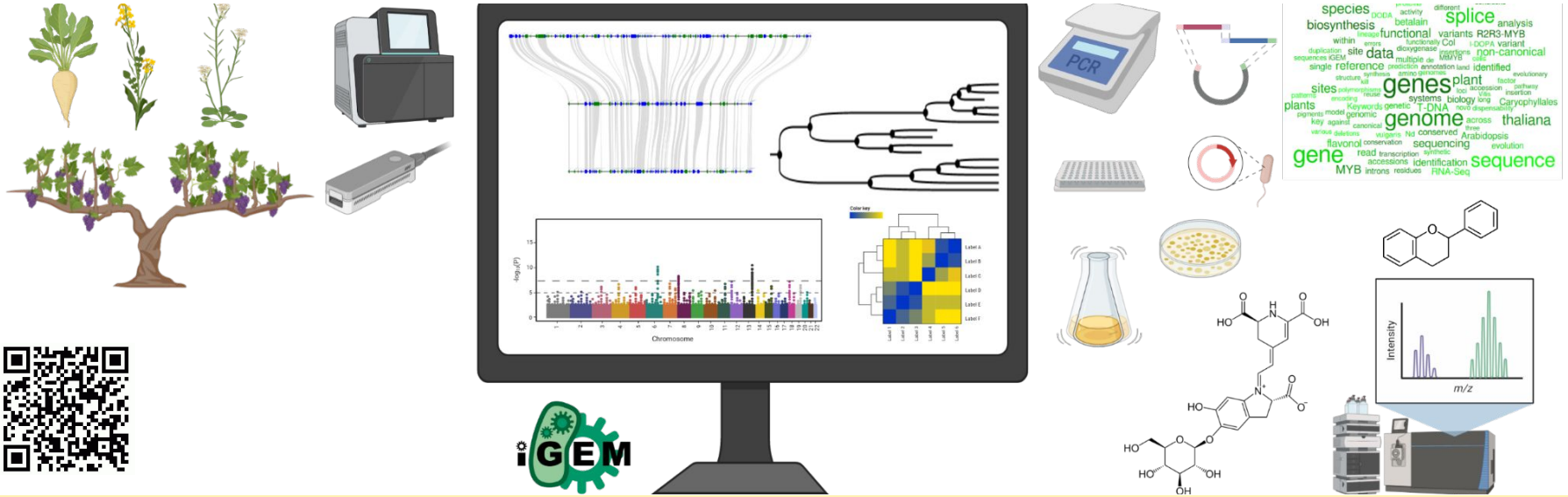
Technische Universität Braunschweig



Evolution of Plant Genomics

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

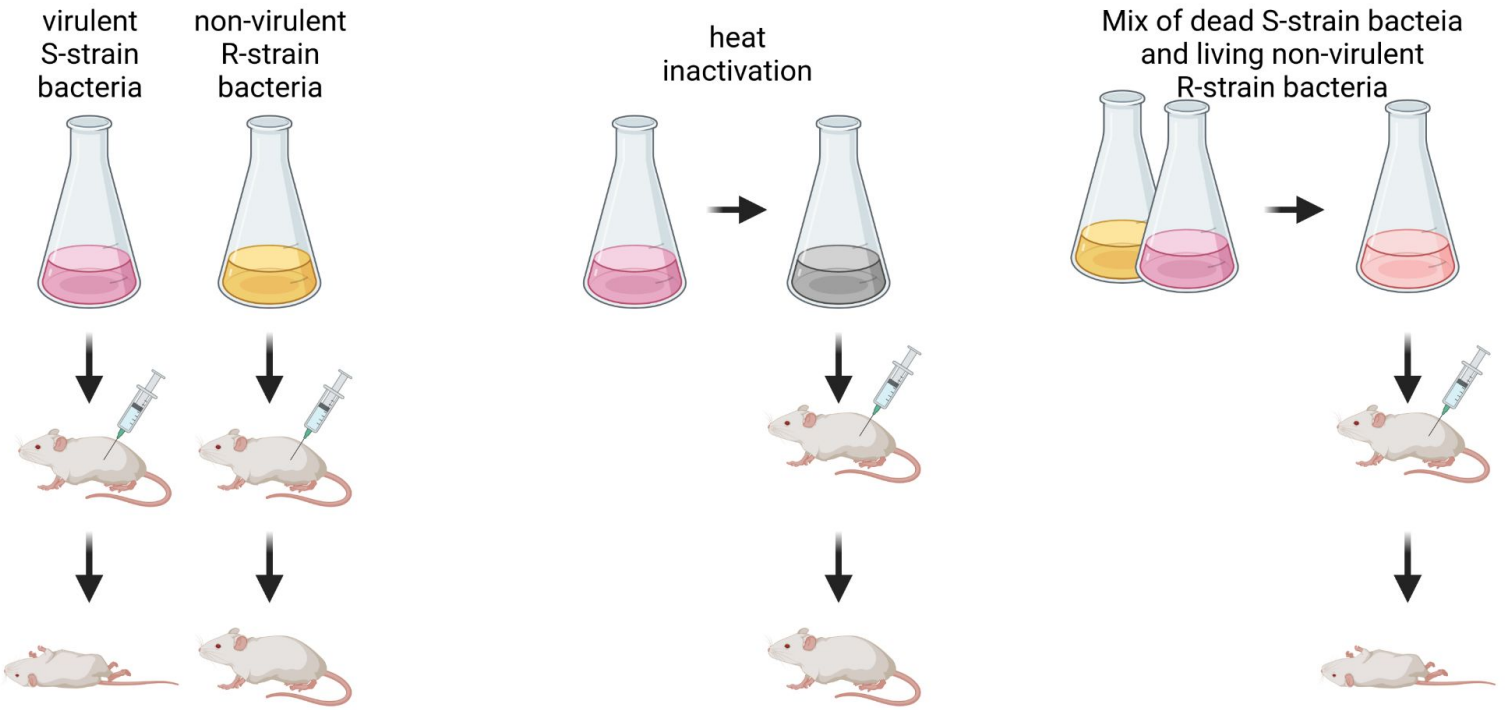
Availability of slides

- All materials are freely available (CC BY) - after the lectures:
  - StudIP: **GE31/MM12**
  - GitHub: https://github.com/bpucker/teaching

- Questions: Feel free to ask at any time

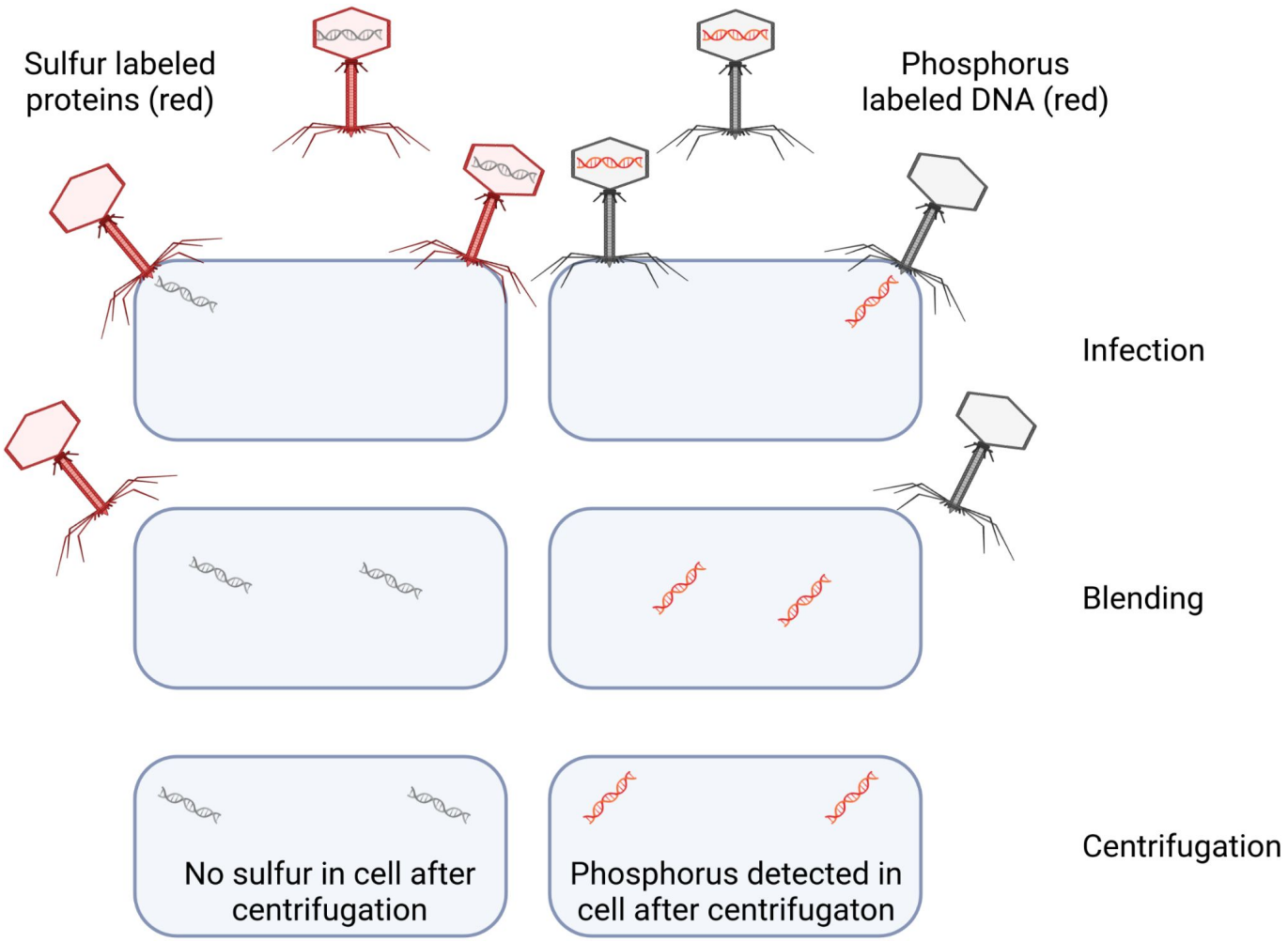
- Feedback, comments, or questions: b.pucker[a]tu-braunschweig.de

Discovery of DNA as genetic material (1)



Griffith, 1928: 10.1017/S0022172400031879

Discovery of DNA as genetic material (2)



Hershey & Chase, 1952: :10.1085/jgp.36.1.39

Discovery of DNA

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

Friedrich Miescher
(1844-1895)

Discovery of DNA structure

- 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

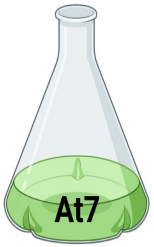
- gDNA from the nucleus
- mtDNA from the mitochondria (chondrome)
- cpDNA from the chloroplast (plastome)
- pDNA (plasmids, only in biotechnological applications)

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

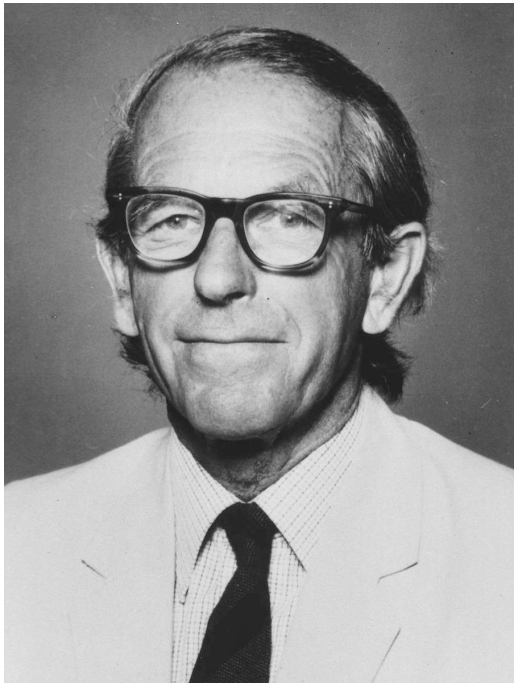
- Plasmid DNA
  - TELT: cheap and good quality for small plasmids
  - Alkaline lysis: cheap and good quality
  - Standard plasmid isolation kit: high quality but expensive

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

Inventor of Sanger sequencing

Nobel prizes for:

1) Protein sequencing (1958)
2) DNA sequencing (1980)



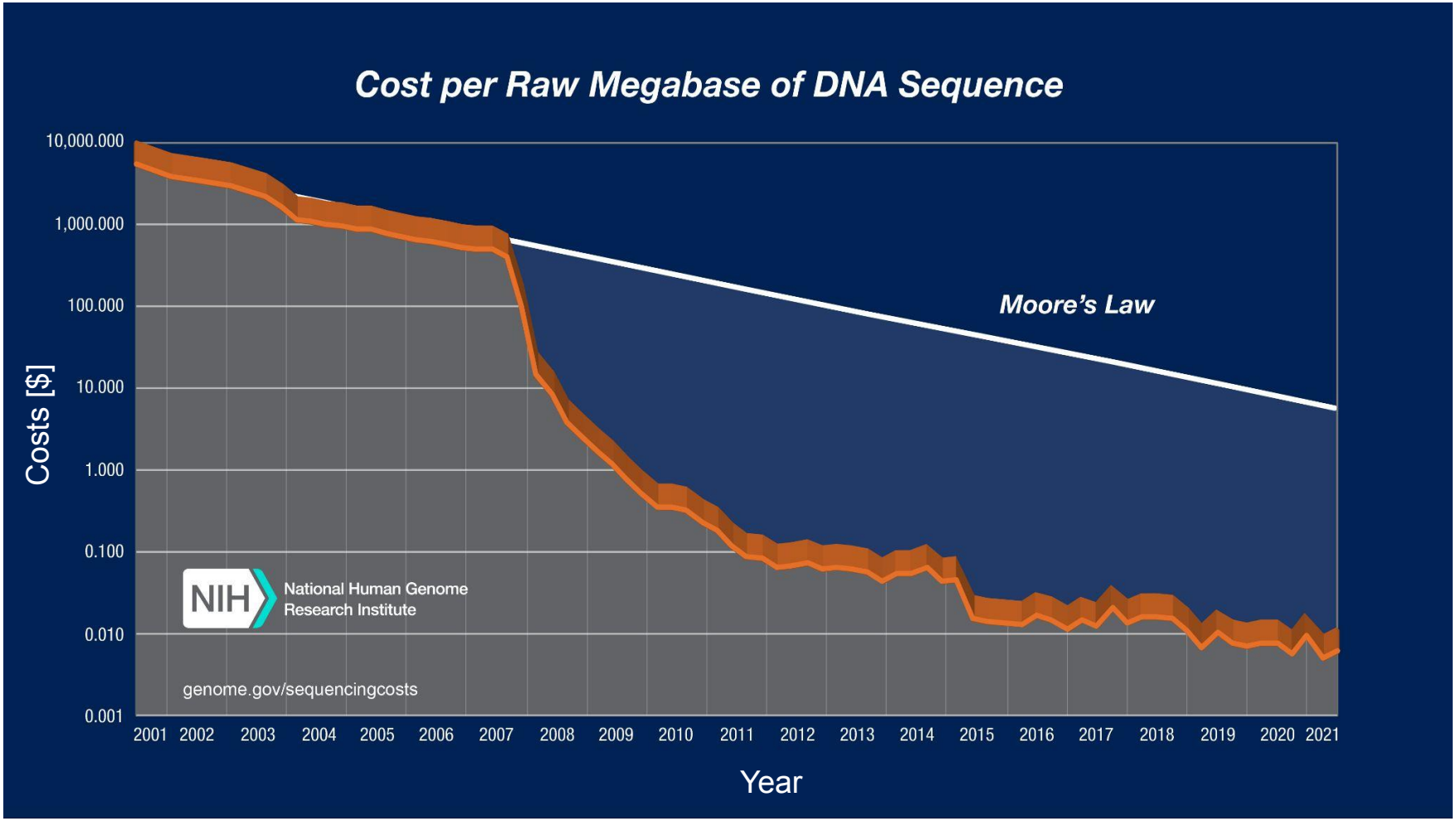
Frederick Sanger
(1918-2013)

Do you know other 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?

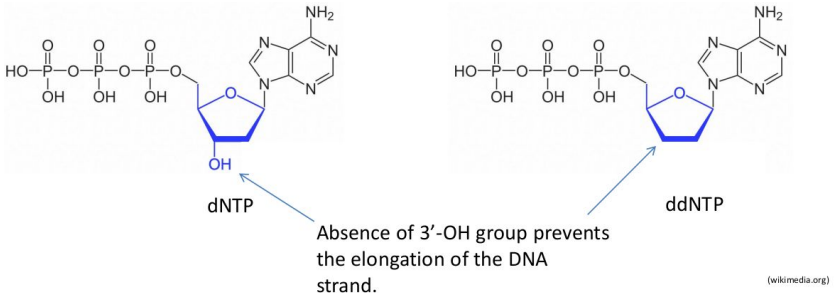
Development of sequencing capacity



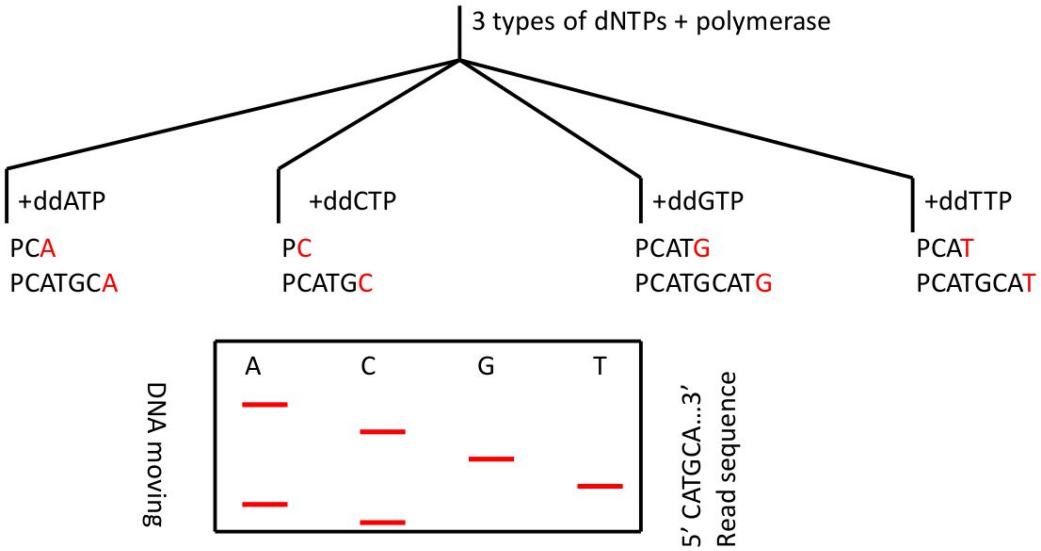
https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Sanger sequencing

Concept of Sanger sequencing

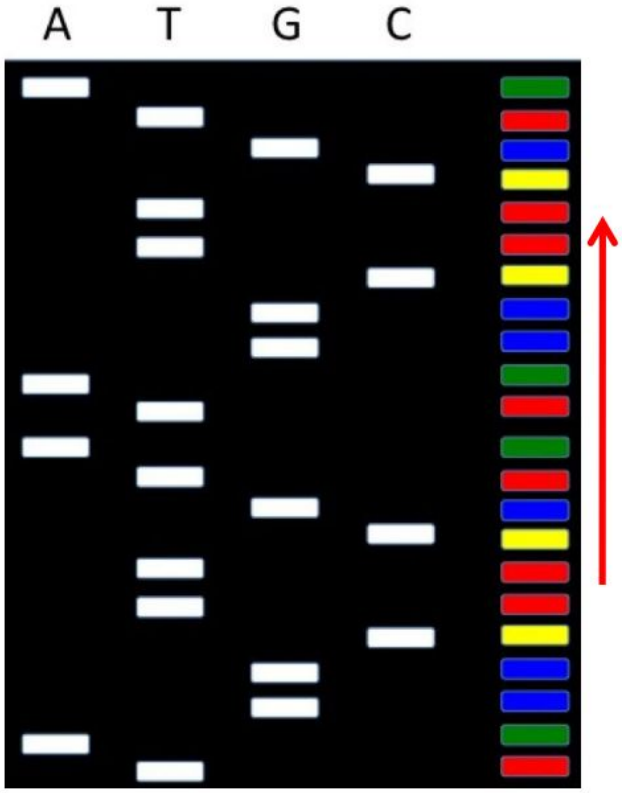


Primer (P): 5' –TGCATGGCATGATGCATG-3'
Template: 3' –ACGTACCGTACTACGTACGTACGTACGTCTAGGT-5'



Sanger sequencing - original version

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



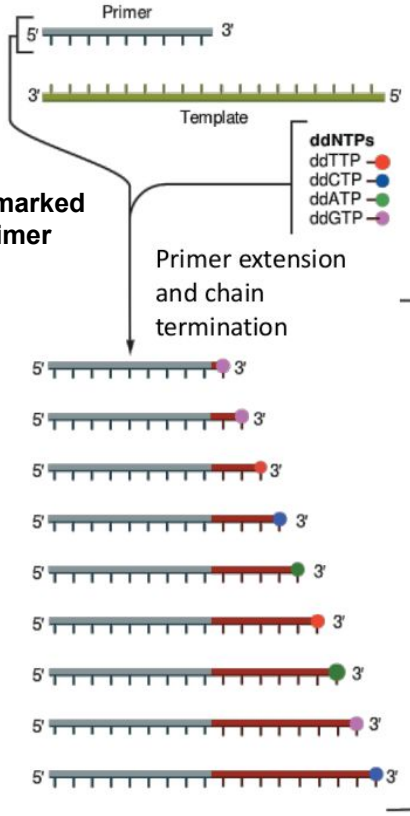
(modified from wikimedia.org)

Sanger sequencing - today

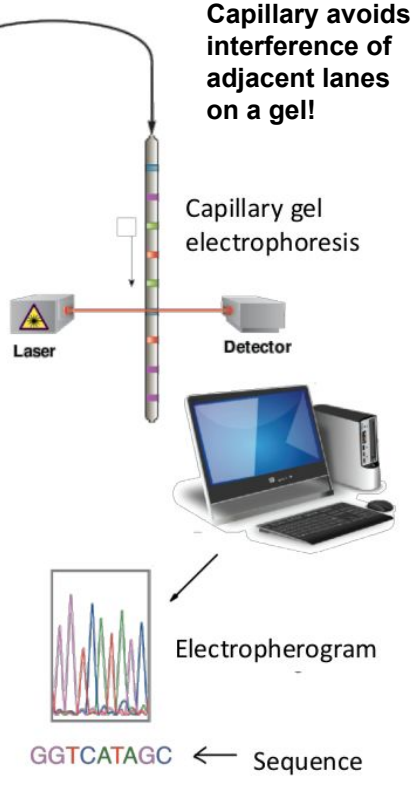
Only one reaction!

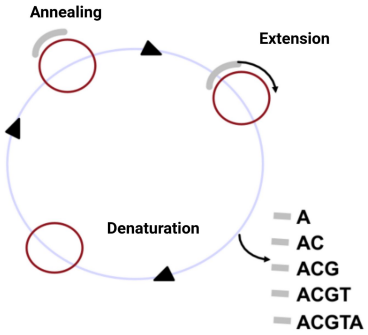
ddNTPs are marked instead of primer

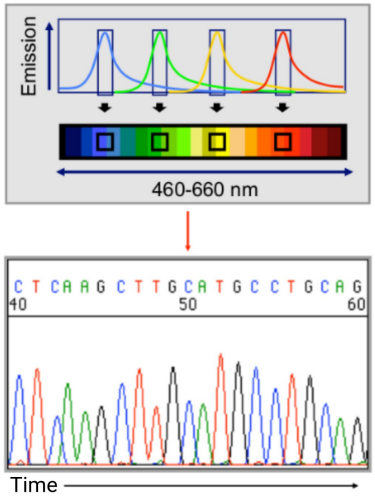
Low input required due to cycle sequencing



Capillary avoids interference of adjacent lanes on a gel!



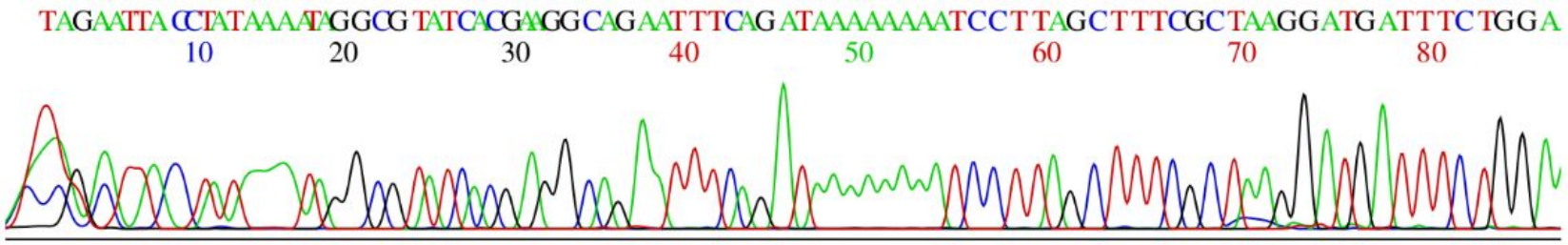




Figures modified from wikipedia

TRACE (.abi/.ab1)

- Original result file of ABI basecaller (Sanger sequencing)
- Contains only one red per file



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
```

QUALA

- There are two types of lines: header and quality
- Header line starts with '>', can contain name and information about sequence
- One entry corresponds to a FASTA file entry
- Example:

```
>seq1 len=5
10 11 12 8 6
>seq2 len=10
10 11 12 11 11
10 10 10 6 4
>seq3 len=1
15
```

Phred-Score

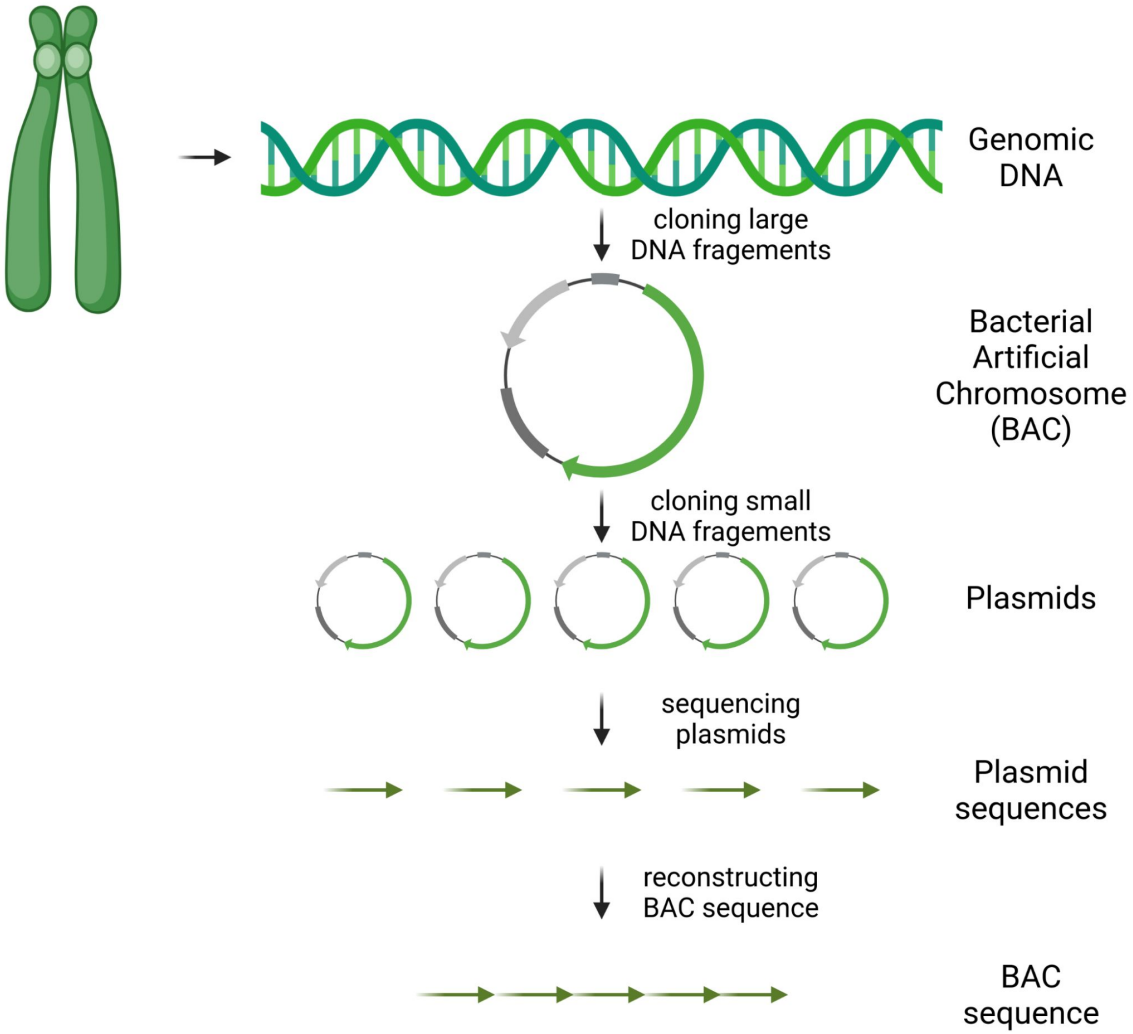
- Negative logarithm of the error probability for given position in read
- Multiplication by 10 to avoid floats

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%

Sanger sequencing - practical consideration

- Only one primer! (not a PCR)
- Primer needs to bind uniquely
- Primer needs to bind at 58°C
- Do NOT submit your samples in TE buffer (EDTA prevents sequencing)
- Amount of required DNA input depends on plasmid/fragment size

BAC-based sequencing strategy



What is a contig/scaffold?

Contigs and scaffolds

- Contig = continuous sequence

```
ACGATACAGTAACCTACCGGATTAGACT
```

- Scaffold = compilation of contigs with interleaved, unknown sequence (gaps)

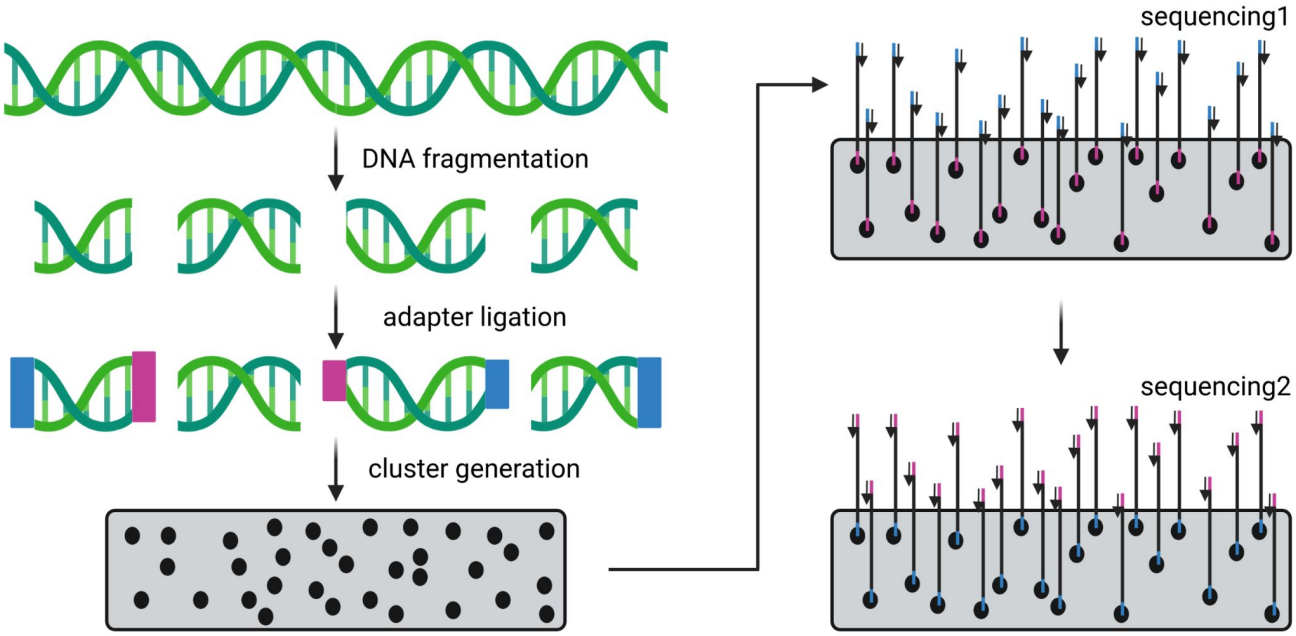
```
ACGATACAGTAACCTACCGGATTAGACTNNNNNNNACGACGGATACGTACAGTNNNNNNNNATTAGACCA
```

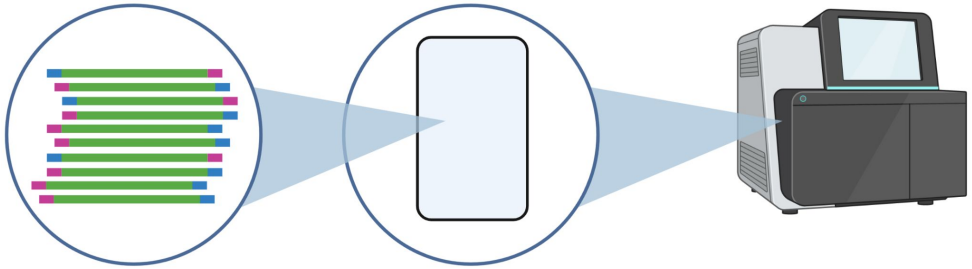
Sequencing the *Arabidopsis thaliana* genome

2000	2005	2010	2015	2020
Col-0 genome sequence release	Annual improvements of reference sequence	Re-sequencing projects	Ler and Nd-1 genome sequence release	Many more genome sequences released
The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408, 796–815 (2000). https://doi.org/10.1038/35048692		Cao, J., Schneeberger, K., Ossowski, S. et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 43, 956–963 (2011). https://doi.org/10.1038/ng.911 Weigel, D., Mott, R. The 1001 Genomes Project for Arabidopsis thaliana. Genome Biol 10, 107 (2009). https://doi.org/10.1186/gb-2009-10-5-107	Zapata et al., 2016. https://doi.org/10.1073/pnas.1607532113 Pucker B, Holtgräwe D, Stadermann KB, Frey K, Huettel B, Reinhardt R, et al. (2019) A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set. PLoS ONE 14(5): e0216233. https://doi.org/10.1371/journal.pone.0216233	Jiao, WB., Schneeberger, K. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nat Commun 11, 989 (2020). https://doi.org/10.1038/s41467-020-14779-y

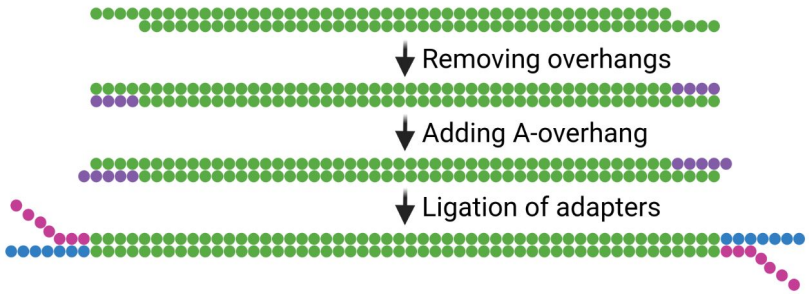
Illumina sequencing

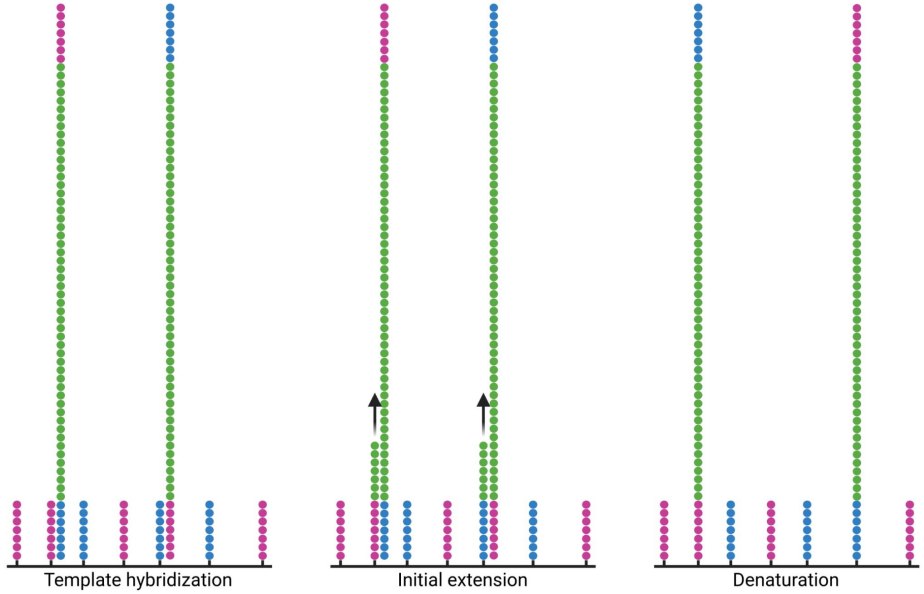
Illumina - sequencing overview



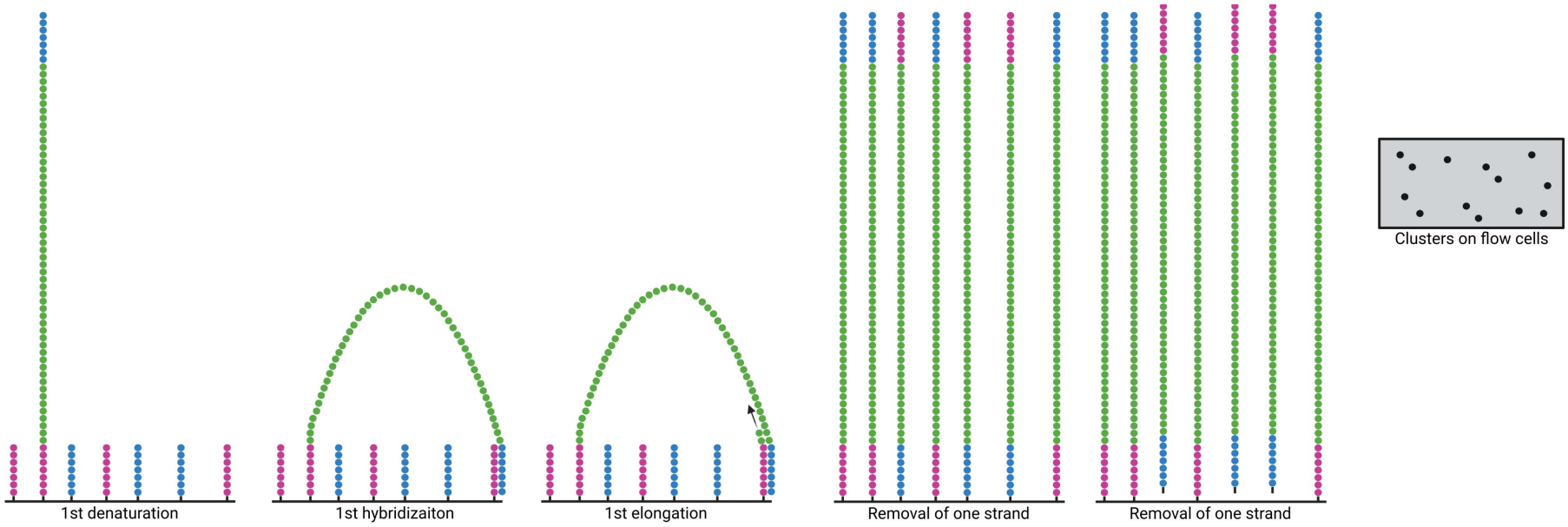


Illumina - sequencing 2

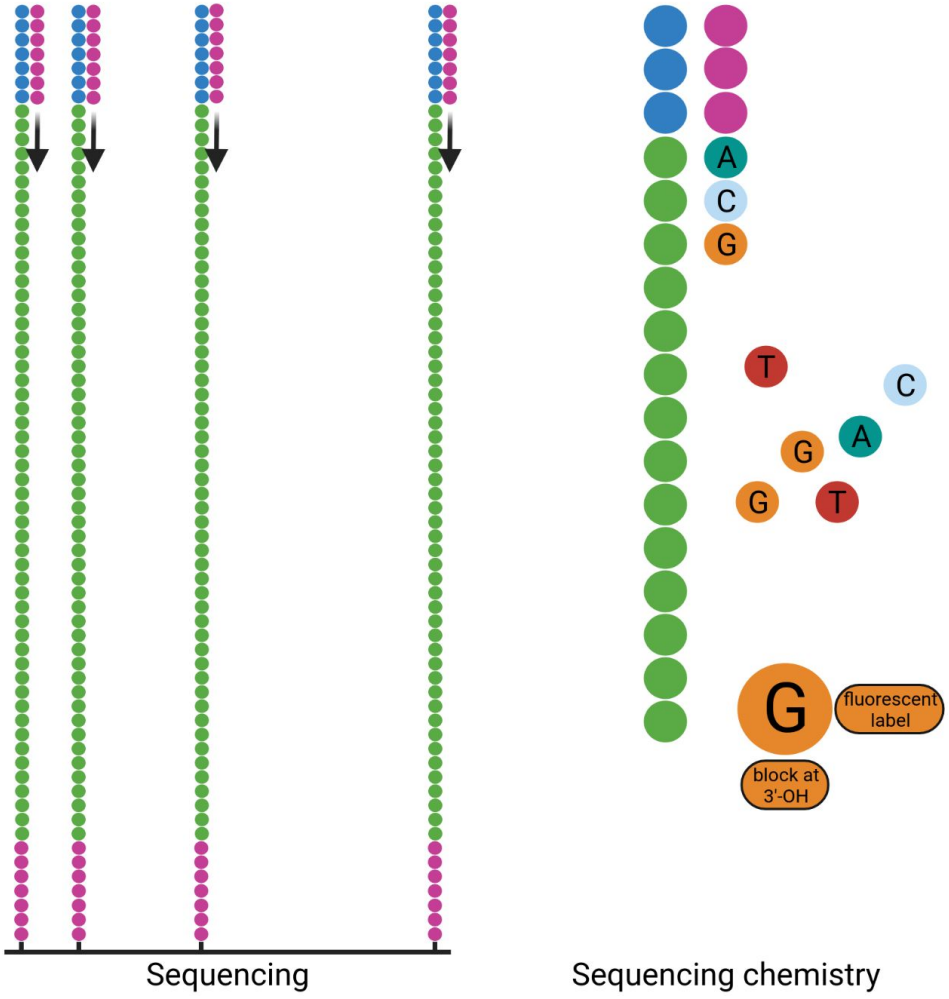




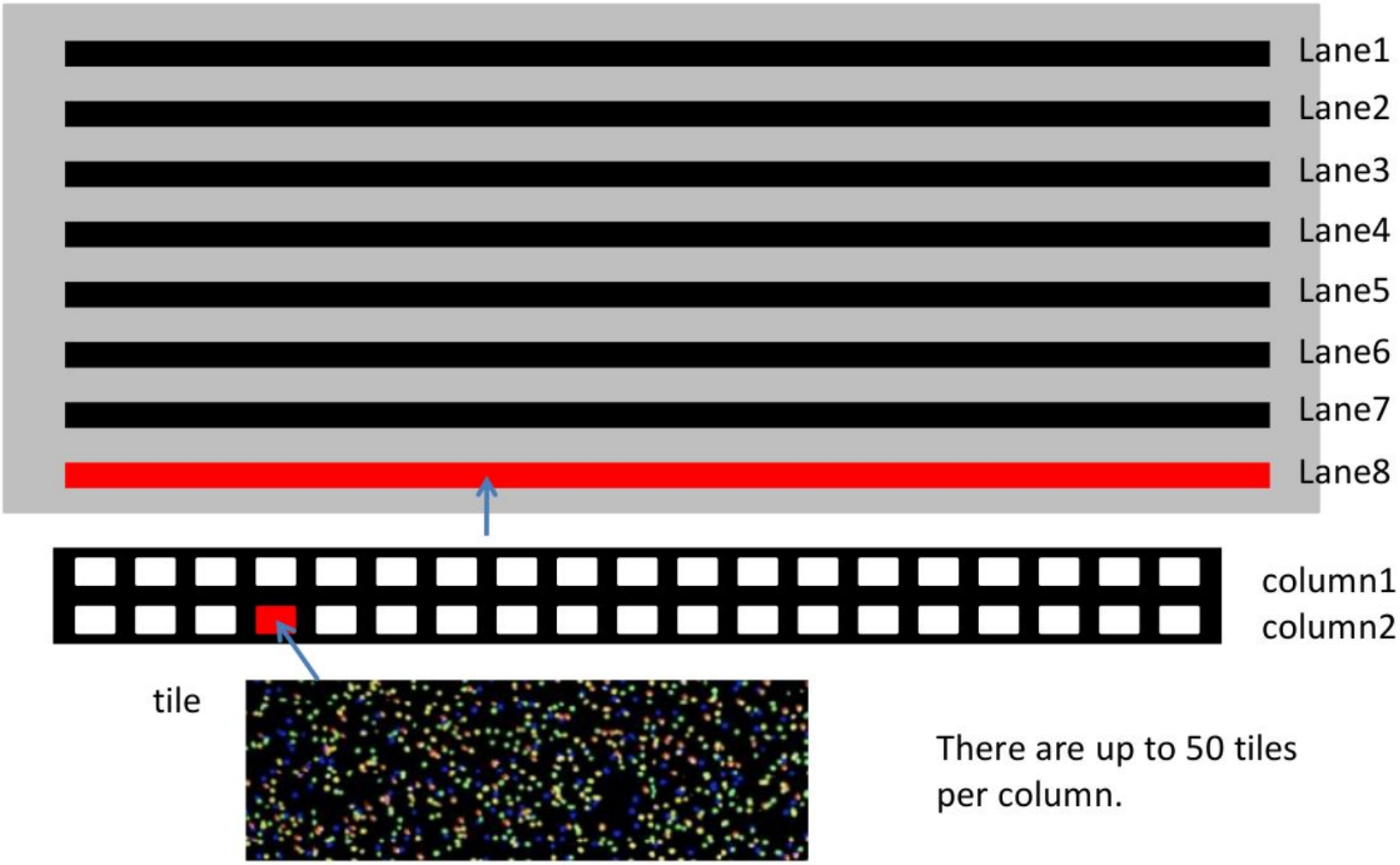
Illumina - sequencing 3



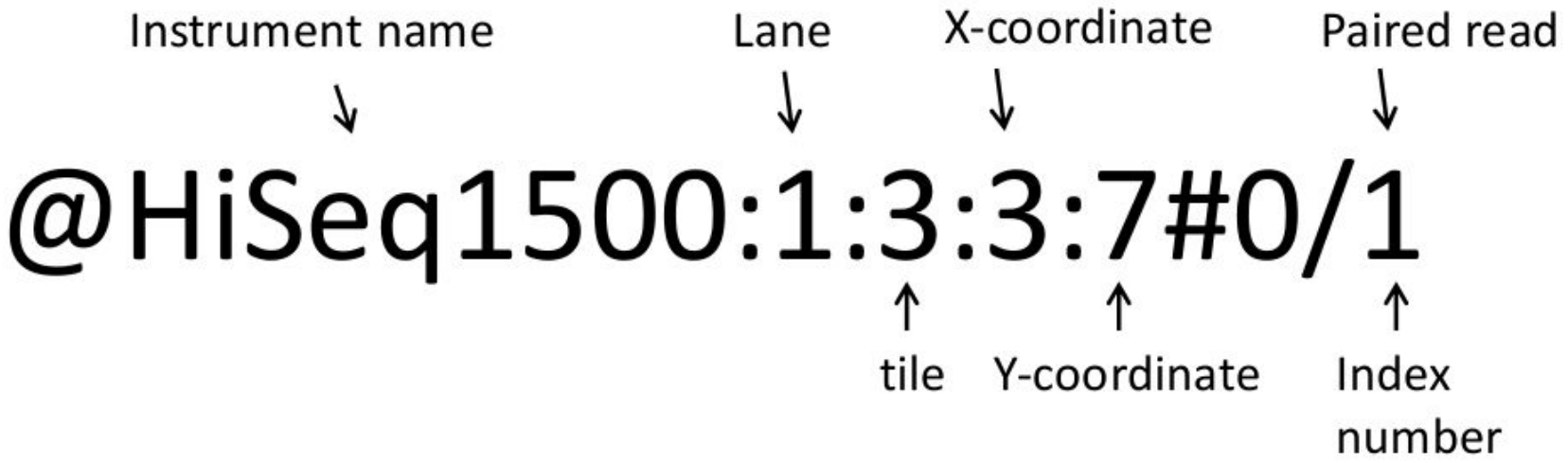
Illumina - sequencing 4



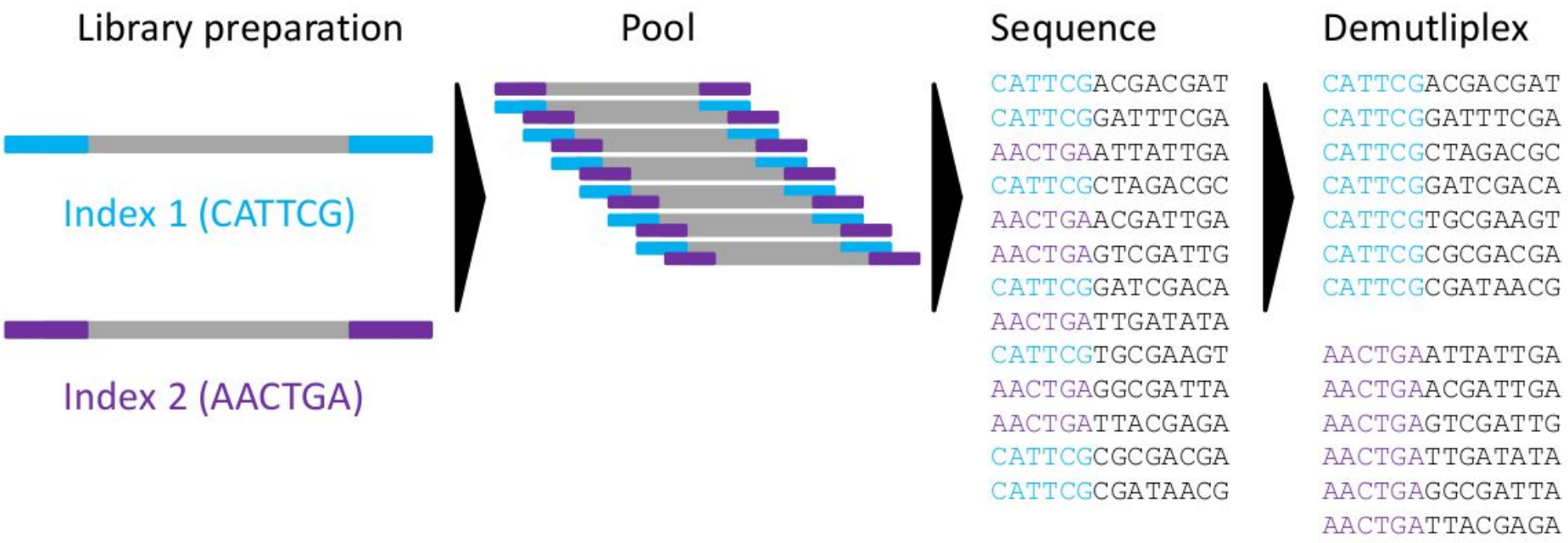
Illumina - flow cell layout



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):

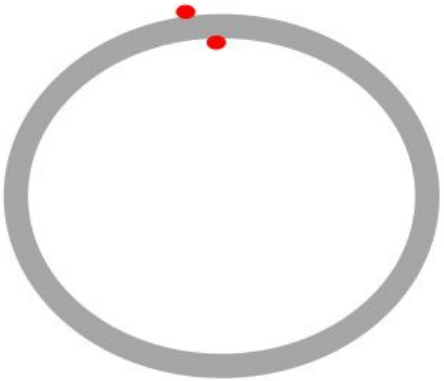
- Paired-end (PE):

Illumina - sequencing modi (mate pair)

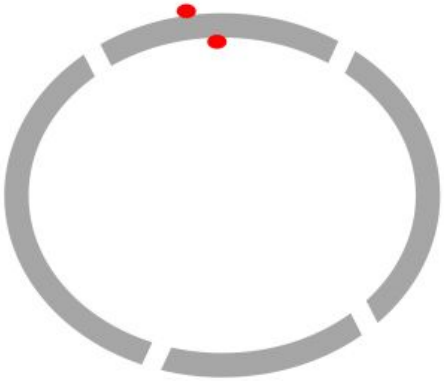
Fragmentation of DNA:

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”

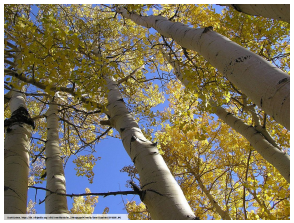
SAM/BAM

- SAM = Sequence Alignment/Map format
- BAM = Binary Alignment/Map format (binary version of SAM)
- Another way to store read information: contains information from FASTA and FASTQ file (reads mapped to reference)

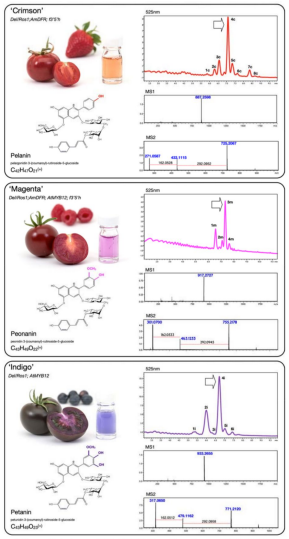
Sequencing crop genomes

- Rice (Goff et al., 2002; Yu et al., 2002)
- Poplar (Tuskan et al., 2006)
- Grapevine (Jaillon et al., 2007)
- Tomato (Sato et al., 2012)
- Sugar beet (Dohm et al., 2014)

Sanjay Acharya, https://en.wikipedia.org/wiki/Black_rice#/media/File:Black_Rice.jpg



https://en.wikipedia.org/wiki/Vitis_vinifera#/media/File:Cabernet_Sauvignon_Gaillac.jpg



https://doi.org/10.3390/horticulturae7090327

Markus Hagenlocher;
https://de.wikipedia.org/wiki/Zucker#%C3%9Cbersetzungen_in_andere_Sprachen

Resequencing projects

- 1001 genome project (*Arabidopsis thaliana*)
  - https://1001genomes.org/

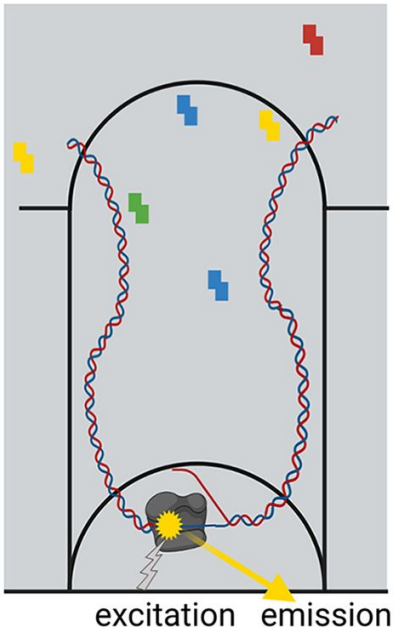
- 150 Tomato Genome ReSequencing project
  - https://www.tomatogenome.wur.nl/

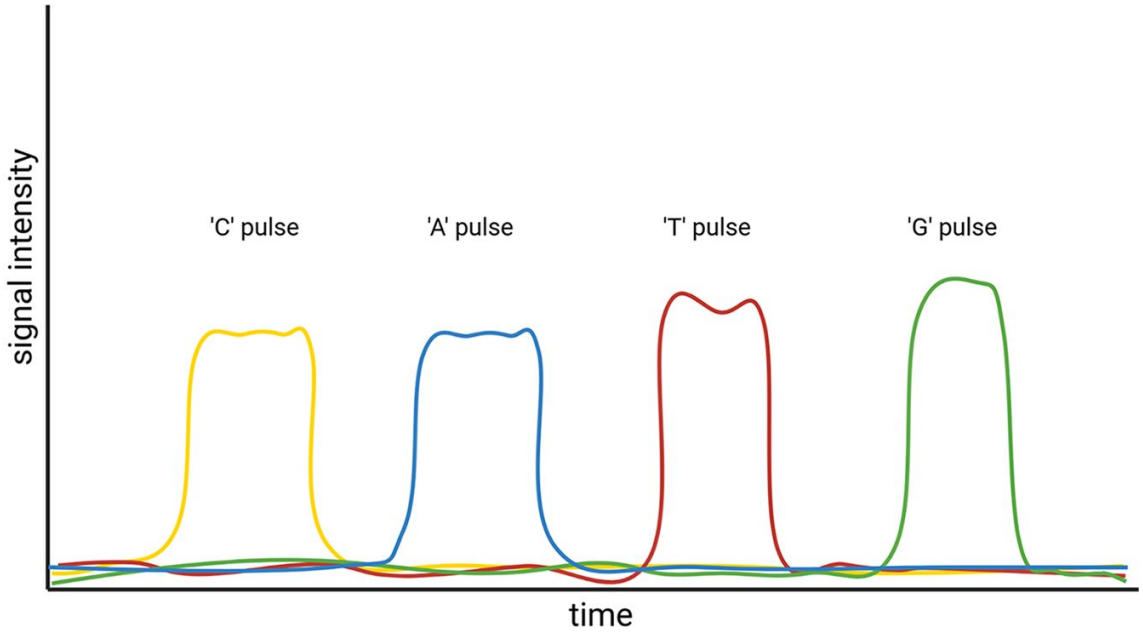
- 3,000 rice genome project
  - http://dx.doi.org/10.5524/200001

PacBio

Pacific Biosciences (PacBio)

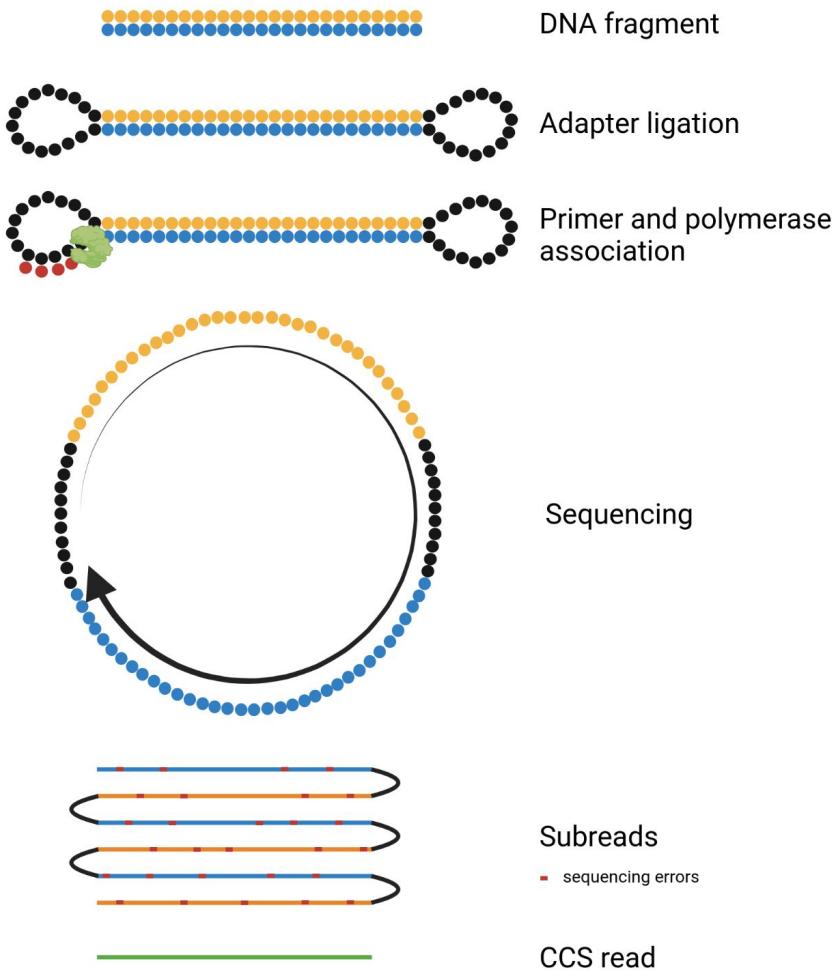
- Polymerase located at bottom of well
- ZMWs = Zero Mode Waveguide





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

PacBio - HiFi

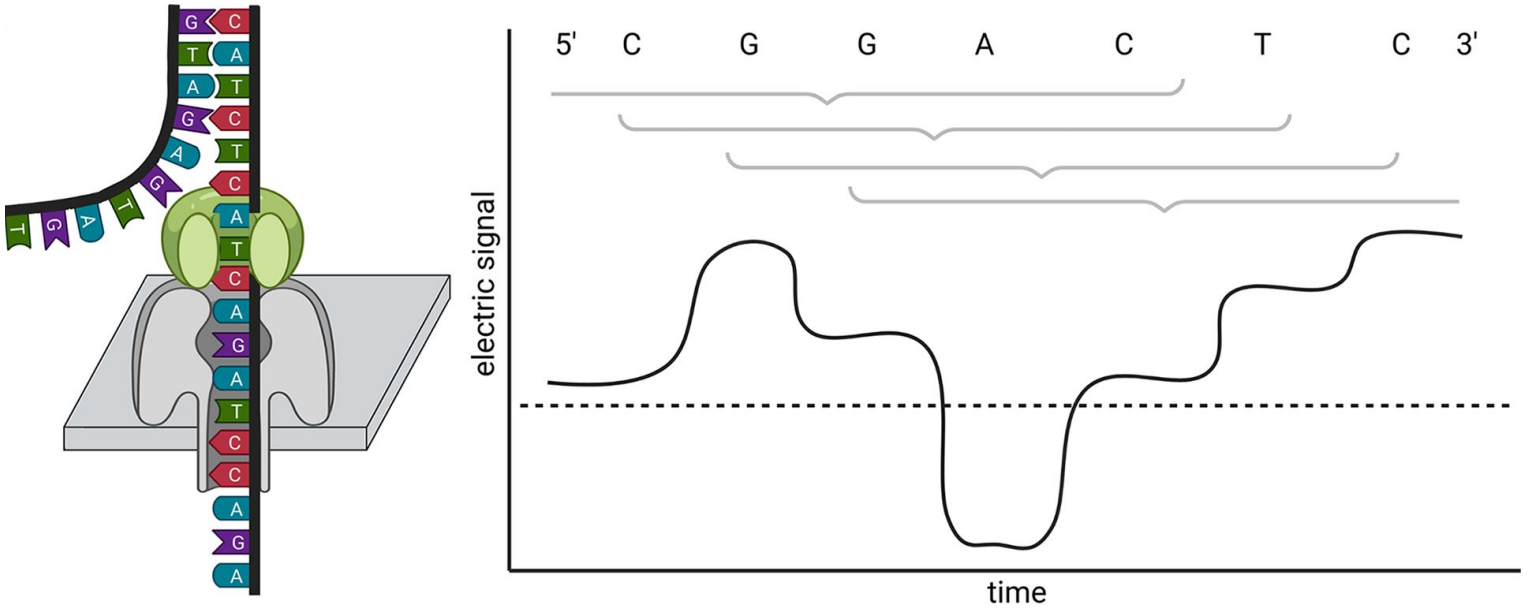

DNA fragment
Adapter ligation
Primer and polymerase association
Sequencing
Subreads
sequencing errors
CCS read

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

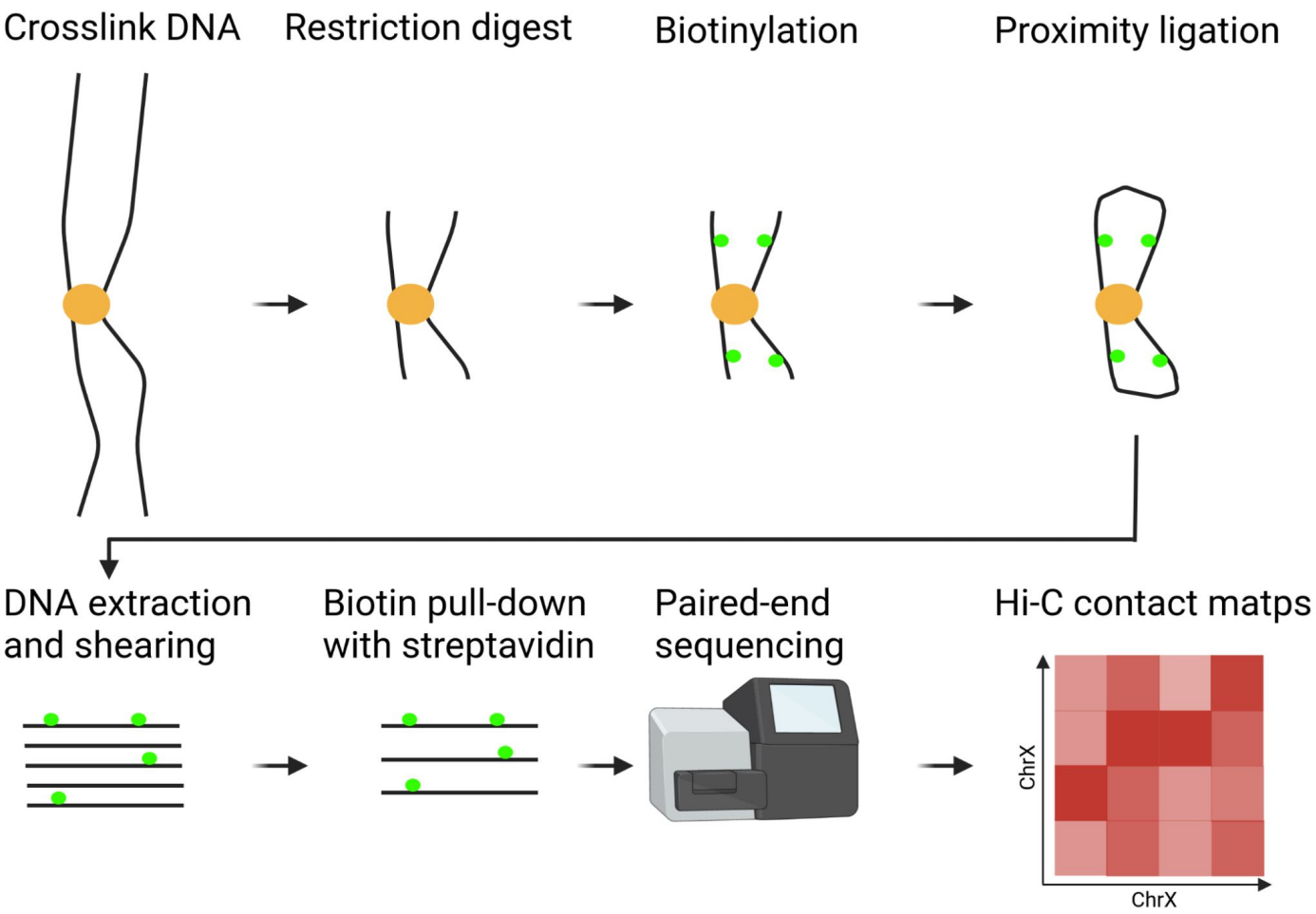
Typical results of ONT sequencing projects

- 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

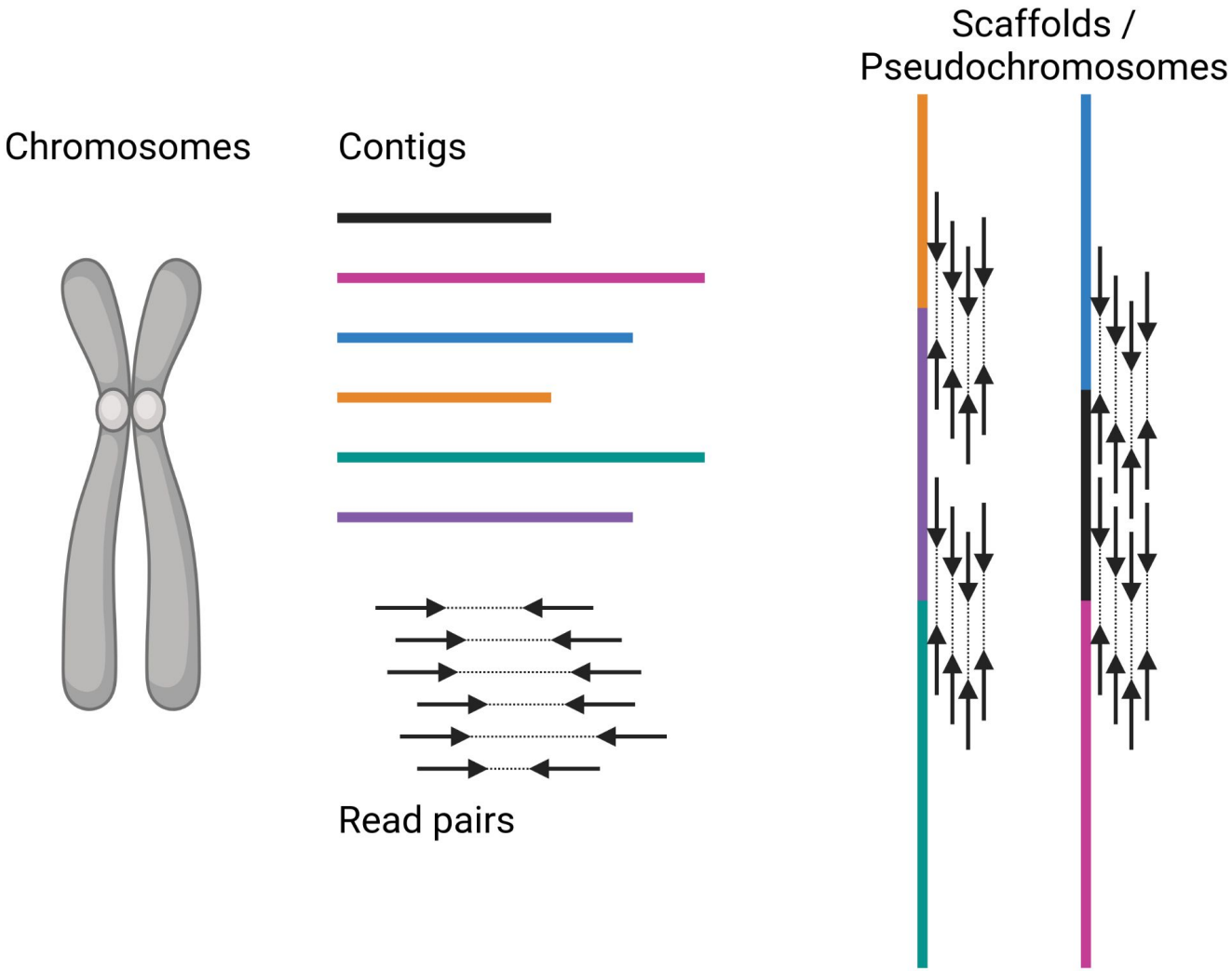
Trends in plant genomics

How can we further improve assemblies?

Hi-C (1)

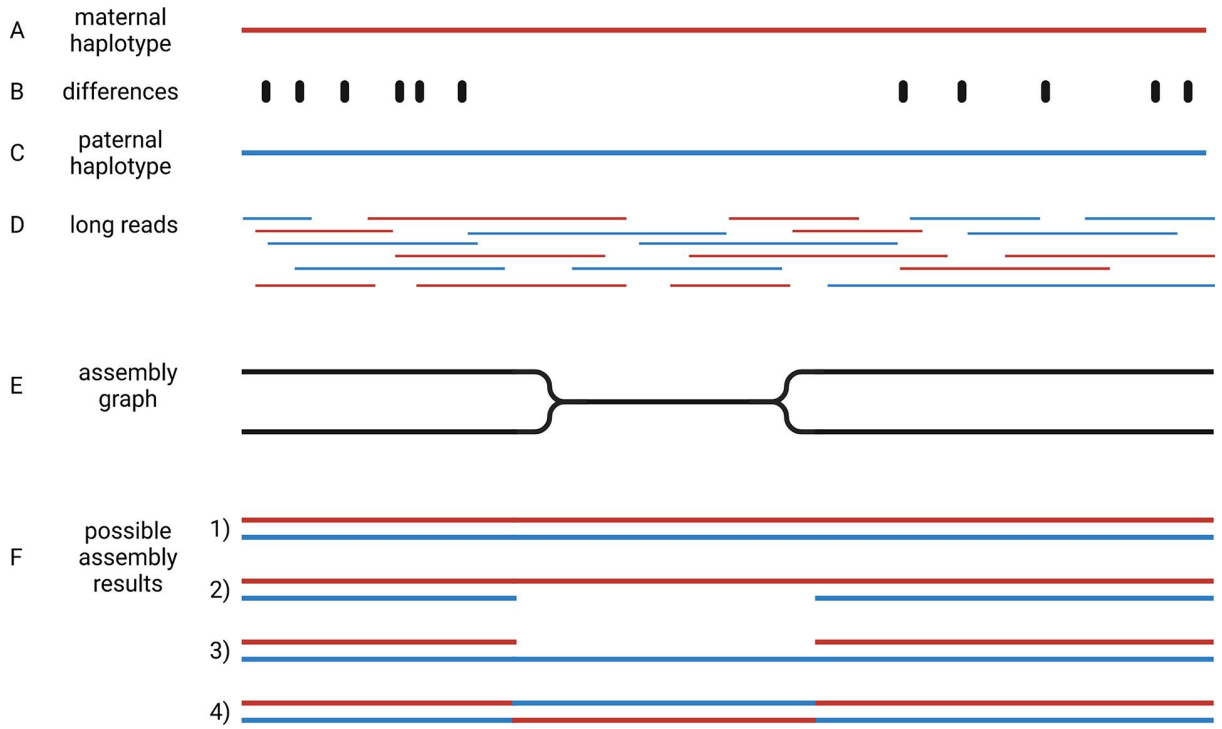

Crosslink DNA
Restriction digest
Biotinylation
Proximity ligation
DNA extraction and shearing
Biotin pull-down with streptavidin
Paired-end sequencing
Hi-C contact matps
ChrX
ChrX

Hi-C (2)


Chromosomes
Contigs
Read pairs
Scaffolds /
Pseudochromosomes

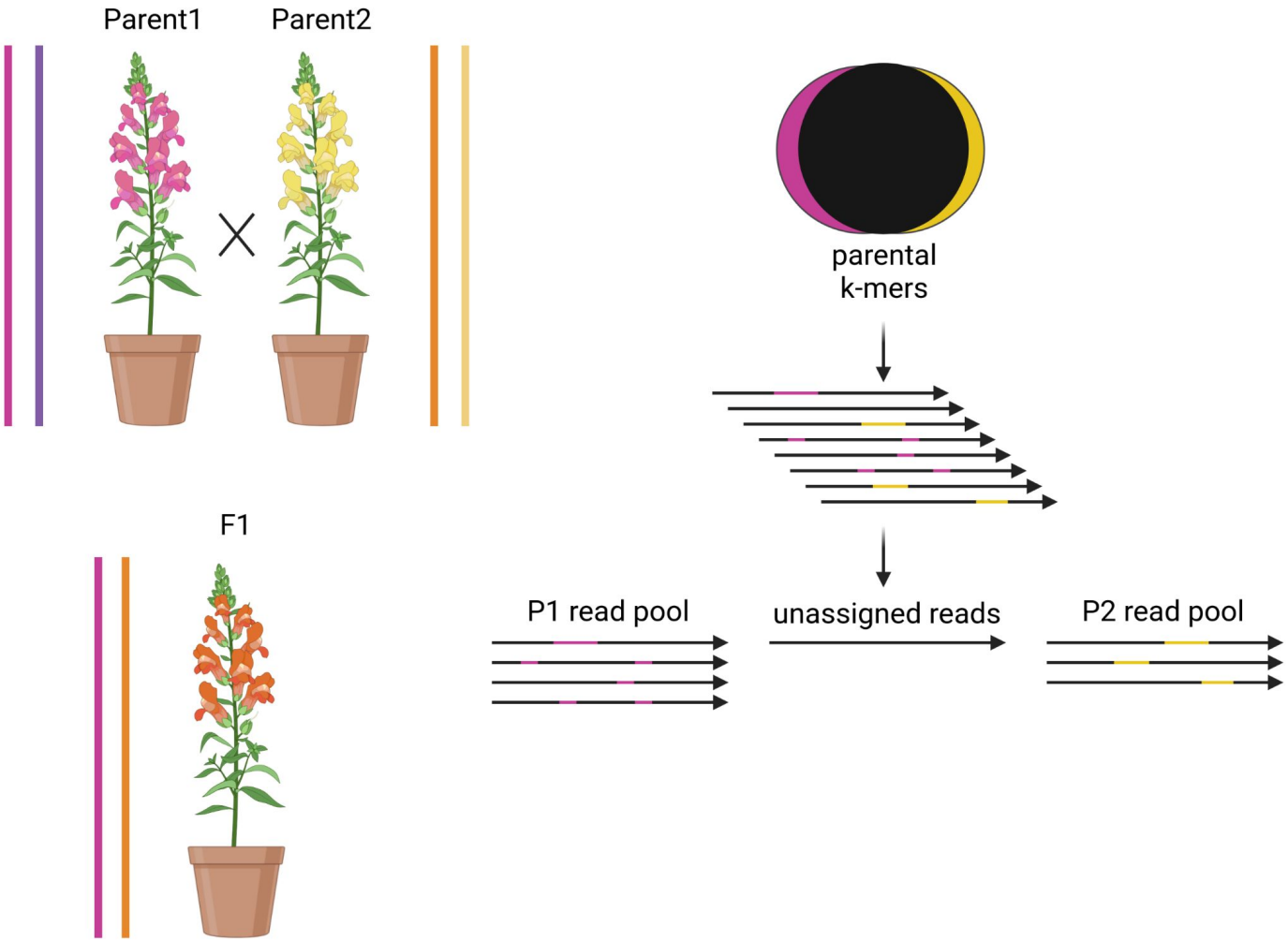
Haplophases

- Haplotype = combination of alleles
- Haplophase = representation of a haplotype



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

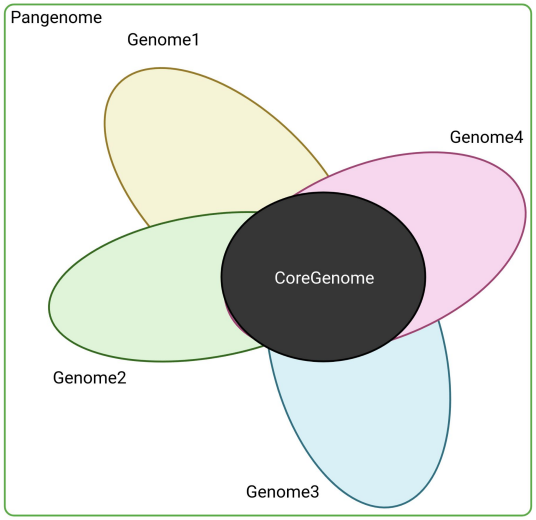
TrioBinning

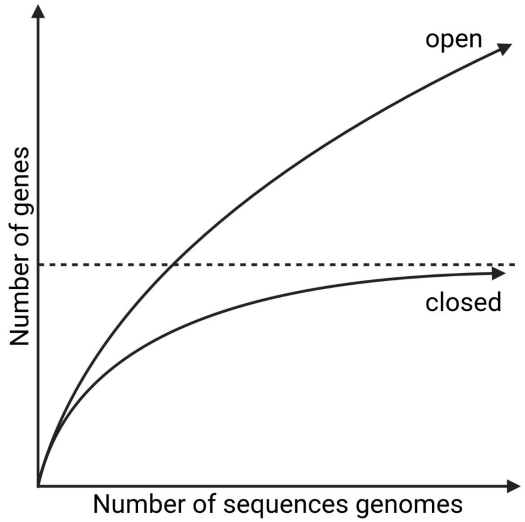


Koren et al., 2018: 10.1038/nbt.4277

Pangenomics

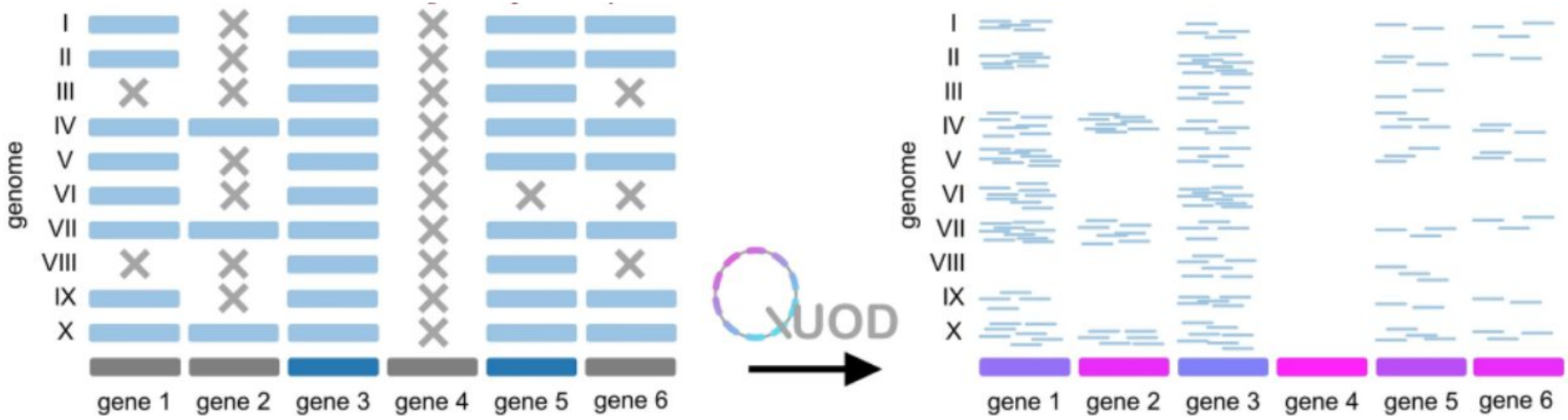
- Pangenome = combination of all genomes in a taxon
- Core genome = present in all species of a taxon
- Shell genome = present in many species of a taxon
- Cloud genome = only present in individual species





Pangenomics

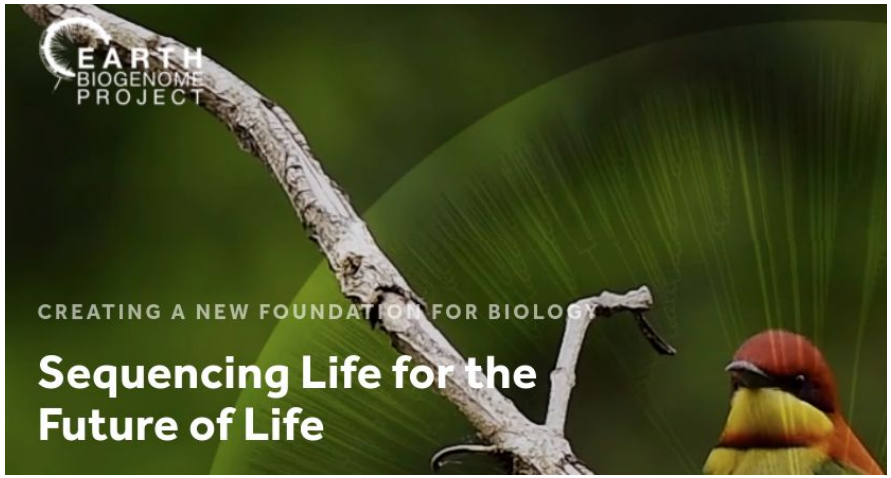
- Core genes = genes are present in all species of a taxon
- Conditionally dispensable genes = genes are present in many, but not in all species
- Dispensable genes = genes only present in a few species of a taxon



Sielemann et al., 2021: 10.1186/s13007-021-00718-5

Sequencing the genomes of all plants (and animals)

- Darwin Tree of Life (Darwin Tree of Life Project, 2021)
- Earth BioGenome Project (Lewin et al., 2018)
- European Research Genome Atlas (ERGA; https://www.erga-biodiversity.eu/)



https://www.darwintreeoflife.org/
https://www.earthbiogenome.org/
https://www.erga-biodiversity.eu/

Democratization of genomics

- Sequencing orphan crop genomes
- Portable and affordable sequencers
- Genome projects conducted by individual labs (not just sequencing centers)

Time for questions!

Questions

1. How does Sanger sequencing work?
2. What are the important steps of the Illumina sequencing workflow?
3. How does PacBio sequencing work?
4. What are the important steps of an ONT sequencing workflow?
5. What is a pangenome?
6. What is TrioBinning?
7.