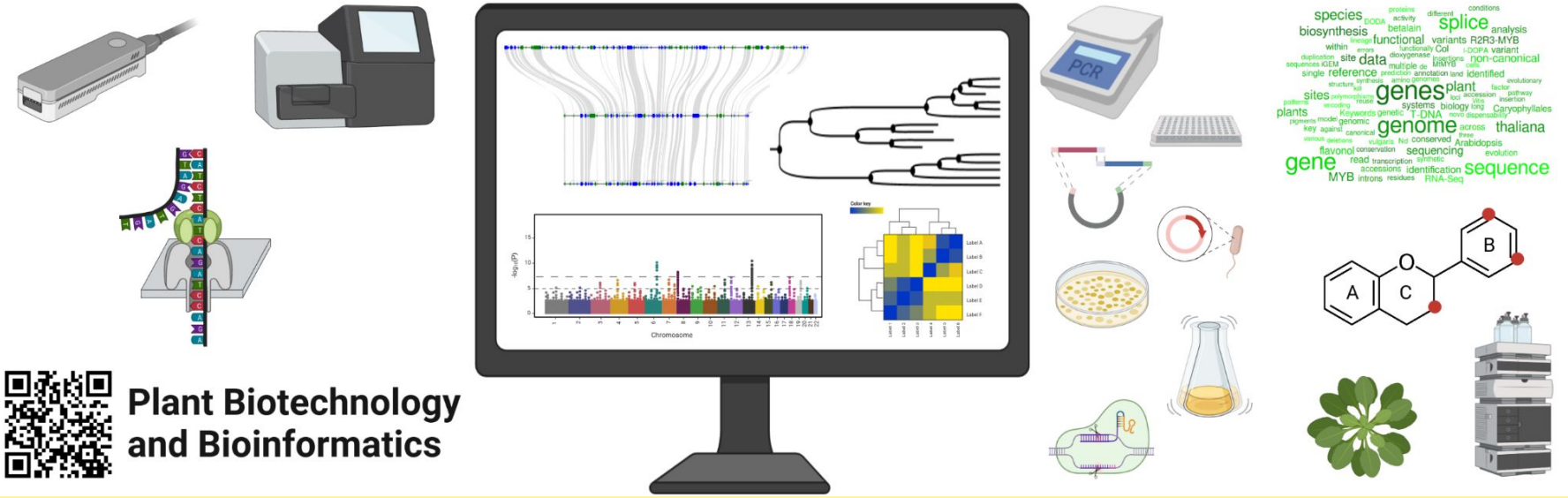
Technische Universität Braunschweig



Plant Biotechnology and Bioinformatics

Genomics #1 - Big Data Analytics in Life Sciences

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

Outline

- Sequencing technologies:
  - Sanger
  - Illumina
  - **ONT**
  - PacBio

- Genome sequence assembly

Why sequence a plant genome?

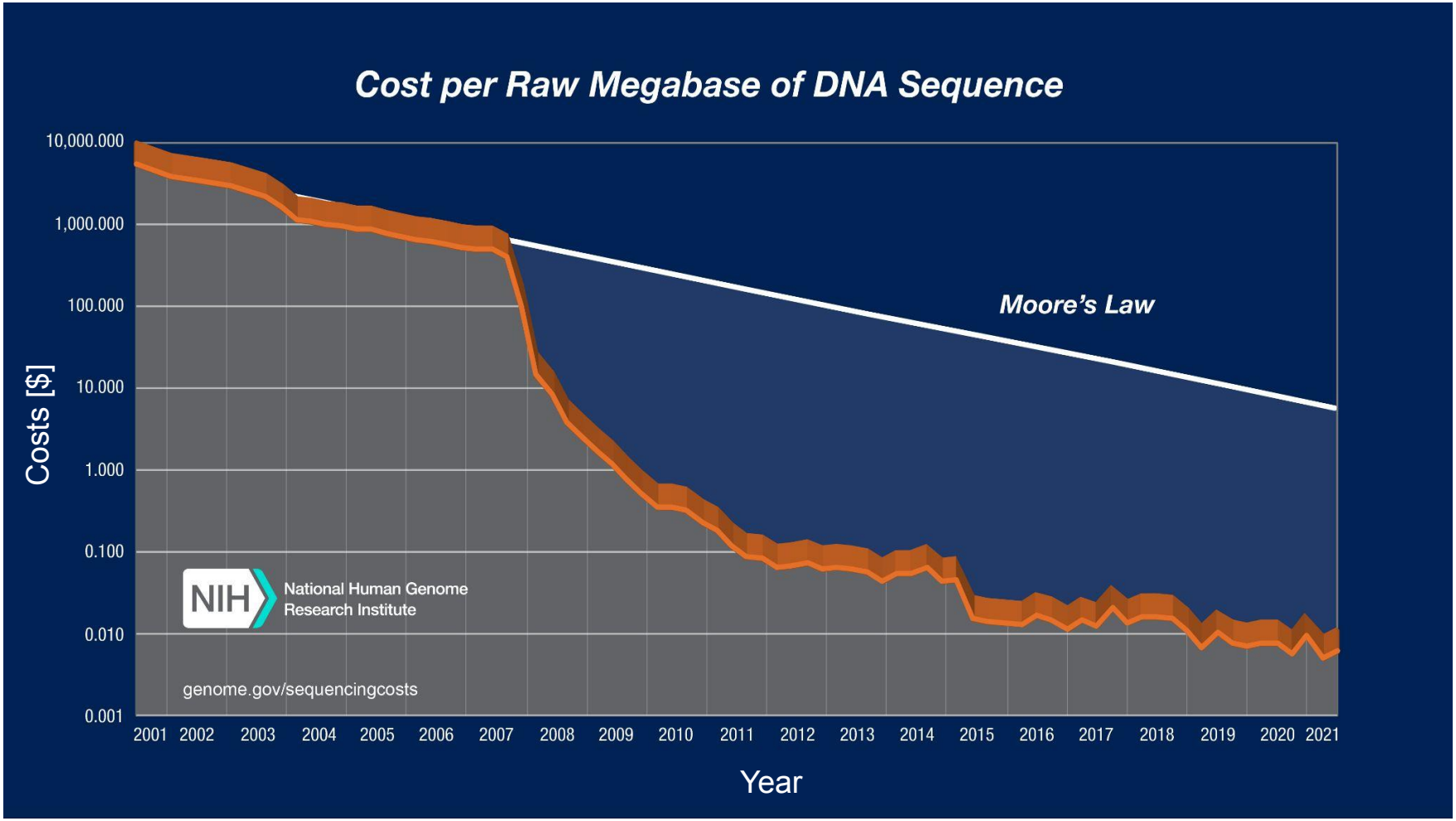
- Primer design (banana)
- Assessment of genetic diversity (population genomics)
- Trait discovery (yam)
- Breeding (sugar beet)
- Understanding the evolution (At7)

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

Overview of sequencing technologies (selection)

- 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?

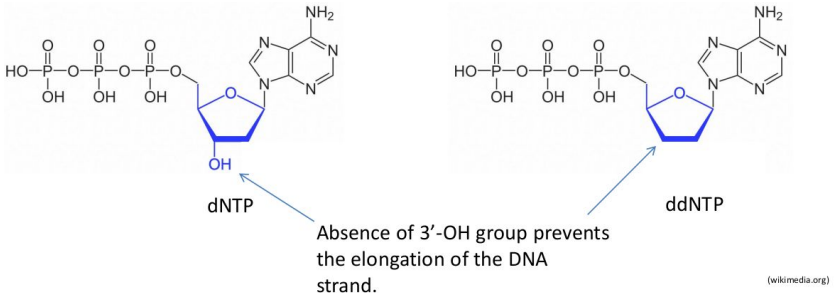
Development of sequencing capacity



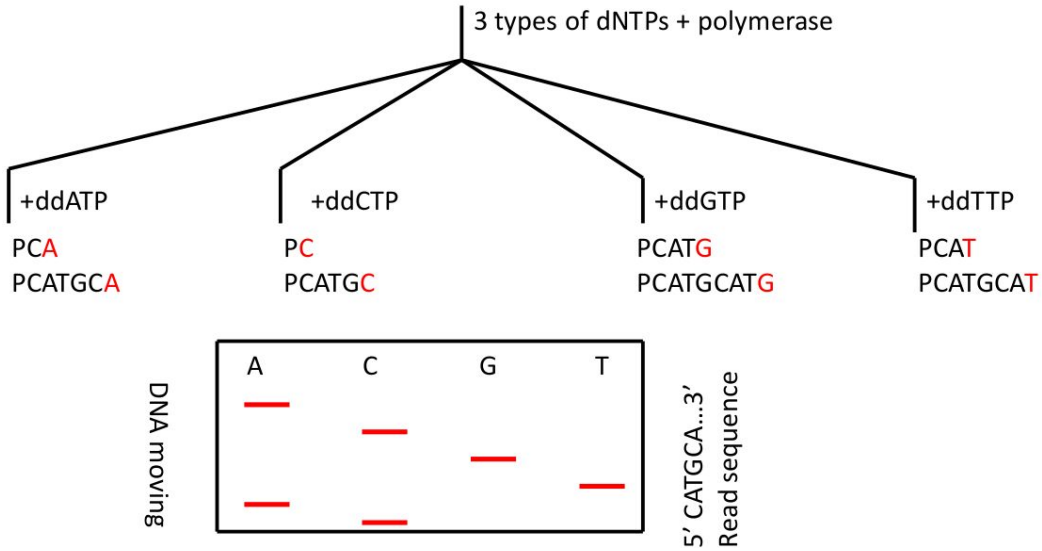
https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Sanger sequencing

Concept of Sanger sequencing

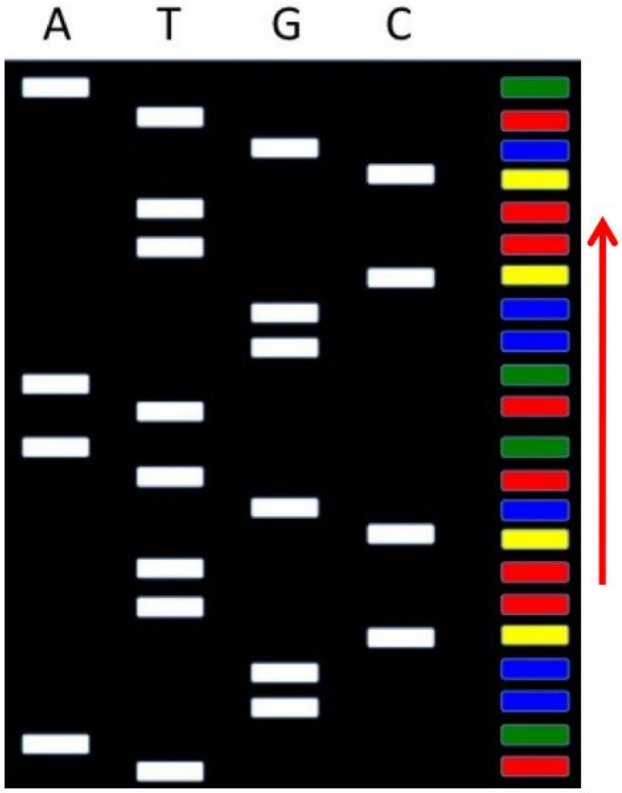


Primer (P): 5' –TGCATGGCATGATGCATG-3'
Template: 3' –ACGTACCGTACTACGTACGTACGTACGTCTAGGT-5'



Sanger sequencing - original version

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



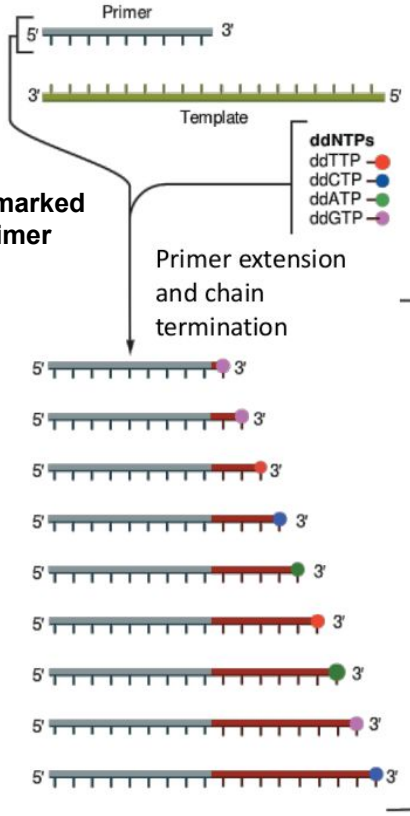
(modified from wikimedia.org)

Sanger sequencing - today

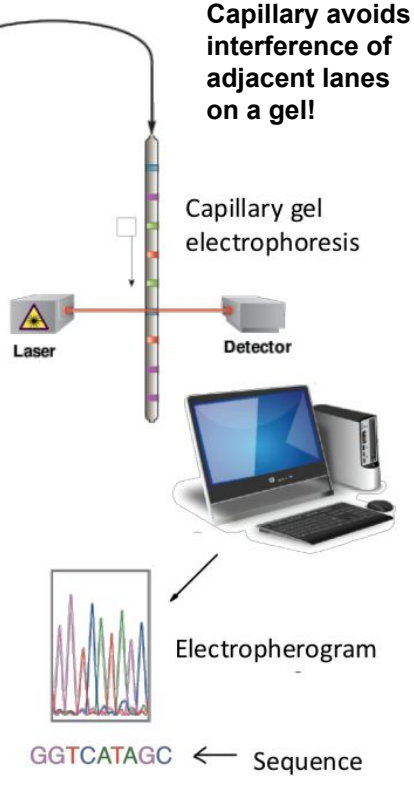
Only one reaction!

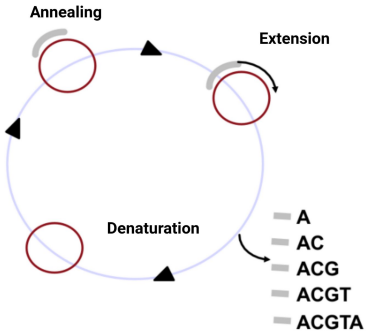
ddNTPs are marked instead of primer

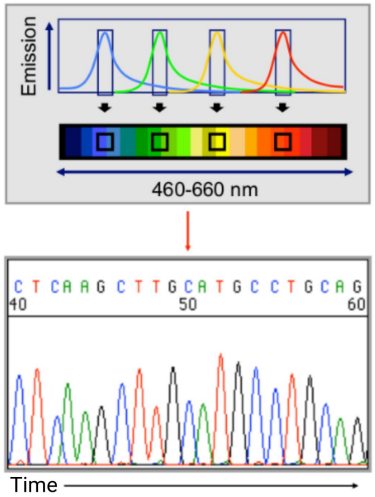
Low input required due to cycle sequencing



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%

ASCII code

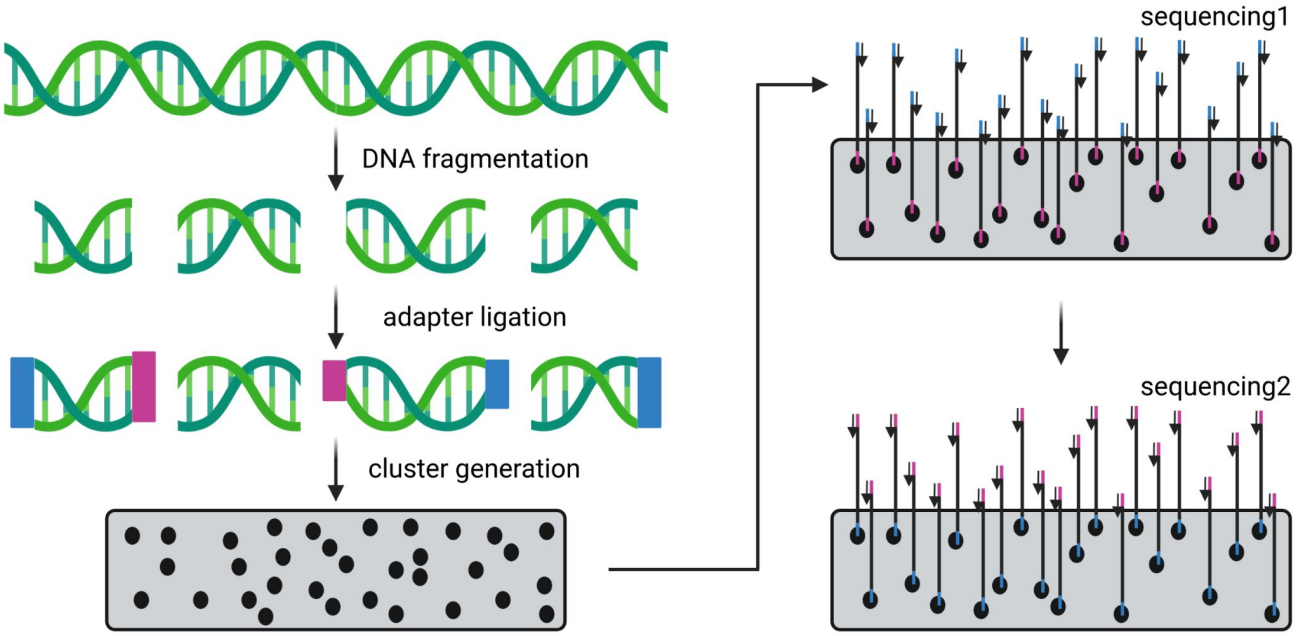
- Phred score encoded with ASCII
- Phred32, phred64 (different offset values to avoid special characters)

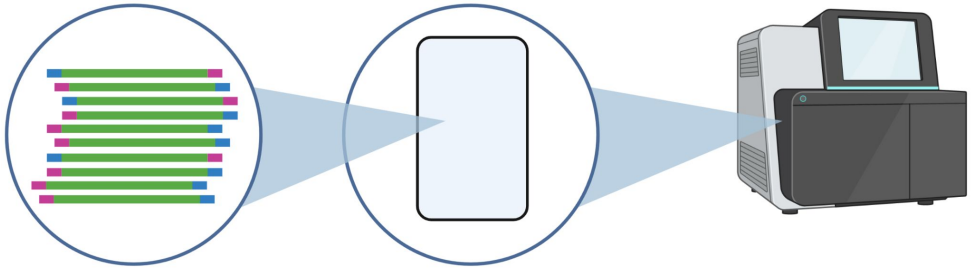
Binary	Oct	Dec	Hex	Glyph		
				1963	1965	1967
010 0000	040	32	20	space		
010 0001	041	33	21	!		
010 0010	042	34	22	"		
010 0011	043	35	23	#		
010 0100	044	36	24	$		
010 0101	045	37	25	%		
010 0110	046	38	26	&		
010 0111	047	39	27	'		
010 1000	050	40	28	(		
010 1001	051	41	29	)		
010 1010	052	42	2A	*		
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	

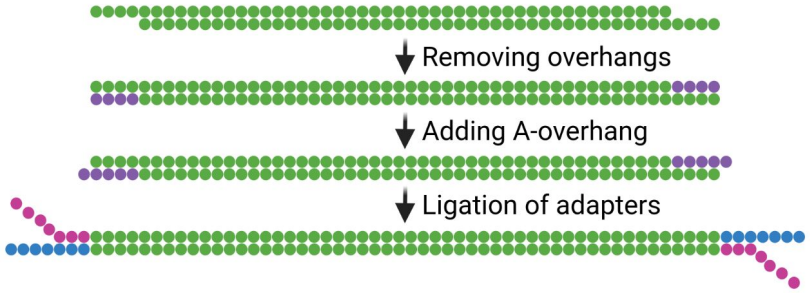
Illumina sequencing

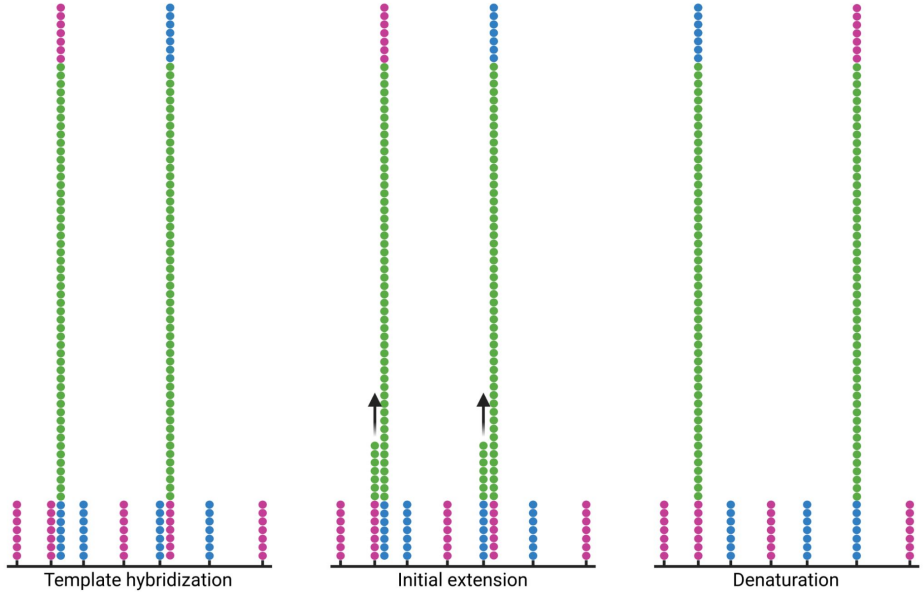
Illumina - sequencing overview



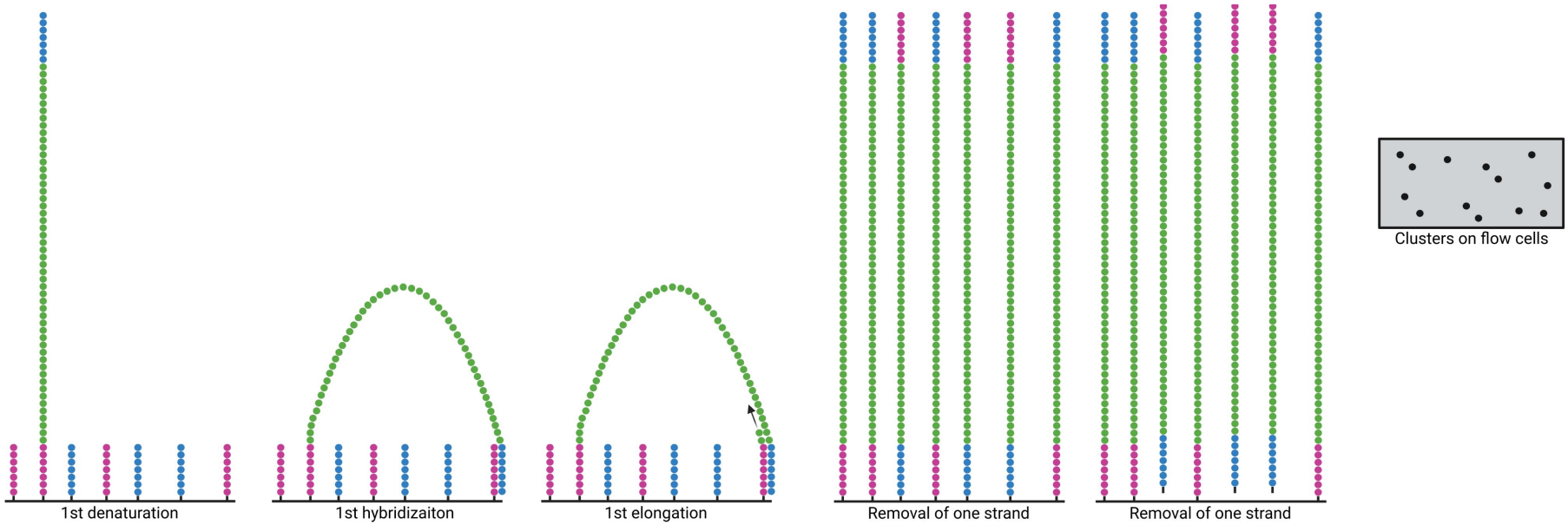


Illumina - sequencing 2

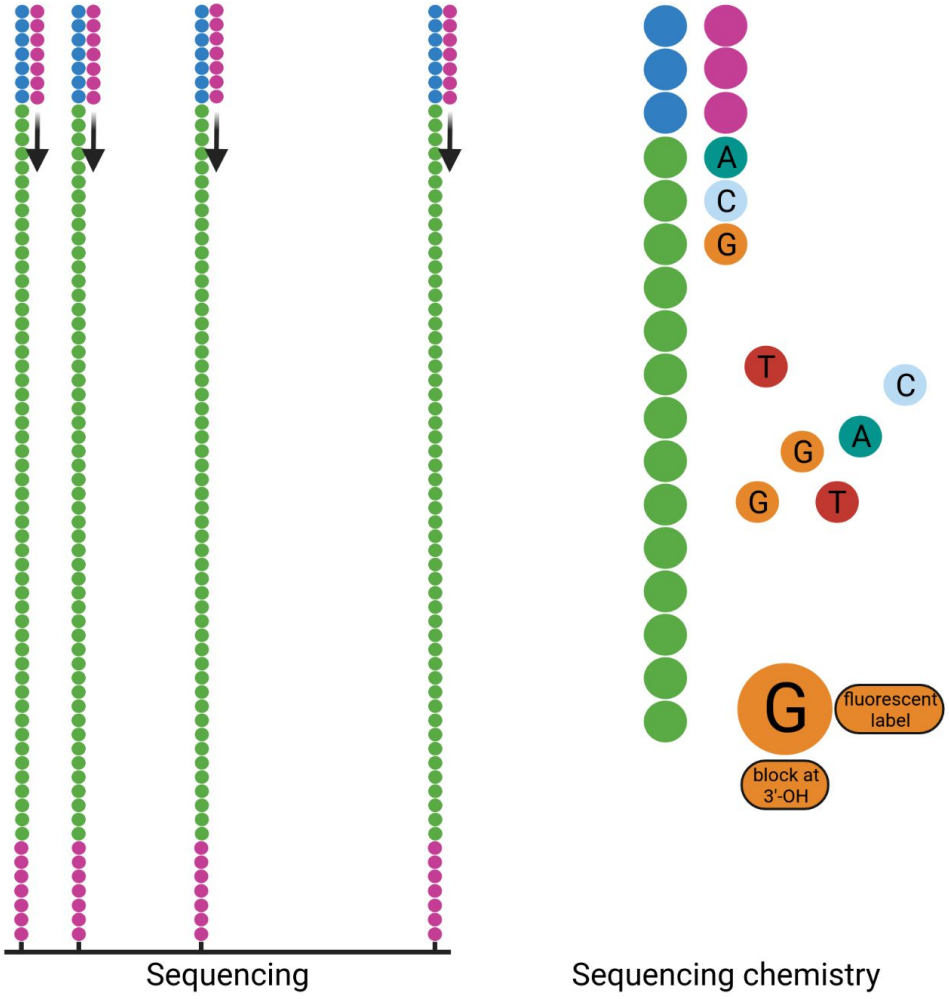




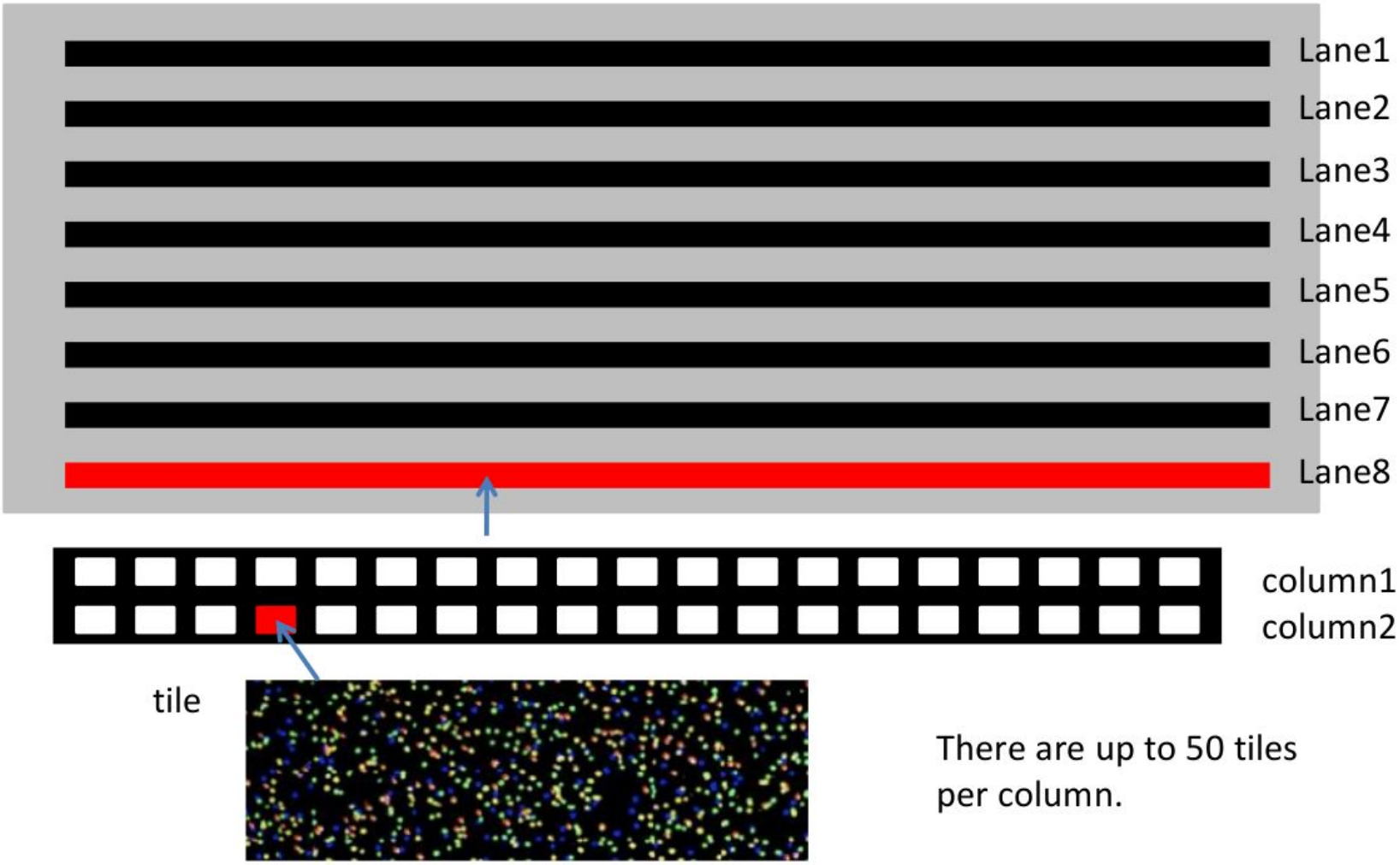
Illumina - sequencing 3



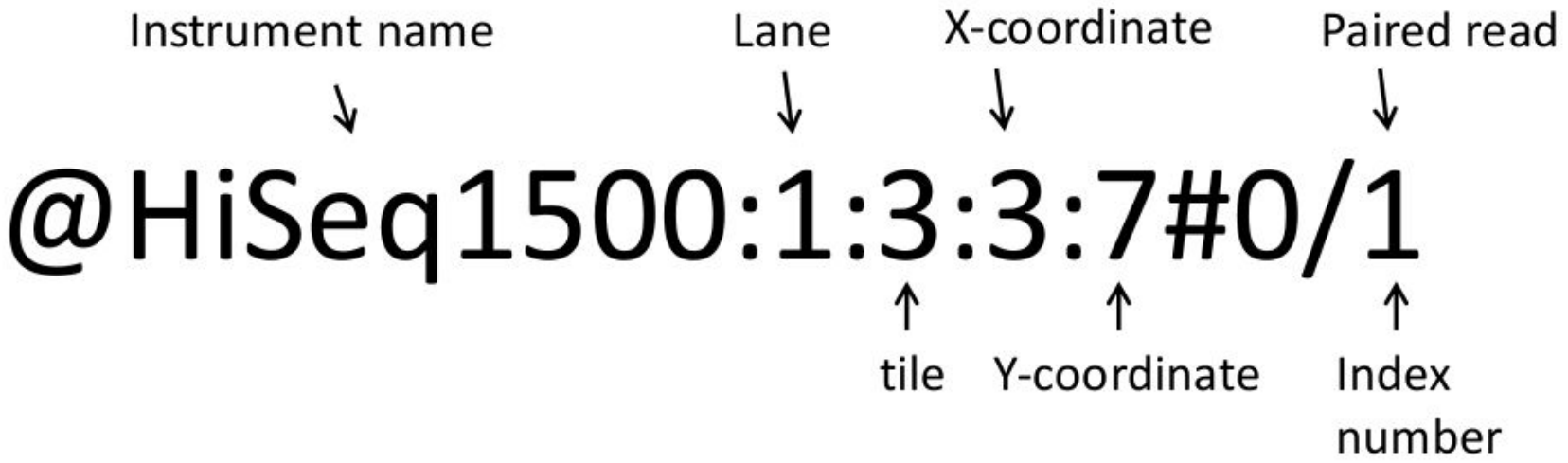
Illumina - sequencing 4



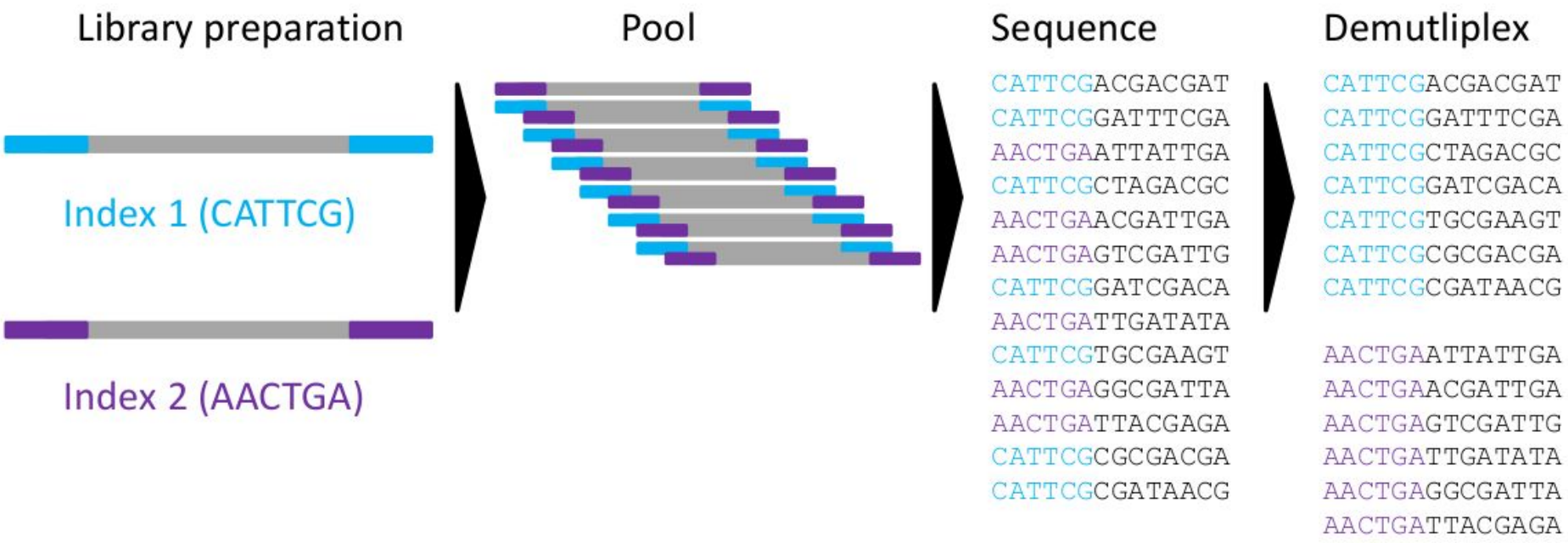
Illumina - flow cell layout



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):

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

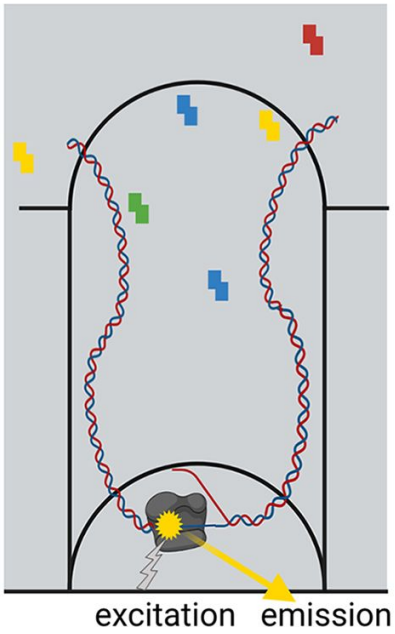
- Example:

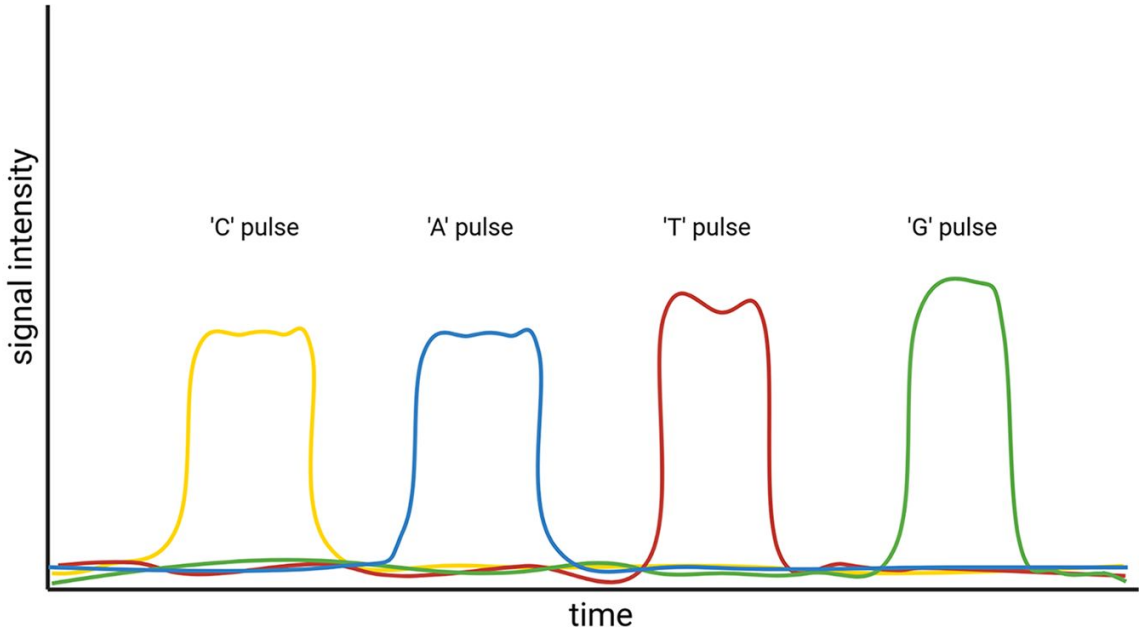
  @seq1
  ACGTACGTACGT
  +
  “”?CB””:DC”

PacBio

Pacific Biosciences (PacBio)

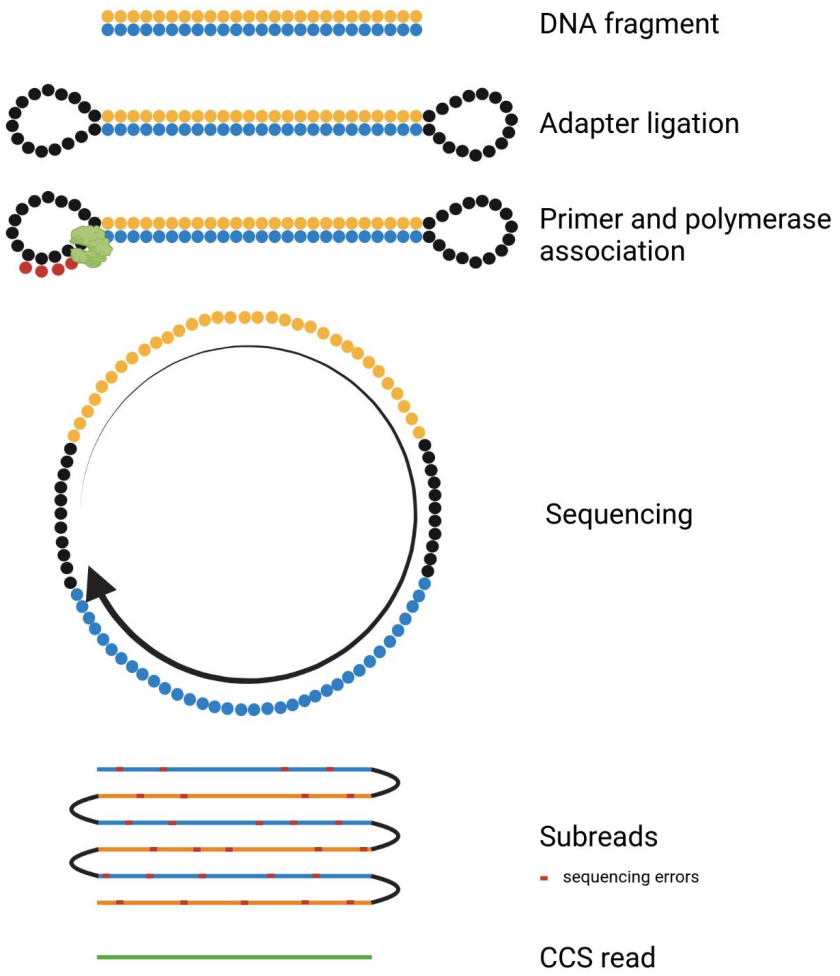
- Polymerase located at bottom of well
- ZMWs = Zero Mode Waveguide





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

PacBio - HiFi

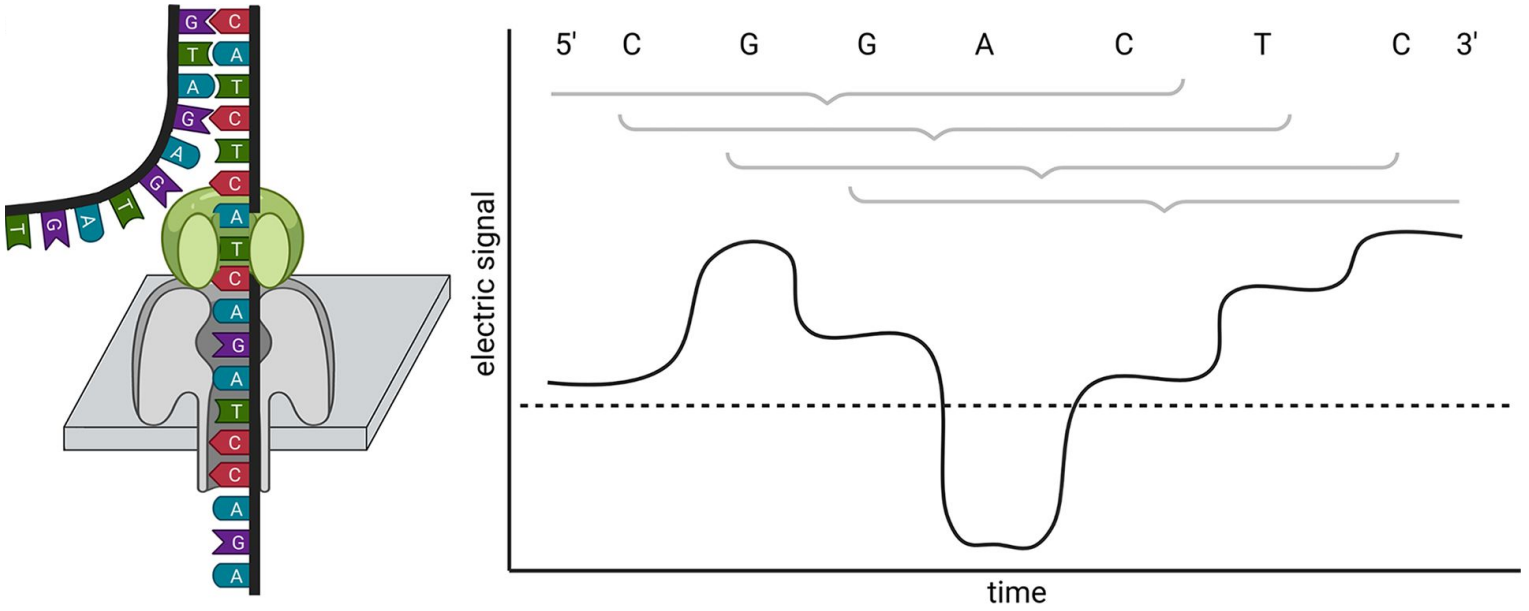

DNA fragment
Adapter ligation
Primer and polymerase association
Sequencing
Subreads
sequencing errors
CCS read

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		

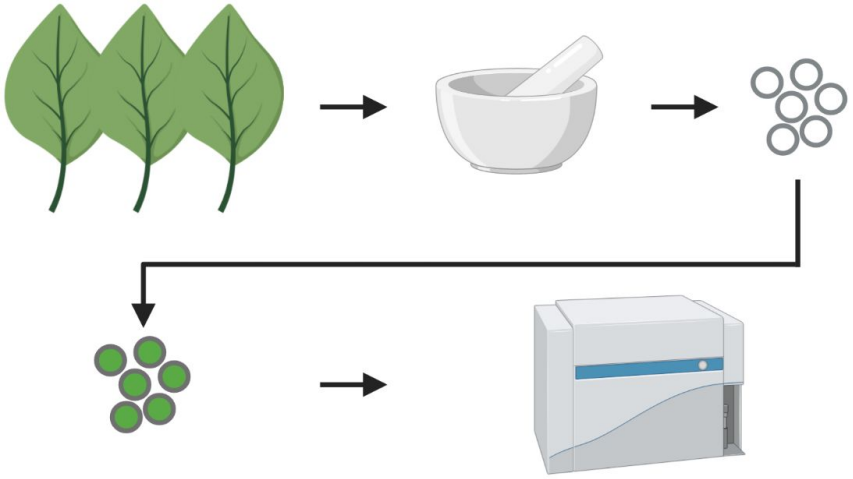
Pucker et al., 2022: 10.1017/qpb.2021.18

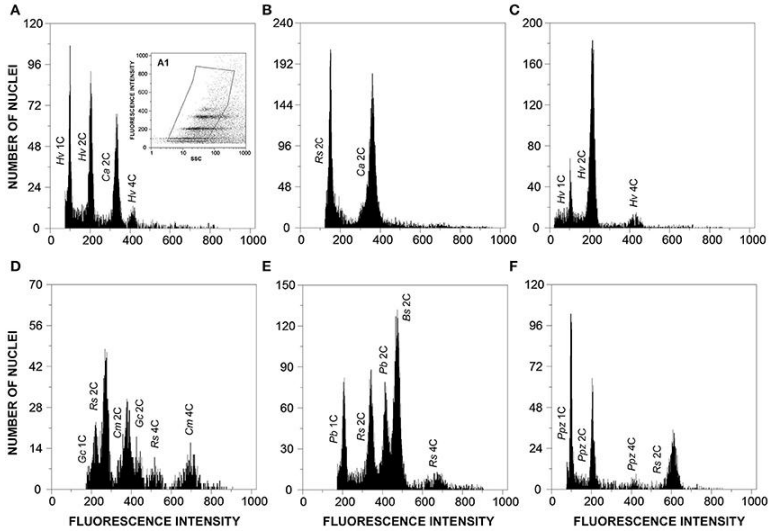
Considerations & preparations

- 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/



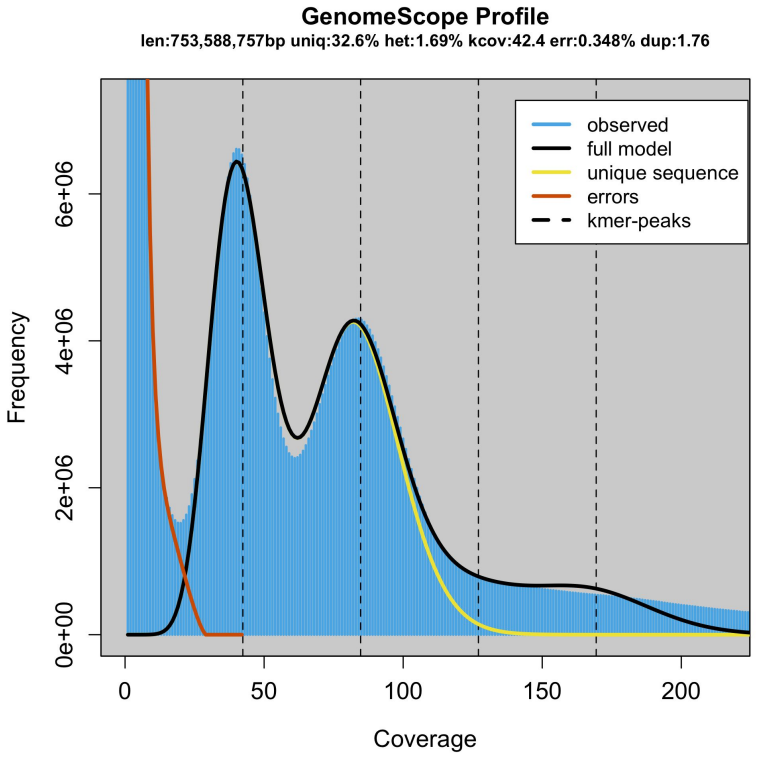


https://doi.org/10.3389/fpls.2014.00422

Genome size - 2

- Tools for genome size estimation based on short reads
  - K-mer-based: GenomeScope2, findGSE, gce

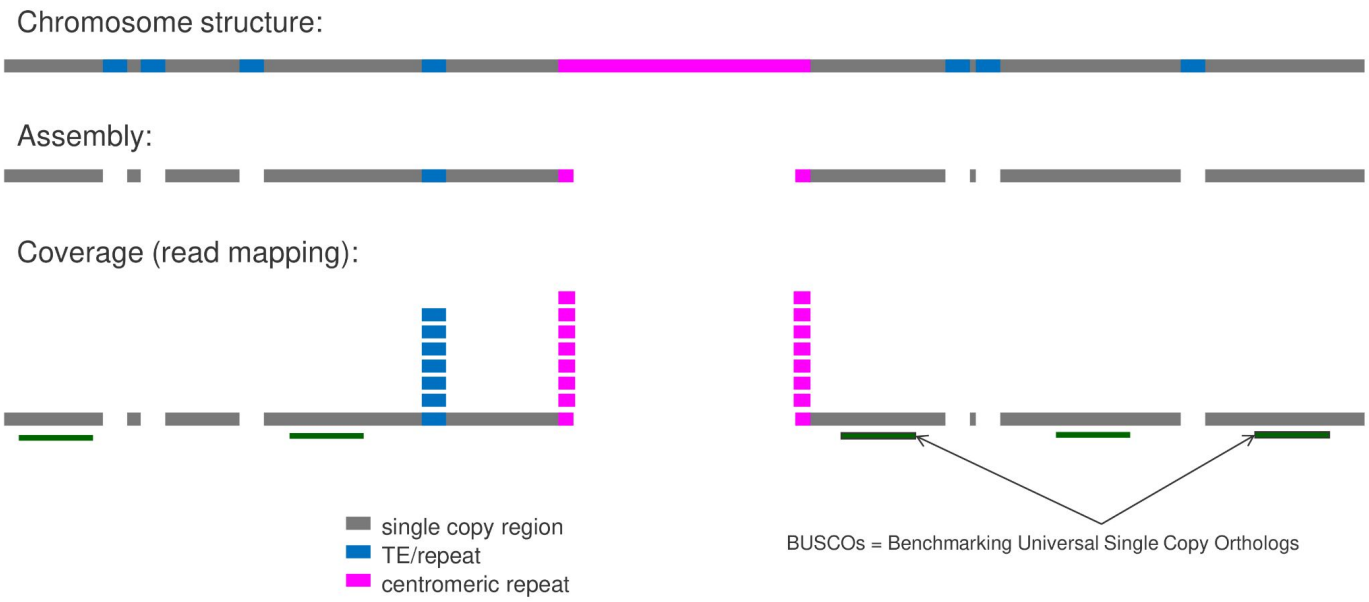
```
ACGAAGCCATAT
ACGAAGC
 CGAAGCC
  GAAGCCA
   AAGCCAT
    AGCCATAT
```



https://github.com/schatzlab/genomescope

Genome size - 3

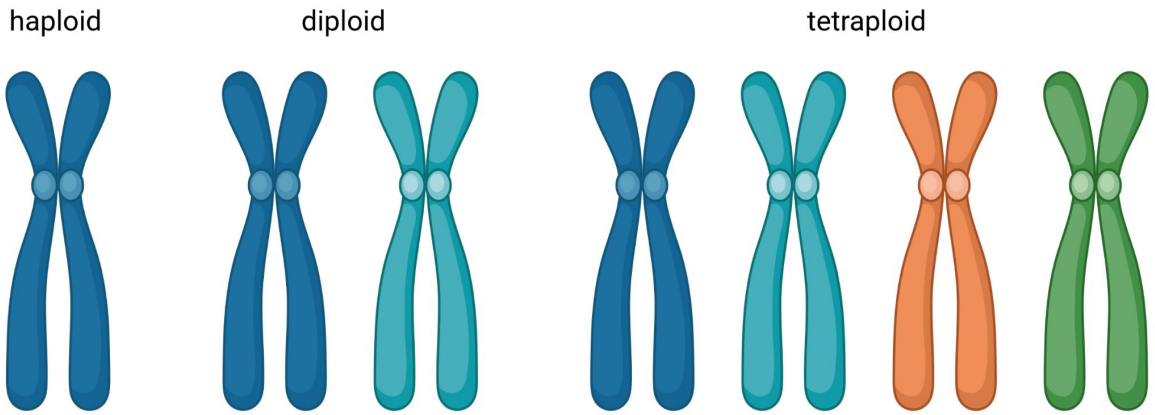
- Tools for genome size estimation based on short/long reads
  - Mapping-based: MGSE, Gnodes



Pucker, 2019: 10.1101/607390

Ploidy

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



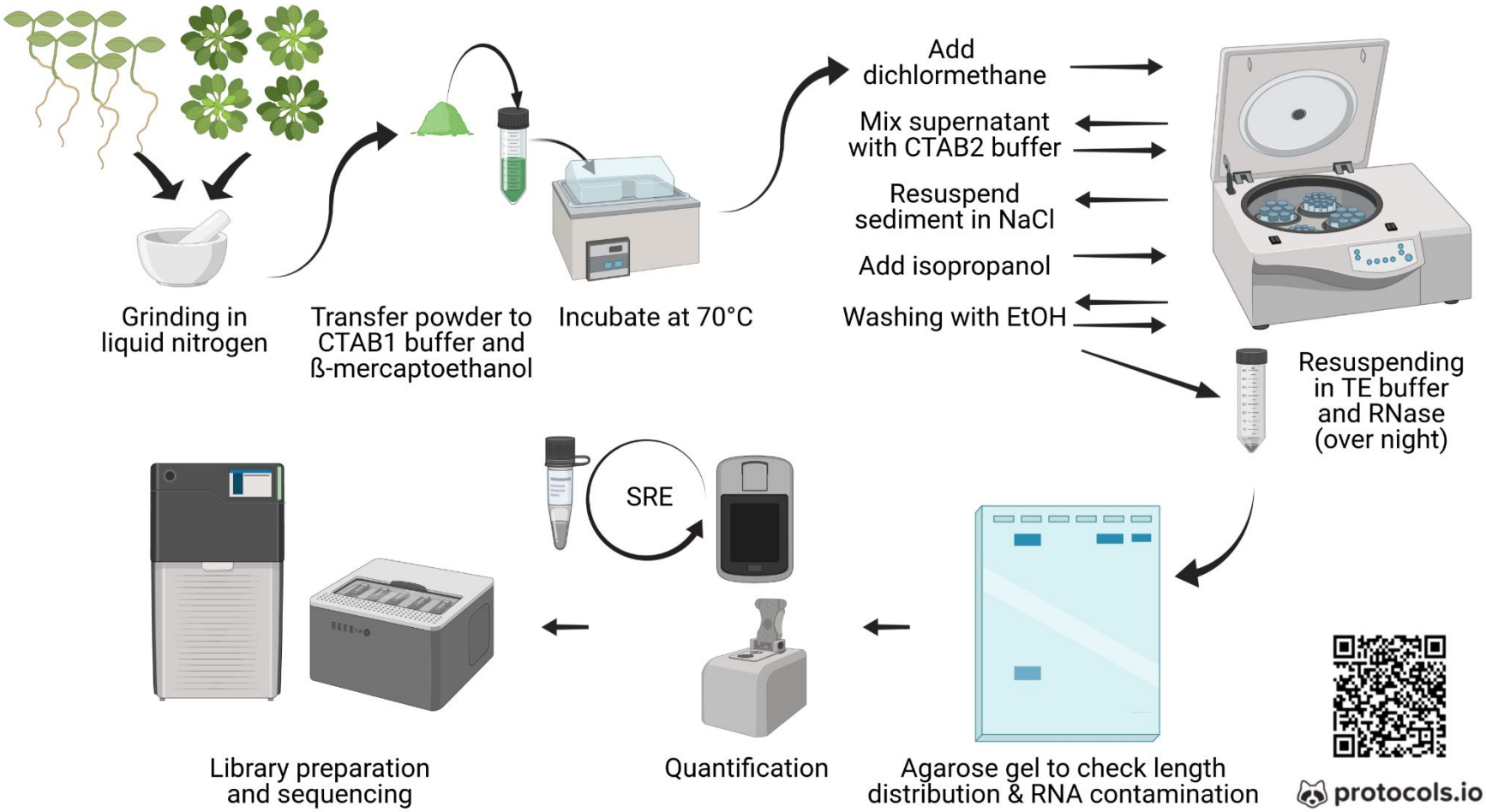
Picking an individual

- 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

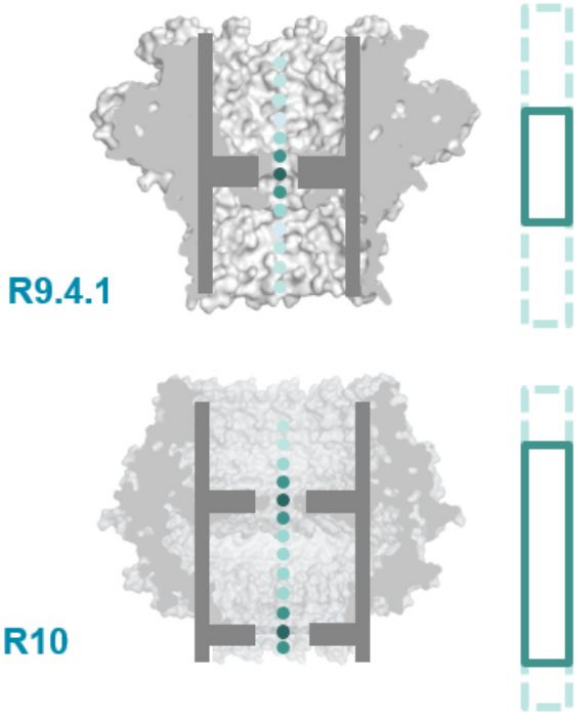
DNA extraction workflow



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	

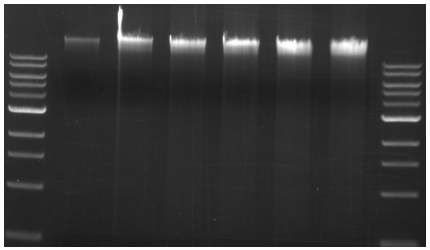
Nanopore comparison



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

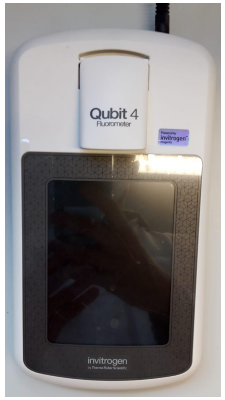
Quality control

- Agarose gel electrophoresis



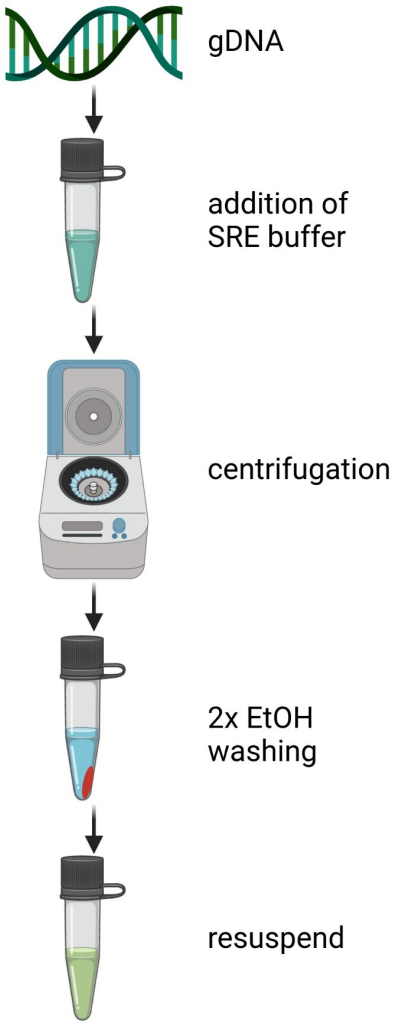
- Photometric measurement via NanoDrop

- Quantification with Qubit

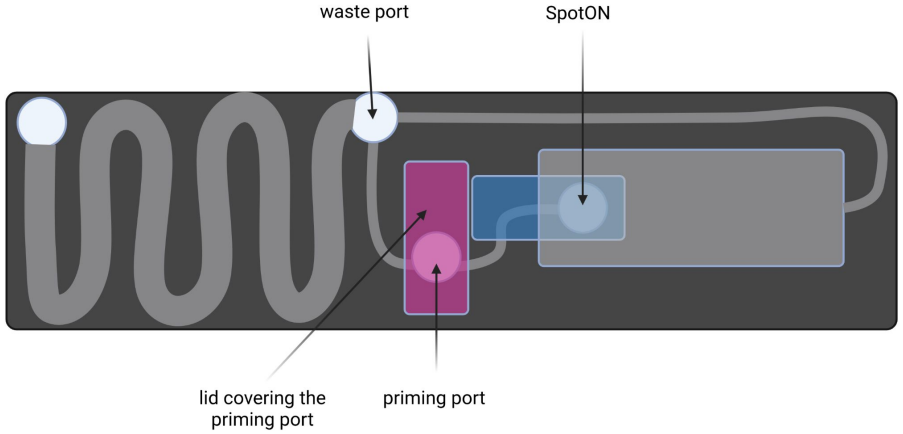


Short Read Eliminator (SRE)

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



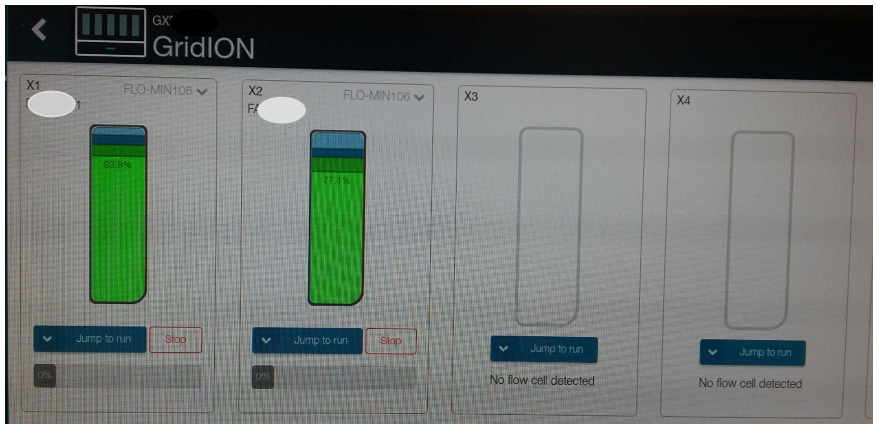
Loading flow cell



- Flow cell check (>800 guaranteed)
- 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

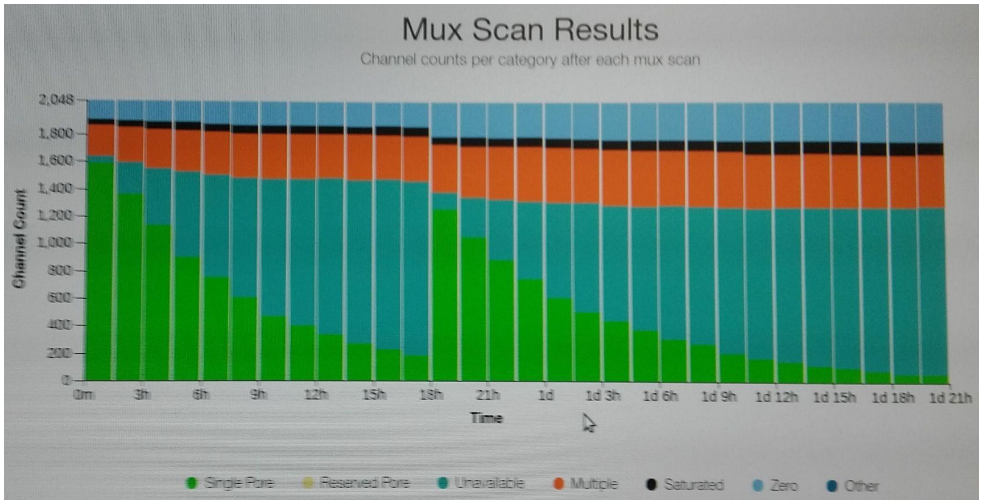
Monitor sequencing

- 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



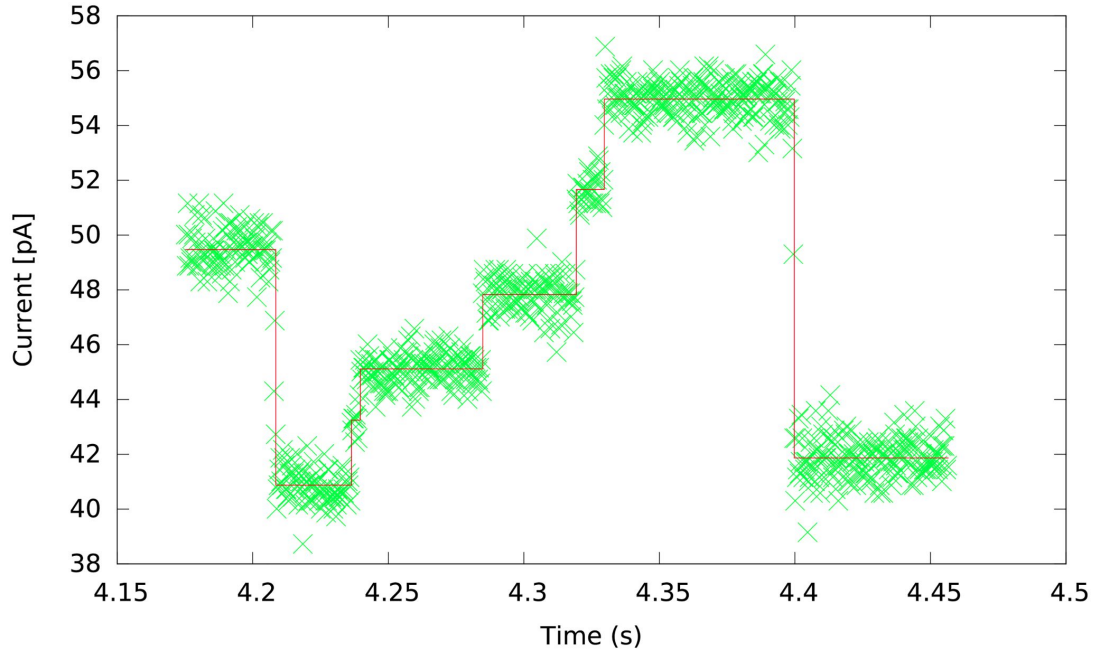
Stop & wash

- 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/POD5) need to be stored



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

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

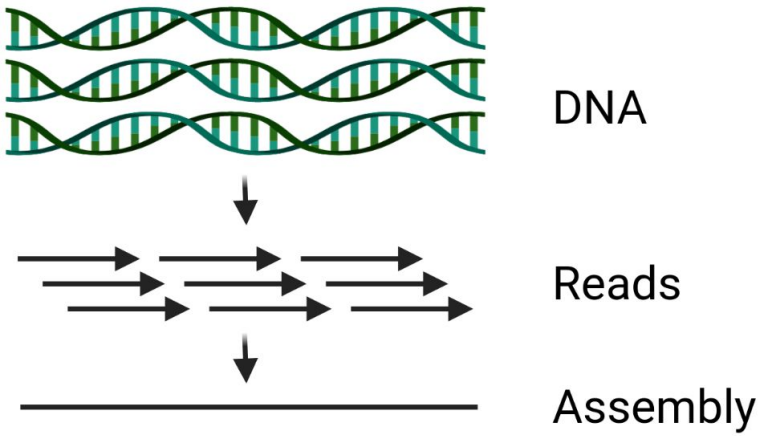
Summary sequencing technologies

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

Genome sequence assembly

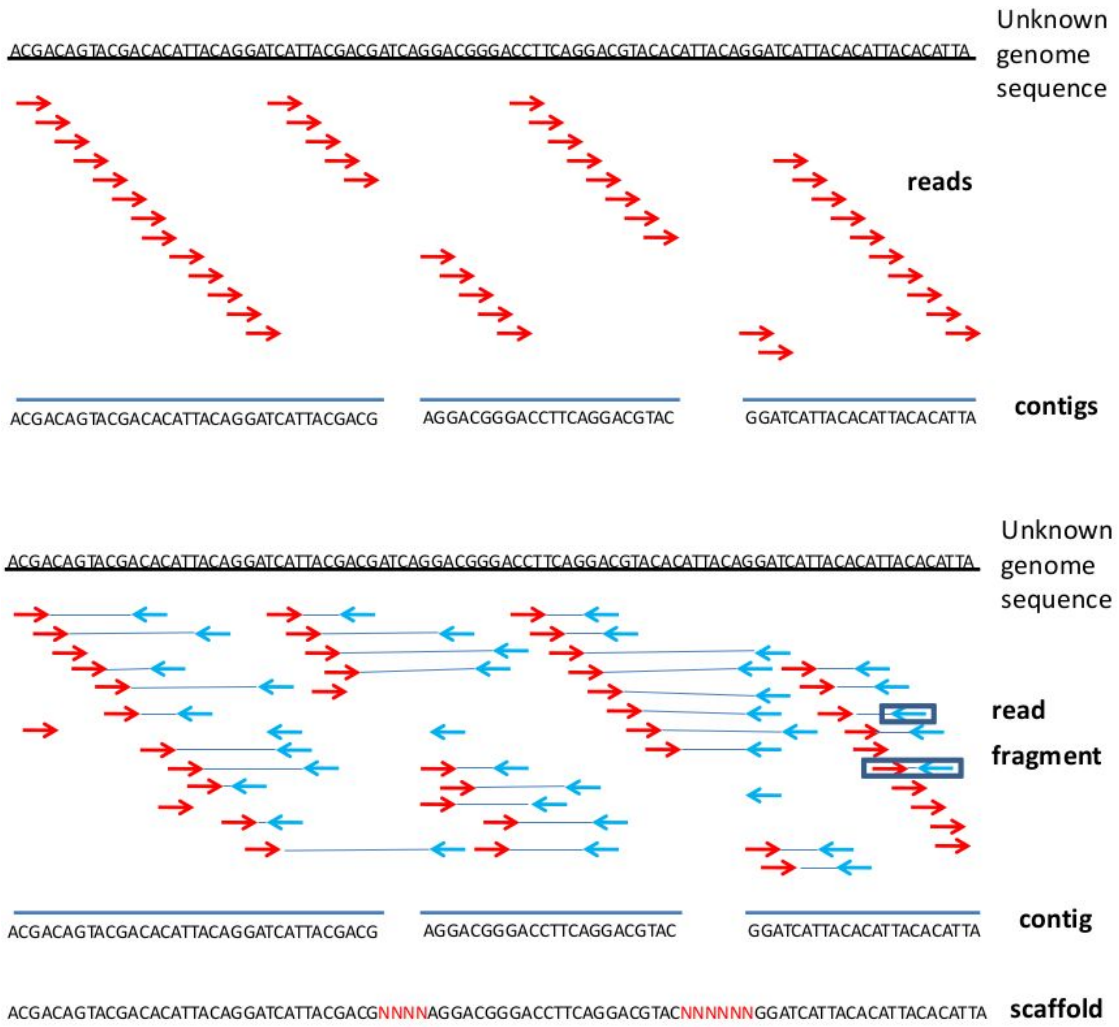
The assembly problem

- Reads are shorter than the chromosome - even long reads
- Multiple copies of the genome (DNA exist that can be subjected to sequencing
- Assembly = putting sequence pieces together (finding the common string of all substrings)
- Genome = DNA in a cell
- Genome sequence = representation of the DNA in a cell; stored in FASTA files




```
>TRINITY_DN100016_c0_g1_i1
MPKKSSNIQQNNQNRIKRKGRPPKKKKYVQQIDSSDEDILSVRHASTRPRIISIRRNEIPMRPEIHI
>TRINITY_DN100019_c0_g1_i1
MPPKAADKKPAAKAPATASKAPEKKDAGKKTAATGDKKKRTKARKETYSSYIYKVLKQVHPDTGISN
>TRINITY_DN10001_c0_g1_i2
MAKVGNPIVDIETDGSVNEPESSEKNIEVSSSSTQAPESTNTTELLVNEKKAFSLATPAVRRVAREH
>TRINITY_DN100025_c0_g1_i1
MVENQDGCFFKPGWKEFVRSNDLEGGDFLVFNLVDKISYQVVIFDGTCCPKDLCPFSIMNPIFQHLR
>TRINITY_DN10002_c0_g1_i1
MSDELNQVFQRCREQKRPVFVAFVTAGYPDSEETVDILLGLEAGGADIIELGIPFTDPMVDGKTIQD
>TRINITY_DN100061_c0_g1_i1
MQQVVAKLKAIITKTNVTNENSPVENSSSSTSATSSINNSLHGDLSRVFDNMELESNVSNSSISSNI
```

Contigs, scaffolds, pseudochromosomes



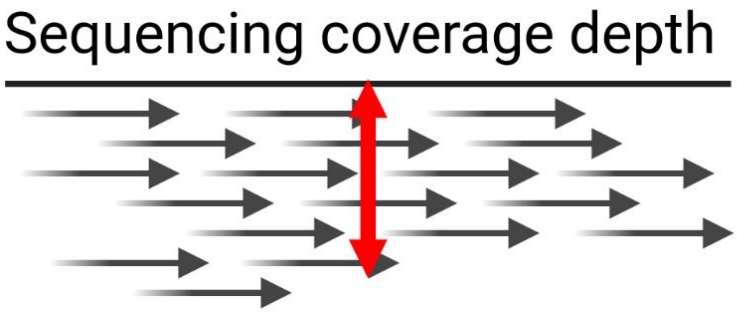
Contig = continuous sequence

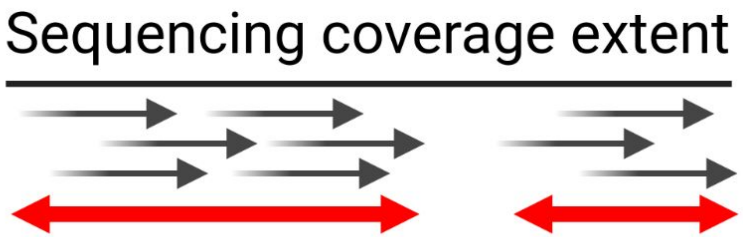
Scaffold = compilation of contigs with interleaved, unknown sequence (gaps)

Gaps = regions between contigs that are represented by Ns

Pseudochromosomes = scaffolds representing an entire chromosome

Sequencing coverage depth vs. coverage extent





Sequencing coverage depth

- Coverage depth = average number of times a given base is being sequenced

- Calculation:
  - N = number of reads
  - L = read length in base pairs
  - G = genome size in base pairs
  - Coverage depth d = N X L / G

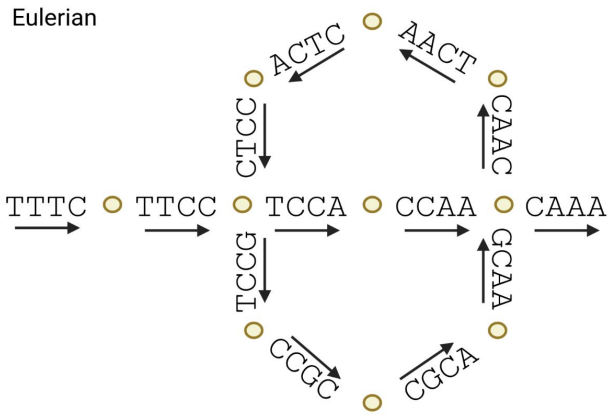
- Coverage (depth) reflects total amount of sequencing data

- Coverage (depth) is very important parameter for sequencing projects

De Brujin graph (DBG)

- 1) Reads are broken into smaller k-mers
- 2) K-mers represented in a “De Bruijn graph”
- 3) Inference of genome sequence from graph
- Paradigm used in many assemblers: Velvet, ABySS, AllPath-LG, SOAPdenovo
- Complexity:
  - Number of nodes and links equal to genome size
  - Eulerian path problem is ‘easy’ to solve

```
TTTCCGCAACTCCAA
TTTC
 TTCC
  TCCG
   CCGC
    CGCA
     GCAA
      CAAC
       AACT
        ACTC
         CTCC
          TCCA
           CCAA
```

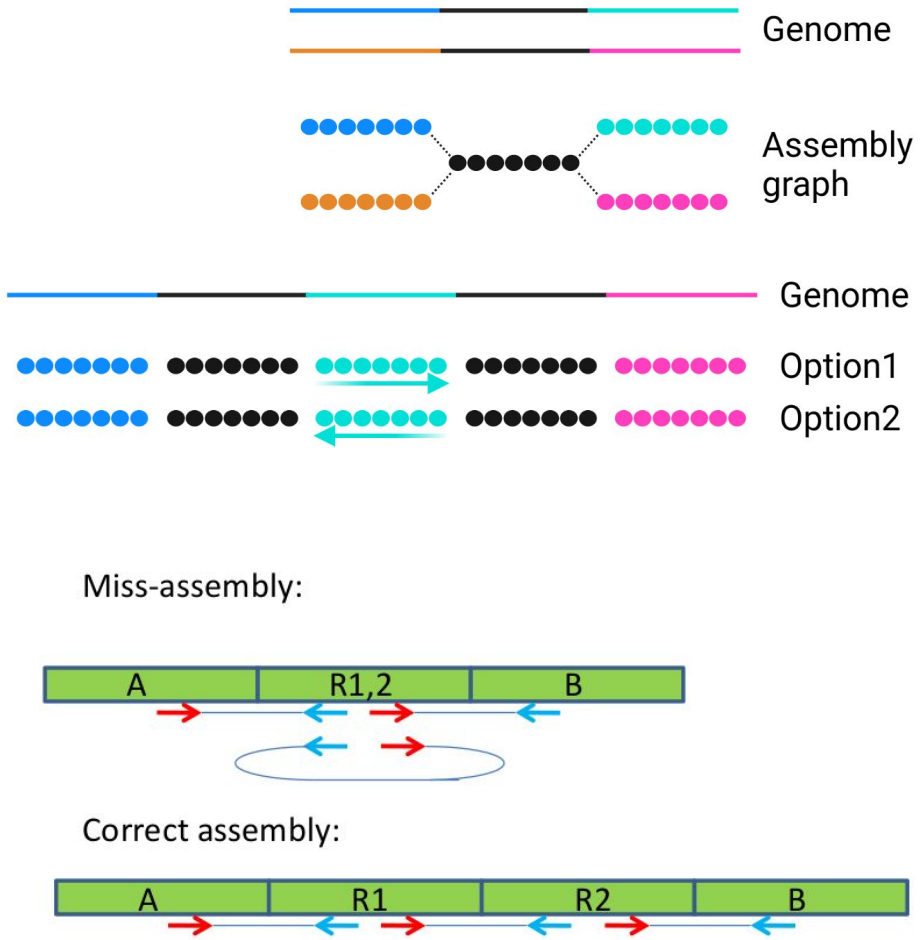


K-mer size

- Larger k-mers increase the assembly continuity by spanning repeats
- Larger k-mers are more sensitive to sequencing errors
- Removal of low abundance k-mers (caused by sequencing errors)
- K-mer size should be 0.5 to 0.75 of the read length
- Typical k-mer sizes:
  - 87 or 97 for 2x150bp reads

Assembly challenges

- Collapsed repeats
- Inversions
- Overstretched repeats

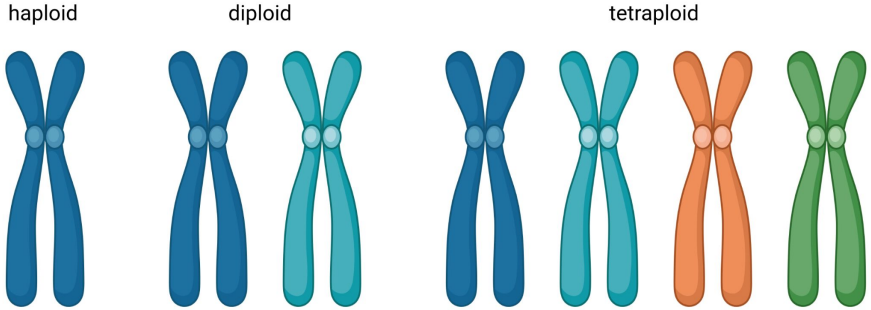


Assembly challenges (2)

- Genome size: variation from 65 Mbp to 149 Gbp

Genlisea violacea: https://commons.wikimedia.org/wiki/File:Genlisea_violacea_giant.jpg
Utricularia gibba: https://commons.wikimedia.org/wiki/File:Utricularia_gibba_flower_01.jpg
Arabidopsis thaliana: https://commons.wikimedia.org/wiki/File:Arabidopsis_thaliana_sl10.jpg
Dioscorea dumetorum: https://commons.wikimedia.org/wiki/File:YamsatBrixtonMarket.jpg
Beta vulgaris: https://de.wikipedia.org/wiki/Zuckerr%C3%BCbe#/media/Datei:Zuckerr%C3%BCbe.jpg
Vitis vinifera: https://commons.wikimedia.org/wiki/File:Vitis_vinifera_20120823.jpg
Sequoia giganteum: https://commons.wikimedia.org/wiki/File:Grizzly_Giant_Mariposa_Grove.jpg
Paris japonica: https://commons.wikimedia.org/wiki/File:Paris_japonica_Kinugasasou_in_Hakusan_2003_7_27.jpg

- Ploidy: haploid/diploid genomes are much easier to analyze than polyploid genomes



Kress et al., 2022: 10.1073/pnas.2115640118

Assembly polishing

ONT assembly ...ACGTTACGGTACGACATGACGTAAAAAAAAACGATAGTACGTGTTAGCGTACGTACGTAACGTT...

Illumina reads

GACGTAAAAAAAA-CGATAGTACGTGTTAGCGTACGTAC
ATGACGTAAAAAAAA-CGATAGTACGTGTT
CGTAAAAAAAA-CGATAGTACGTGTTAGCGTAC
ACATGACGTAAAAAAAA-CGATAGTACGT
CGTAAAAAAAA-CGATAGTACGTGTTAGCGT
GACGTAAAAAAAA-CGATAGTA

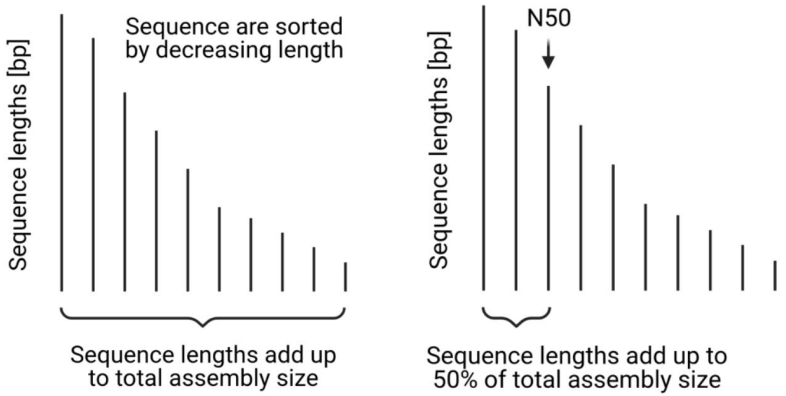
Assembly evaluation (CCC)

- Continuity: Does the assembly represent a genome in a small number of contigs?
- Completeness: Are all parts of the genome represented?
- Correctness: Is the assembly a correct representation of the genome?

Evaluation: continuity

- Check assembly continuity: calculation based on sequences in FASTA file
- Number of contigs, assembly size, N50

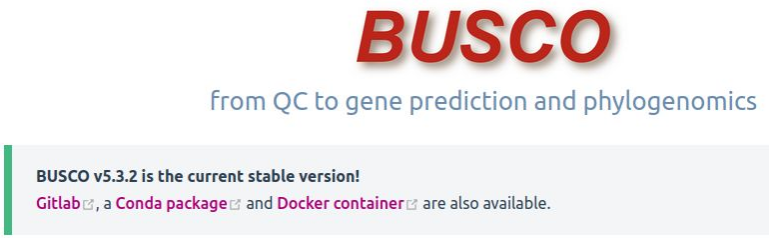
assembler	Canu	FALCON	Miniasm	Flye
number of contigs	69	26	72	44
assembly size	123.5 Mbp	119.5 Mbp	120.2 Mbp	117 Mbp
maximal contig length	15.9 Mbp	15.9 Mbp	14.3 Mbp	14.9 Mbp
N50	13.4 Mbp	9.3 Mbp	8.6 Mbp	10.6 Mbp
N90	2.9 Mbp	2.8 Mbp	1.4 Mbp	2.5 Mbp



Pucker et al., 2019: 10.1371/journal.pone.0216233

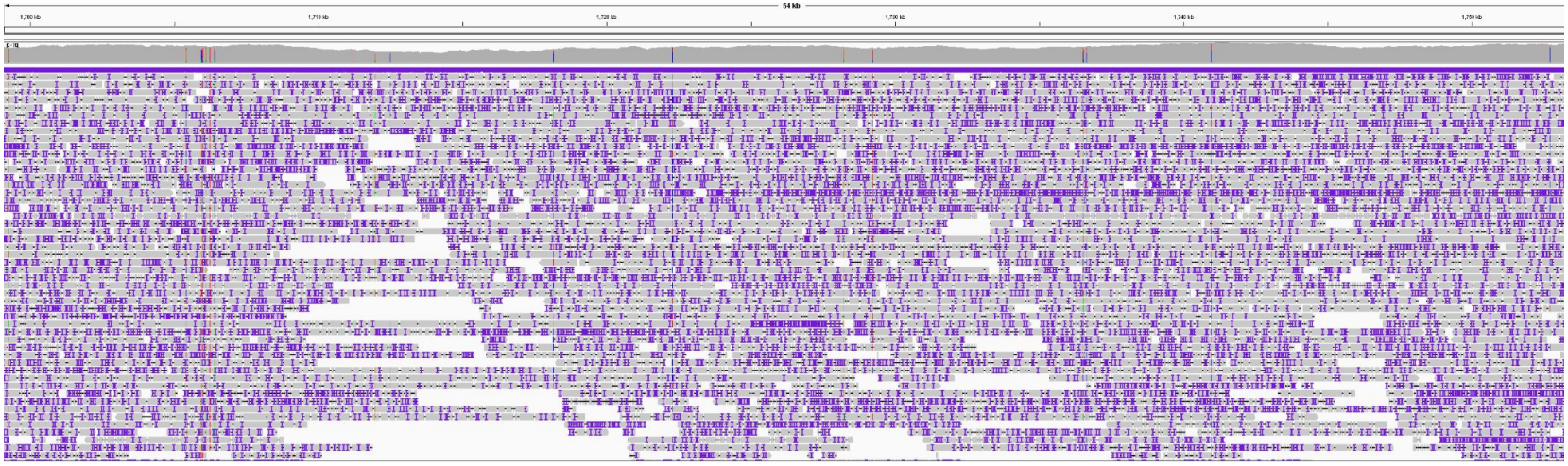
Evaluation: completeness

- Check assembly completeness: inspection for presence of conserved genes
- BUSCO = Benchmarking Universal Single-Copy Orthologs
- BUSCO genes are used to assess assembly completeness



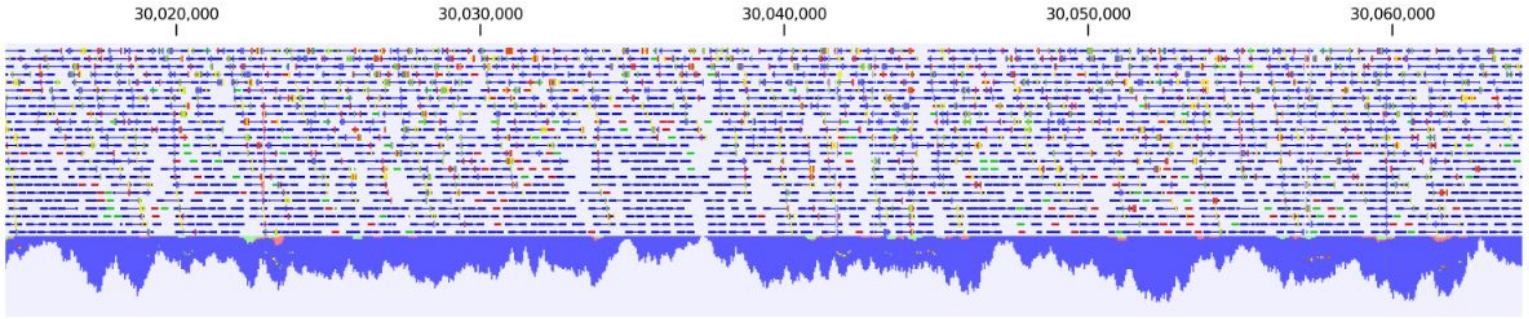
Evaluation: assembly correctness

- Check assembly correctness: analyses of read mappings
- Integrative Genomics Viewer (IGV) can visualize read mappings
- Tools: REAPR, SQUAT



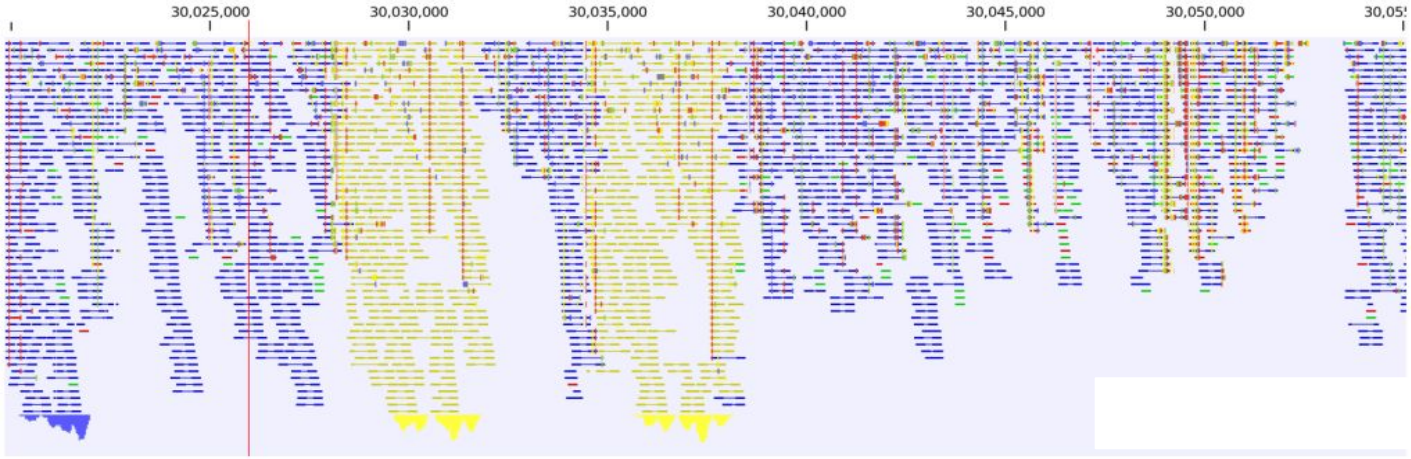
IGV: Thorvaldsdottir et al., 2013: 10.1093/bib/bbs017
REAPR: Hunt et al., 2013: 10.1186/gb-2013-14-5-r47
SQUAT: Yang et al., 2019: 10.1186/s12864-019-5445-3

Read mappings (1)

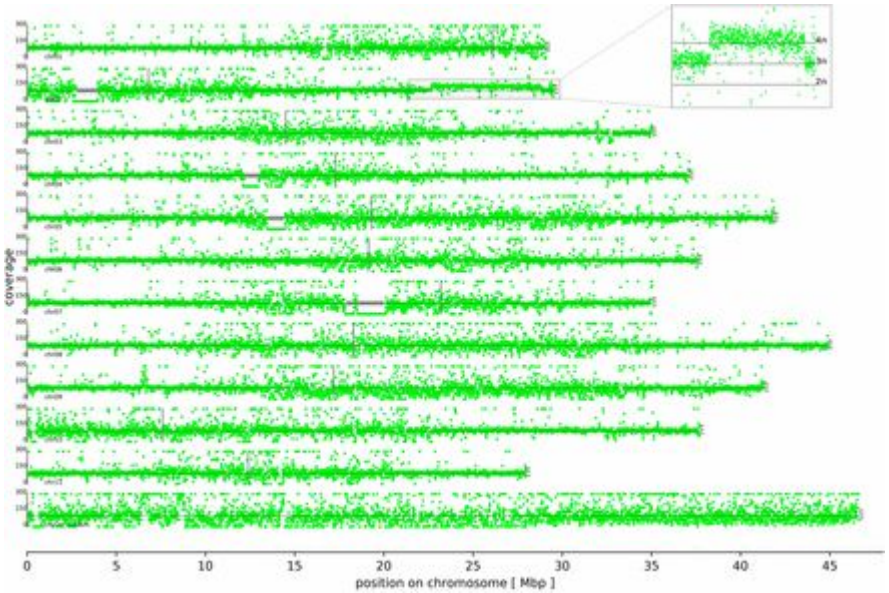


Read mapping of paired-end sequenced fragments (blue) to assembly

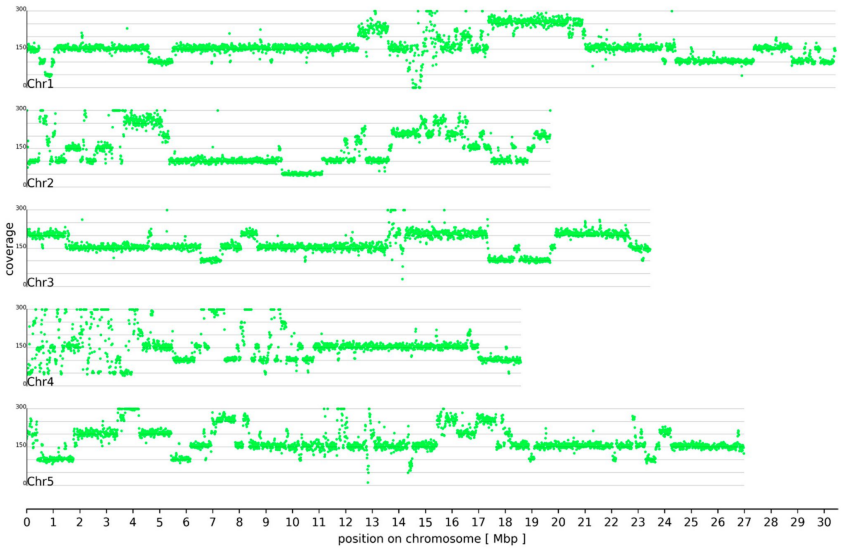
Coverage is too high to show all individual fragments at some positions



Read mappings (2)



Musa acuminata (banana) read mapping



Arabidopsis thaliana At7 read mapping

Pucker et al., 2019: 10.3390/genes10090671
Busche & Pucker et al., 2020: 10.1534/g3.119.400847

Advantages of long reads

- Span larger regions and enable assembly of repeats
- Specific mapping to repetitive regions possible
- Generation of larger contigs (no scaffolding)
- Contigs can represent entire chromosomes

Long read assemblers

- Canu: https://github.com/marbl/canu
  Nurk et al., 2020: 10.1101/gr.263566.120

- Miniasm: https://github.com/lh3/miniasm
  Li, 2016: 10.1093/bioinformatics/btw152

- FALCON: https://github.com/PacificBiosciences/FALCON
  Chin et al., 2016: 10.1038/nmeth.4035

- Shasta: https://github.com/chanzuckerberg/shasta
  Shafin et al., 2020: 10.1038/s41587-020-0503-6

- NextDenovo2: https://github.com/Nextomics/NextDenovo

Importance of error rate

- Correction step is computationally extremely intense
  - Computation of all-vs-all alignments
- Direct assembly is possible with >99% raw read accuracy
- Higher accuracy allows to filter read overlaps more strictly

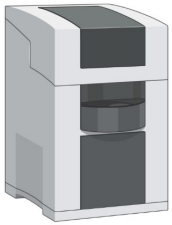
Integration of genetic linkage information (1)

- Classical genetic markers: SSR, CAPS, KASP
  - SSR = Simple Sequence Repeats
  - CAPS = Cleaved Amplified Polymorphic Sequences
  - KASP = Kompetitive Allele Specific PCR

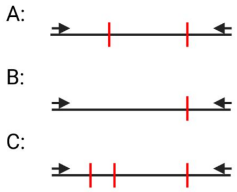
SSR: Simple Sequence Repeats

```
A: CATAGAGAGAGAGAGATGAC
B: CATAGAGAGAGAGATGAC
C: CATAGAGAGATGAC
D: CATAGAGAGAGAGAGATGAC
E: CATAGAGAGAGAGATGAC
F: CATAGAGAGAGAGAGATGAC
G: ACATAGAGATGAC
```

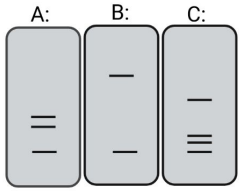
Length analysis via capillary electrophoresis



CAPS: Cleaved Amplified Polymorphic Sequences



PCR & restriction digest



KASP: Kompetetive Allele Specific PCR

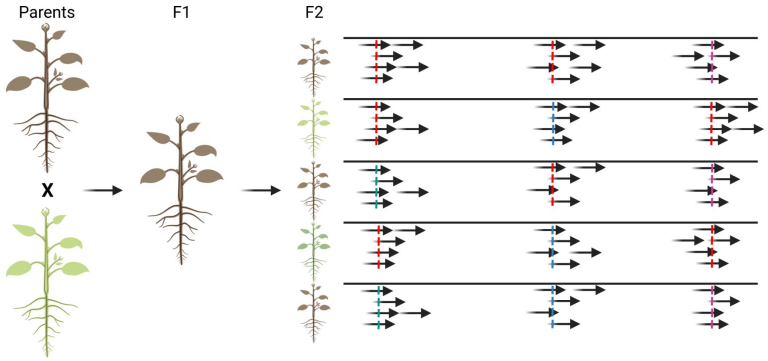
CATAGACACTG[G/A]GTTAGAGATGAC

qPCR with fluorescently labeled primers

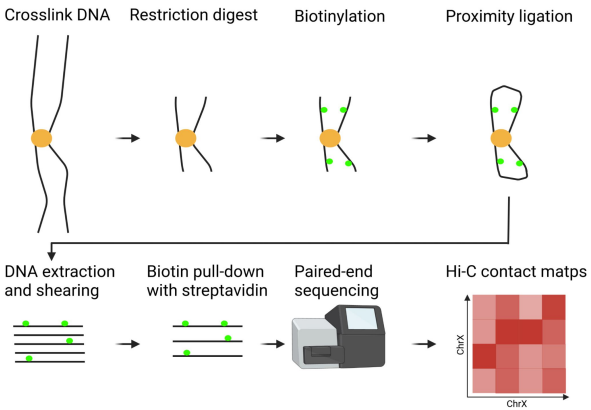
SSR: Holtgräwe et al., 2020: 10.3389/fpls.2020.00156
CAPS: Konieczny & Ausubel, 1993: 10.1046/j.1365-313x.1993.04020403.x
KASP: He et al., 2014: 10.1007/978-1-4939-0446-4_7

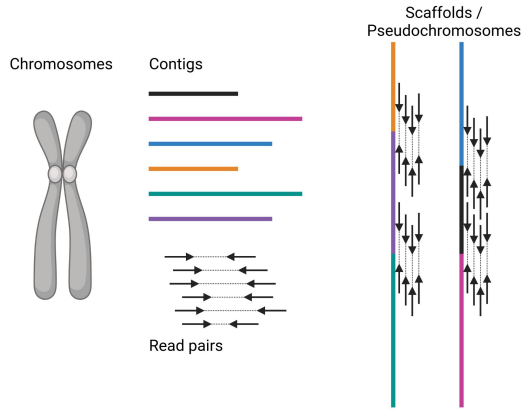
Integration of genetic linkage information (2)

- Genotyping-by-sequencing: SNPs inferred from sequencing data



- Hi-C: chromatin interaction information for long range scaffolding



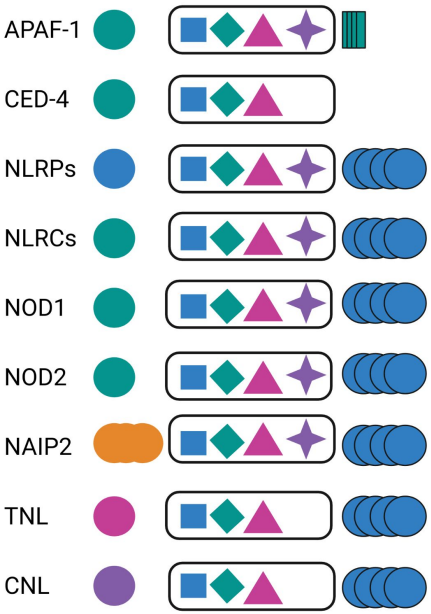


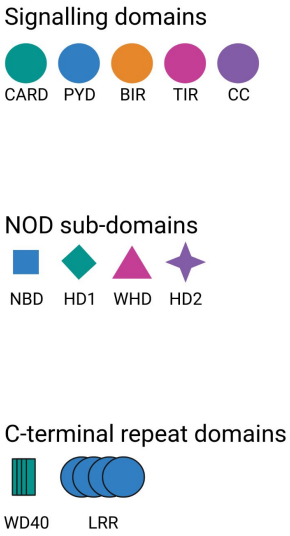
Checking challenging regions

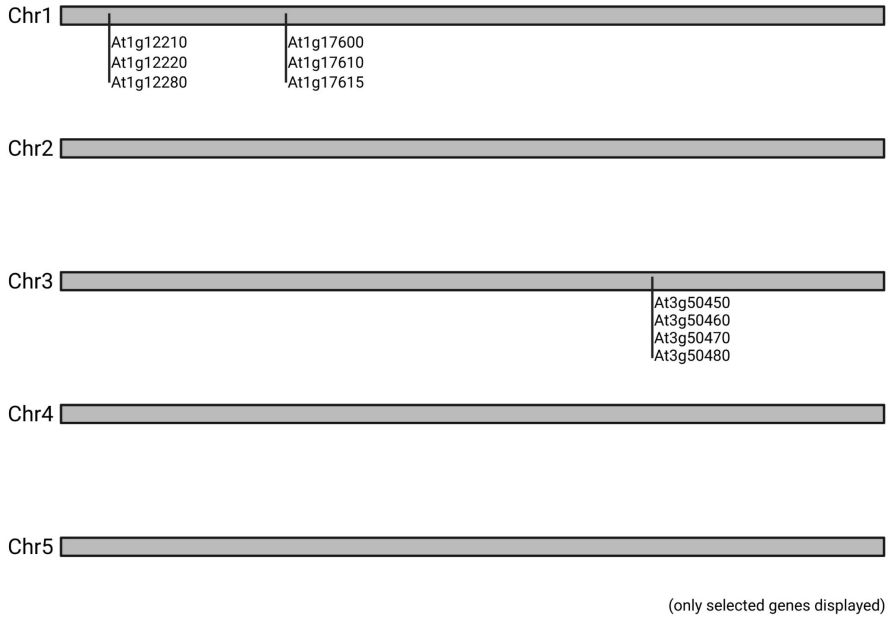
- Resolving the centromeres and NORs are the last big challenges
- Centromeres = repeats in the middle of chromosomes
- Nucleolus Organizing Regions (NORs) = ribosomal RNA encoding repeats (rDNA)
- Checking presence of telomeres at contig ends

NLRome

- NLR genes (NLRome) are often clustered in tandem repeat arrays
- NLR gene clusters are particularly tricky to assemble
- NLRome used for benchmarking high quality assembly





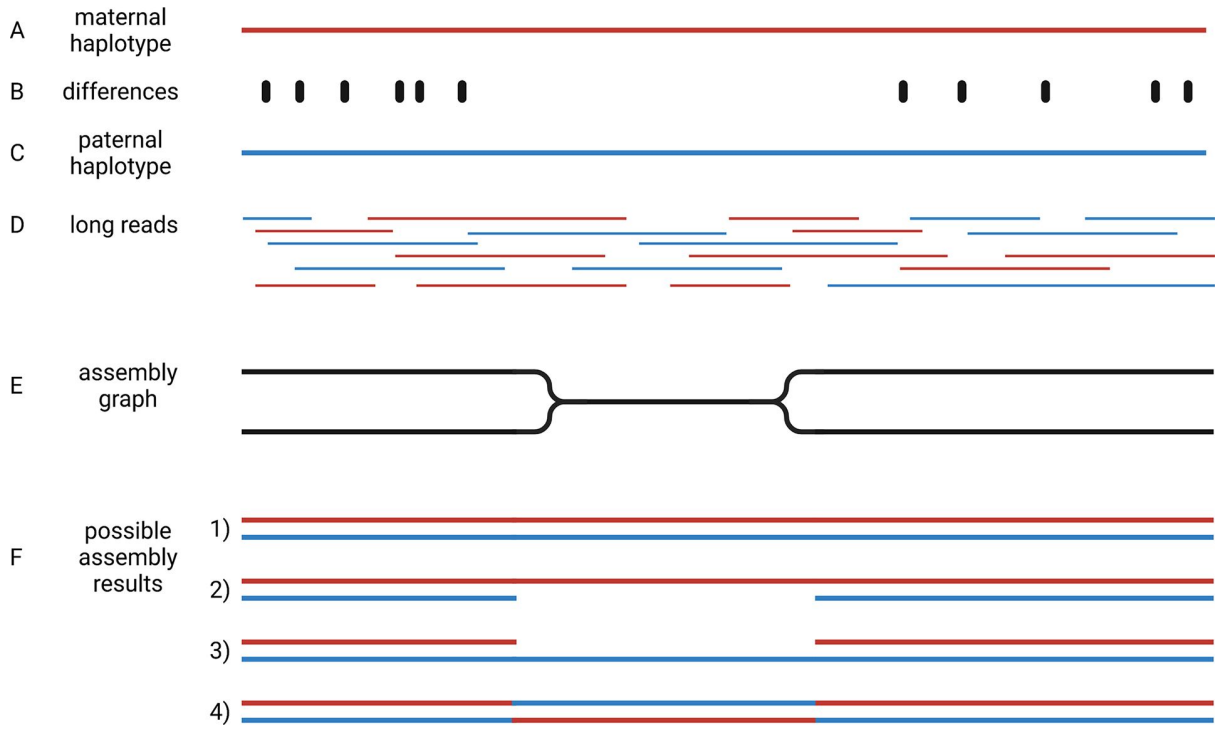


(only selected genes displayed)

Van de Weyer et al., 2019: 10.1016/j.cell.2019.07.038

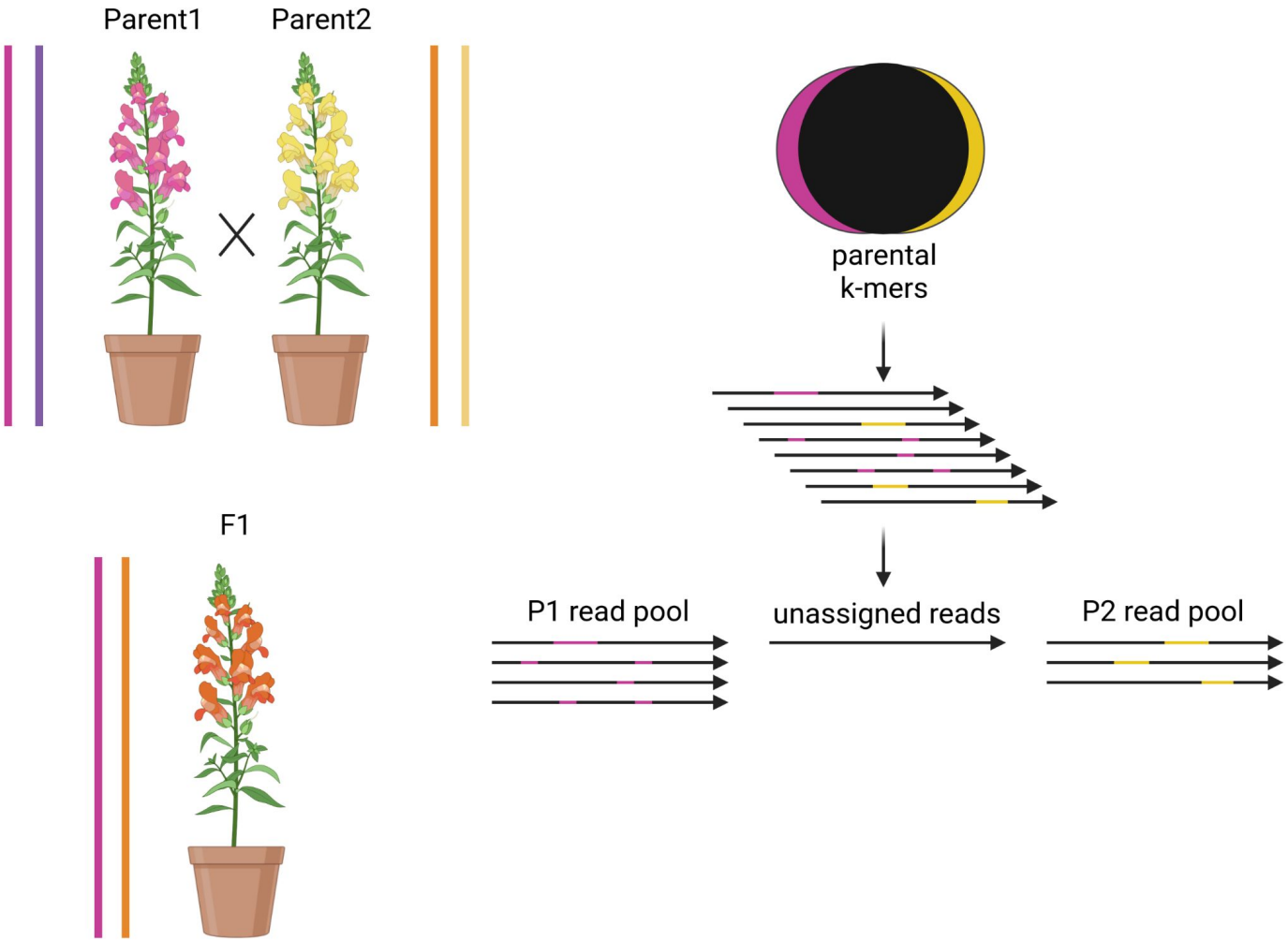
Haplophases

- 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

Summary

- Sequencing technologies:
  - Sanger
  - Illumina
  - **ONT**
  - PacBio

- Genome sequence assembly

Time for questions!

Practical part: outline day2

- QC & trimming/filtering of ONT long reads (Nanostat, porechop, chopper)
- Assembly (Shasta)
- Assess data set completeness (BUSCO)

Example data set: *Digitalis purpurea*

- Common foxglove is an ornamental and medicinal plant
- Flower coloration is an interesting trait
- Genome of multiple plants sequenced by ONT
- Data sets: digi1.fastq.gz & digi4.fastq.gz

https://doi.org/10.1101/2024.02.14.580303

Exercises: 2-1

- Find and install a tool to check read quality and perform trimming if necessary
- Run an assembly of digi1 and digi4 with Shasta (fastq2fasta.py)
- Calculate assembly statistics (contig_stats3.py)
- Run BUSCO on assembly

QC & trimming of long reads

- Removal of low quality read parts
- Removal of adapters sequences
- Chopper: https://github.com/wdecoster/chopper

```
gunzip -c reads.fastq.gz | chopper -q 10 -l 500 | gzip > filtered_reads.fastq.gz
chopper -q 10 -l 500 -i reads.fastq > filtered_reads.fastq
chopper -q 10 -l 500 -i reads.fastq.gz | gzip > filtered_reads.fastq.gz
```

- Porechop: https://github.com/rrwick/Porechop
- ProwlerTrimmer: https://github.com/ProwlerForNanopore/ProwlerTrimmer

Genome sequence assembly with Shasta

- Very fast and resource efficient assembly
- GitHub: https://github.com/paoloshasta/shasta

Download the executable for the latest release:

```
curl -O -L
https://github.com/paoloshasta/shasta/releases/download/0.11.1/shasta-Linux-0.11.1
```

Grant execute permissions:

```
chmod ugo+x shasta-Linux-0.11.1
```

Run an assembly:

```
./shasta-Linux-0.11.1 \
--input reads.fasta \
--config Nanopore-May2022 \
--assemblyDirectory /vol/data/shasta_assembly/ \
>> /vol/data/shasta_doc.txt \
2>&1 &
```

Calculate assembly stats

- Calculate assembly statistics with contig_stats3.py
- Availability: https://github.com/bpucker/GenomeAssembly
- Usage:

```
python3 contig_stats3.py \
--in assembly.fasta \
--min 1000 \ #minimal contig length
--out /vol/data/stats_folder/
```

- Output:

```
number of contigs:      4267
average contig length:  34380.992734942585
minimal contig length:  1001
maximal contig length:  576842

total number of bases:  146703696
total number of bases without Ns:       146703696
GC content:     0.3364022880514203

N25:    101772
N50:    51874
N75:    24480
N90:    15257
```

DOCKER

- Docker enables installation of software with all dependencies
- Follow instructions here to install docker: https://docs.docker.com/engine/install/ubuntu/#install-using-the-repository
- Run docker command:

```
sudo docker run -v /vol/data:/vol/data --user 1000:100 --rm -it teambraker/braker3:latest bash
```

- Mapping of outside to inside paths via -v /path/to/outside_dir:/path/to/inside_dir
- Type exit to leave the docker container

Assess data set completeness with BUSCO

- BUSCO checks sequence data set for presence of well conserved genes
- Availability: https://busco.ezlab.org/
- Usage:

```
python3 busco \
--in assembly.fasta \
--tmp_path /vol/data/waste/ \
--cpu 10 --mode genome \
--long --force --out busco_run \
> log.txt
```

- Output: C:95.7%[S:63.5%,D:32.2%],F:0.6%,M:3.7%,n:2326

```
INFO    C:95.7%[S:63.5%,D:32.2%],F:0.6%,M:3.7%,n:2326
INFO    2225 Complete BUSCOs (C)
INFO    1477 Complete and single-copy BUSCOs (S)
INFO    748 Complete and duplicated BUSCOs (D)
INFO    13 Fragmented BUSCOs (F)
INFO    88 Missing BUSCOs (M)
INFO    2326 Total BUSCO groups searched
```

Time for questions!