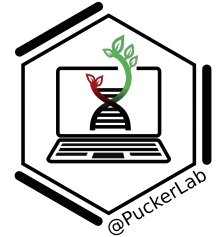
UNIVERSITÄT BONN



Prof. Dr. Boas Pucker

PBPM-BP-02

Availability of slides

- All materials are freely available (CC BY) - after the lectures:
  - eCampus: PBPM0 - Plant Biochemistry, Physiology and Molecular Biology (LEC)
  - GitHub: https://github.com/bpucker/teaching/PBPM

- Questions: Feel free to ask at any time

- Feedback, comments, or questions: pucker[a]uni-bonn.de

My figures and content can be re-used in accordance with CC BY 4.0, but this might not apply to all images/logos. Some figure were constructed using bioRender.com.

UNIVERSITÄT BONN

Repetition

- Global importance of plants (biomass)
- Role of plants during evolution (oxygen)
- Plant systematics / nomenclature
- Properties of model organisms
- Examples of model organisms

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.

UNIVERSITÄT BONN

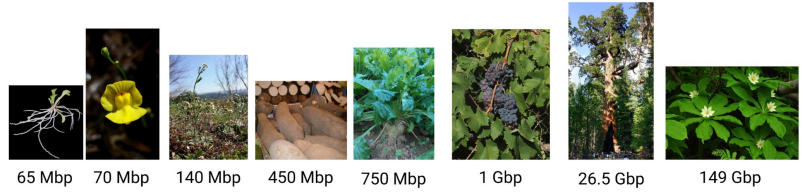
Major types of DNA in plants

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

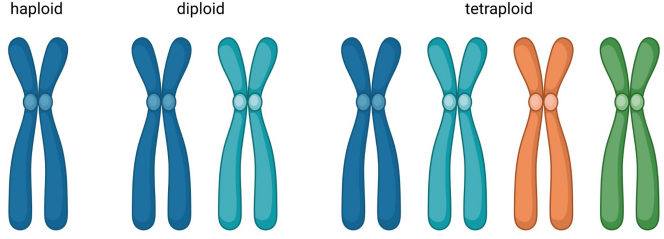
Plant genome sizes

- Genome size: variation from 65 Mbp to 149 Gbp

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



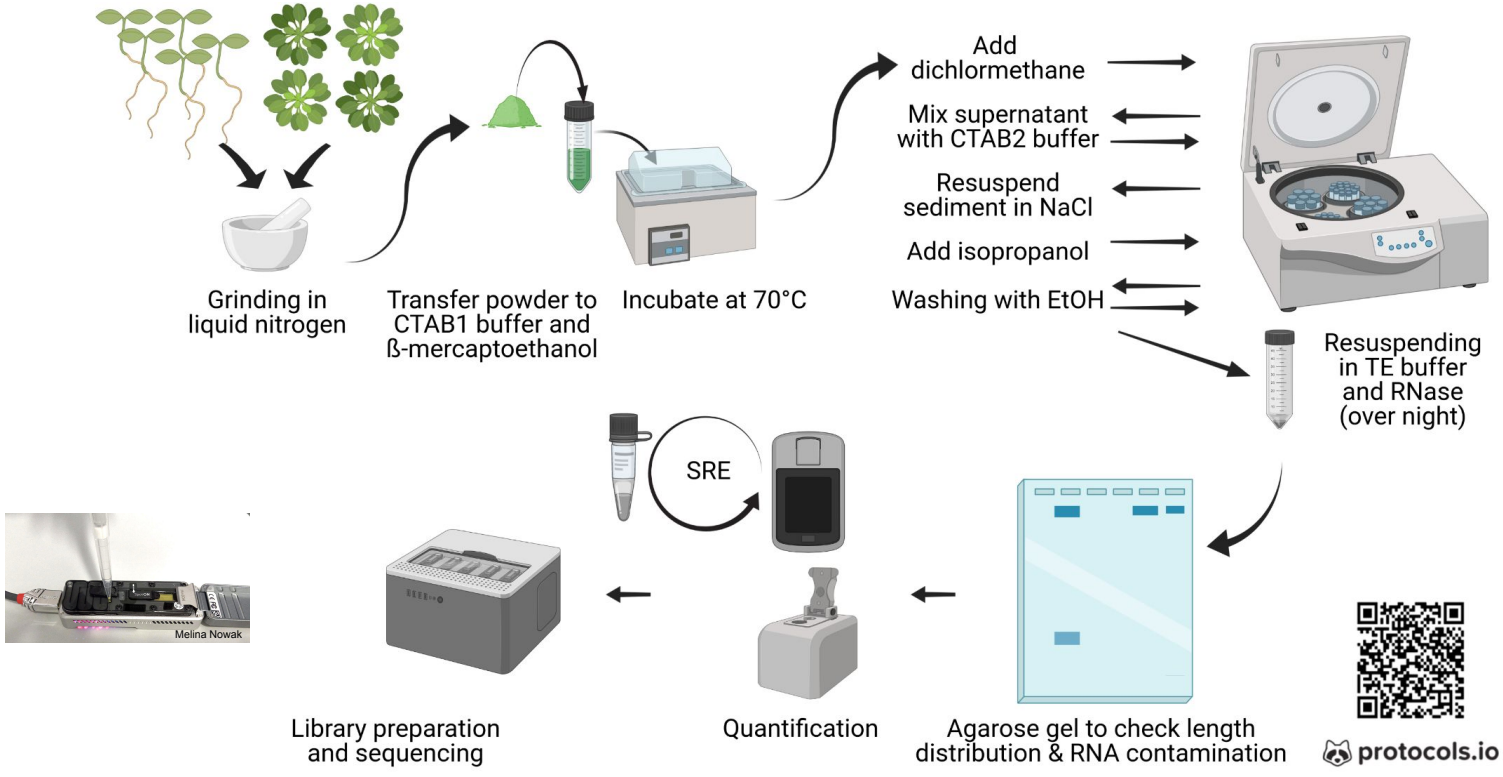
Genlisea violacea: https://commons.wikimedia.org/wiki/File:Genlisea_violacea_giant.jpg
Utricularia gibba: https://commons.wikimedia.org/wiki/File:Utricularia_gibba_flower_01.jpg
Arabidopsis thaliana: https://commons.wikimedia.org/wiki/File:Arabidopsis_thaliana_sl10.jpg
Dioscorea dumetorum: https://commons.wikimedia.org/wiki/File:YamsatBrixtonMarket.jpg
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
Sequica 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



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

UNIVERSITÄT BONN

CTAB-based DNA extraction



https://dx.doi.org/10.17504/protocols.io.bcvyiw7w

UNIVERSITÄT BONN

Challenges in plant DNA extraction

- Breaking the cell wall
- Specialized metabolites
- Sugar/starch
- Chloroplastic DNA

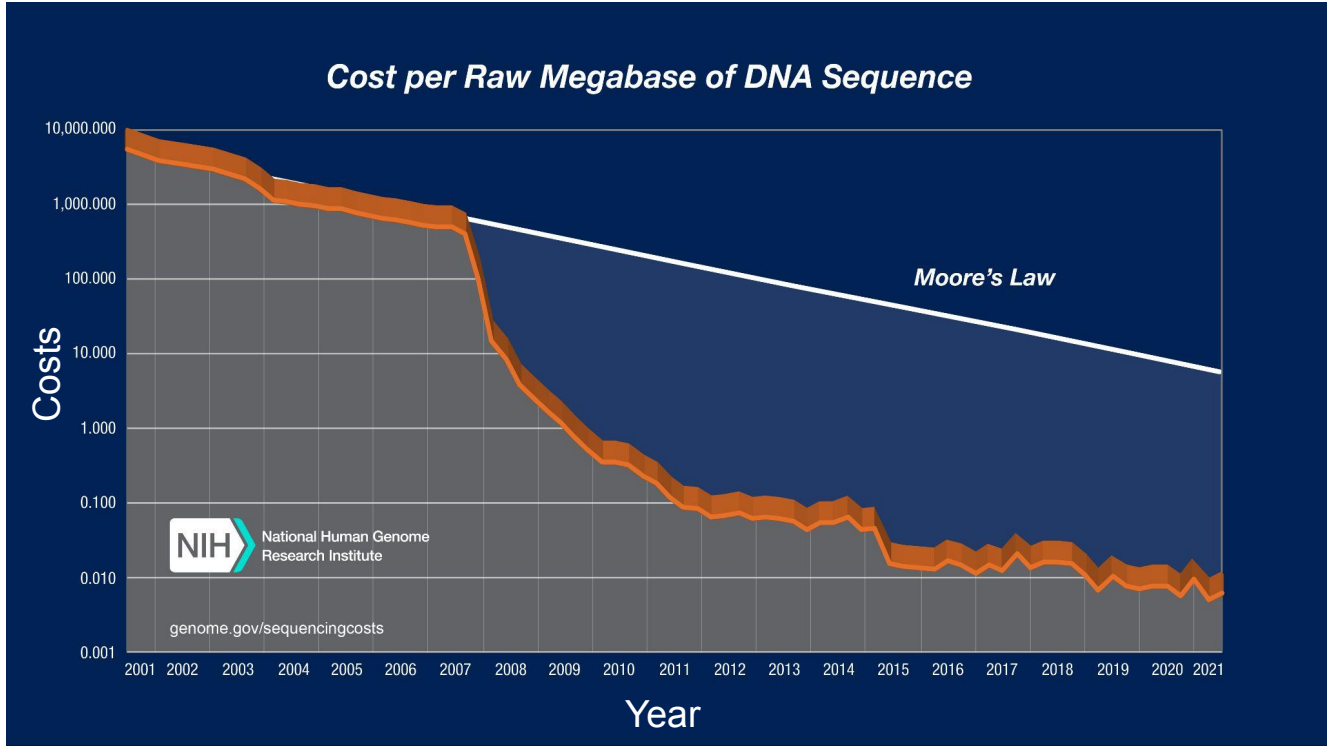
UNIVERSITÄT BONN

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?

UNIVERSITÄT BONN

Development of sequencing capacity

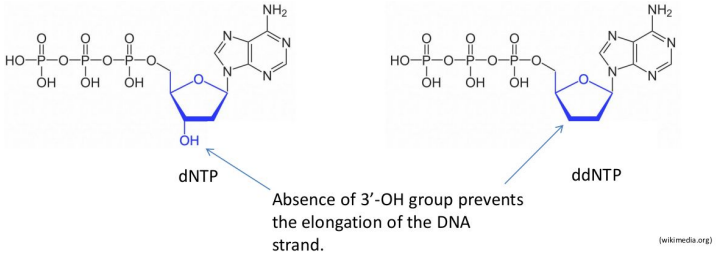


https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

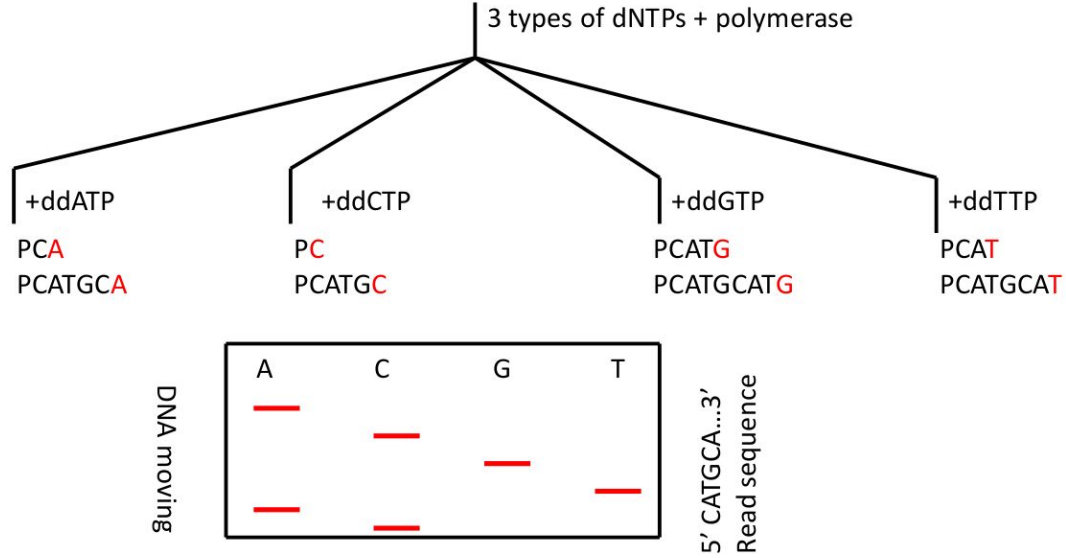
UNIVERSITÄT BONN

Sanger sequencing

Concept of Sanger sequencing

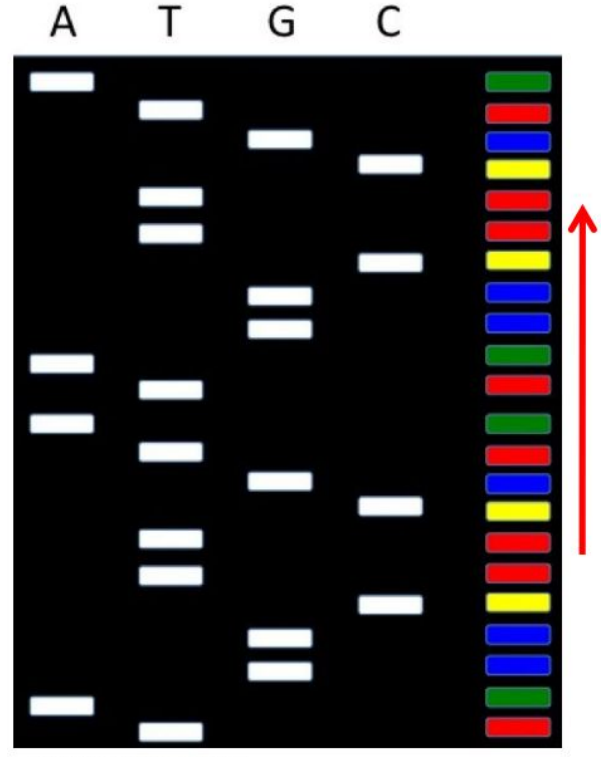


Primer (P): 5' –TGCATGGCATGATGCATG-3'
Template: 3' –ACGTACCGTACTACGTACGTACGTACGTCTAGGT-5'



Sanger sequencing - original version

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



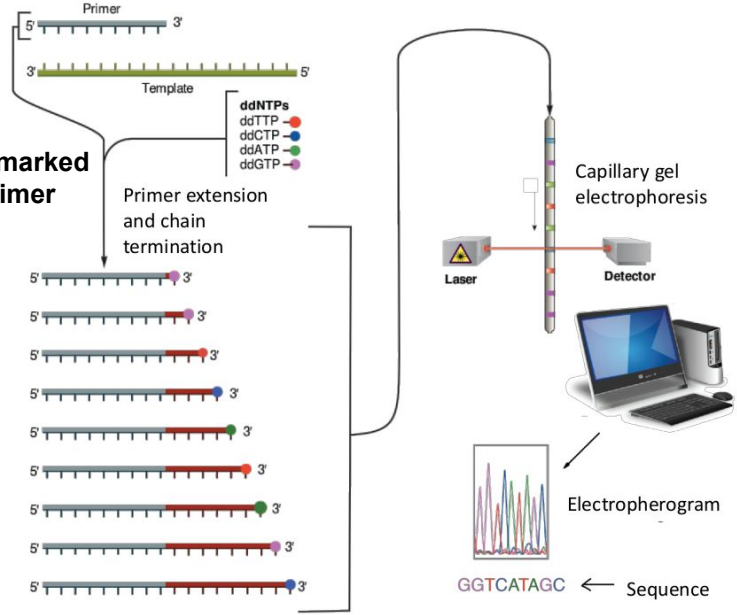
(modified from wikimedia.org)

Sanger sequencing - today

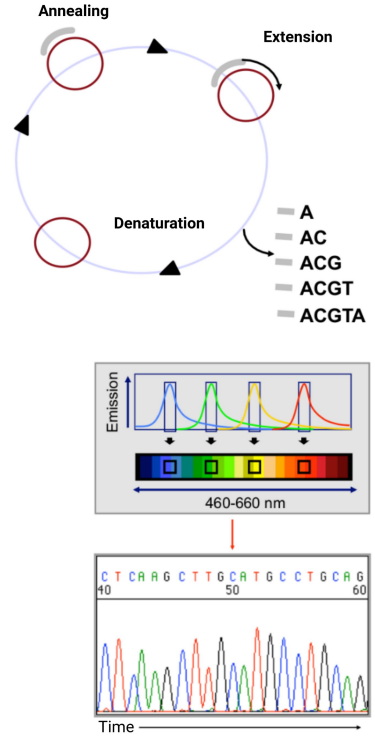
Only one reaction!

ddNTPs are marked instead of primer

Low input required due to cycle sequencing



Capillary avoids interference of adjacent lanes on a gel!

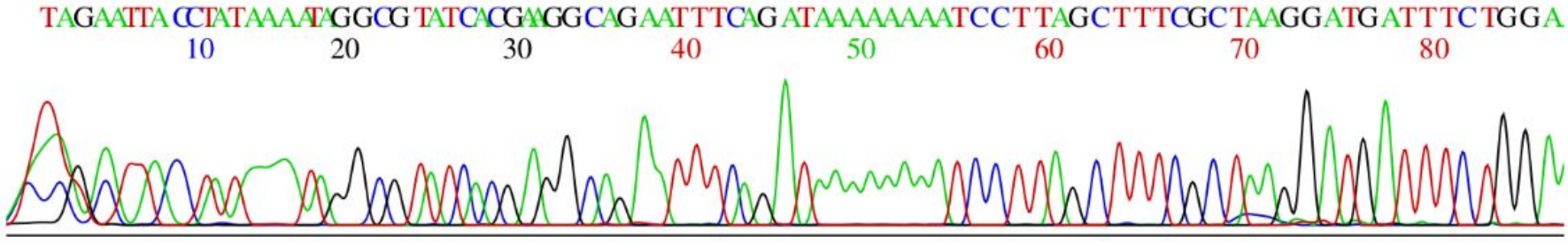


Figures modified from wikipedia

TRACE (.abi/.ab1)

- Original result file of ABI basecaller (Sanger sequencing)

- Contains only one red per file



UNIVERSITÄT BONN

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
```

- 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
```

UNIVERSITÄT BONN

Phred-Score

- Negative logarithm of the error probability for given position in read

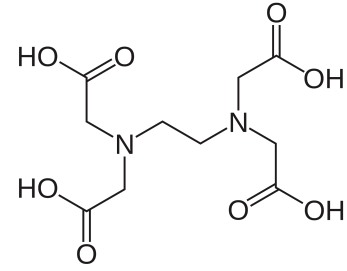
- Multiplication by 10 to avoid floats

$$Q = -10 \log_{10}(P)$$

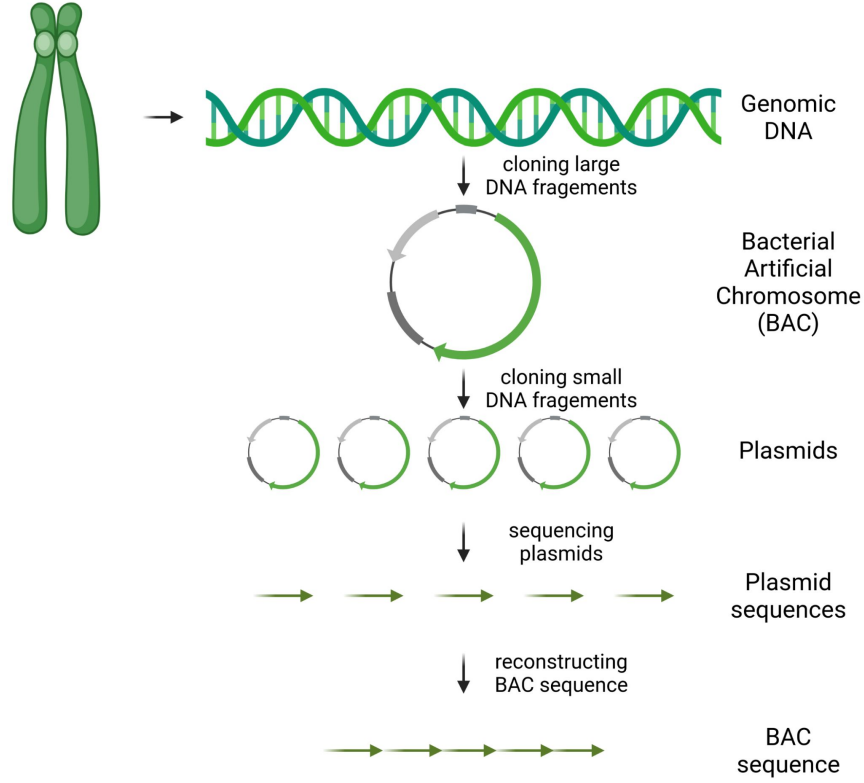
Phred quality score	Error probability	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

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



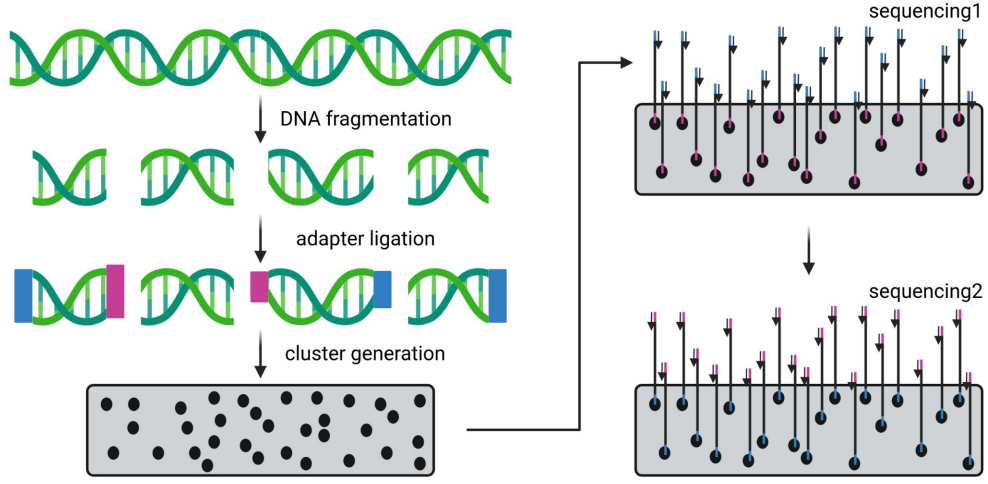
BAC-based sequencing strategy

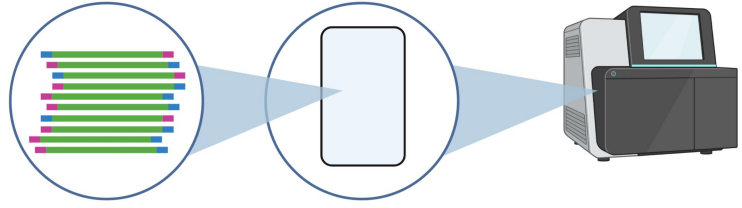


UNIVERSITÄT BONN

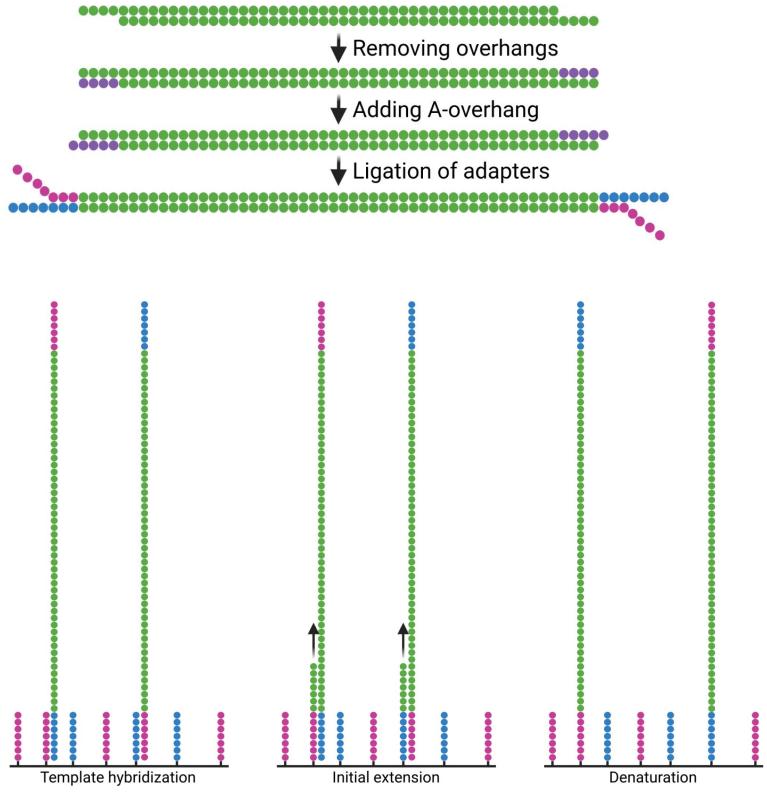
Illumina sequencing

Illumina sequencing (overview)

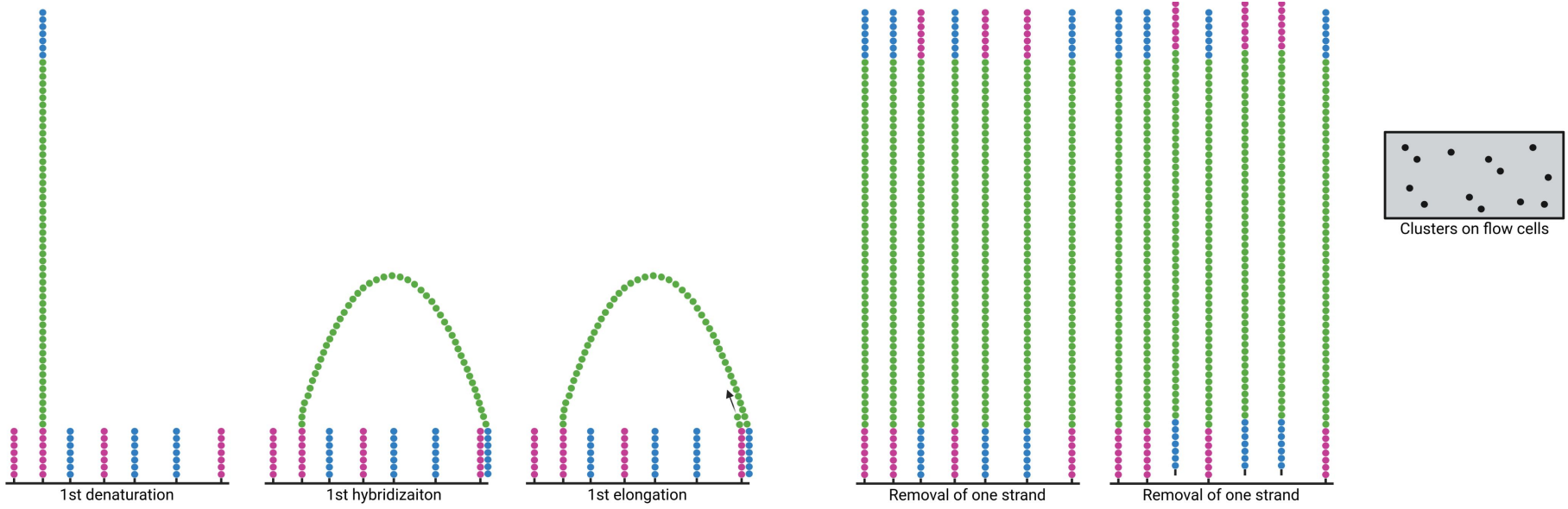




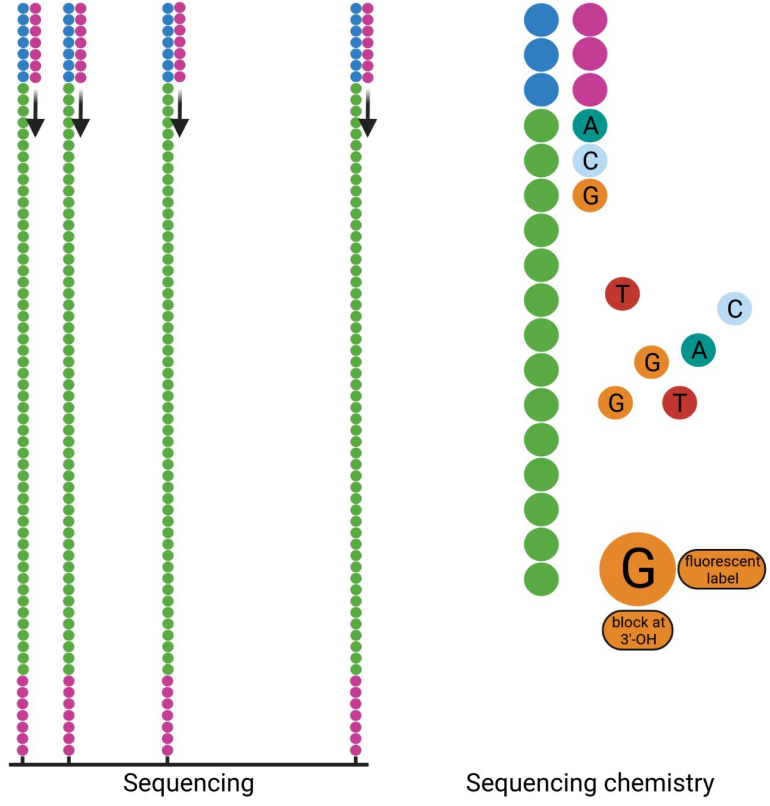
Illumina - sequencing 2



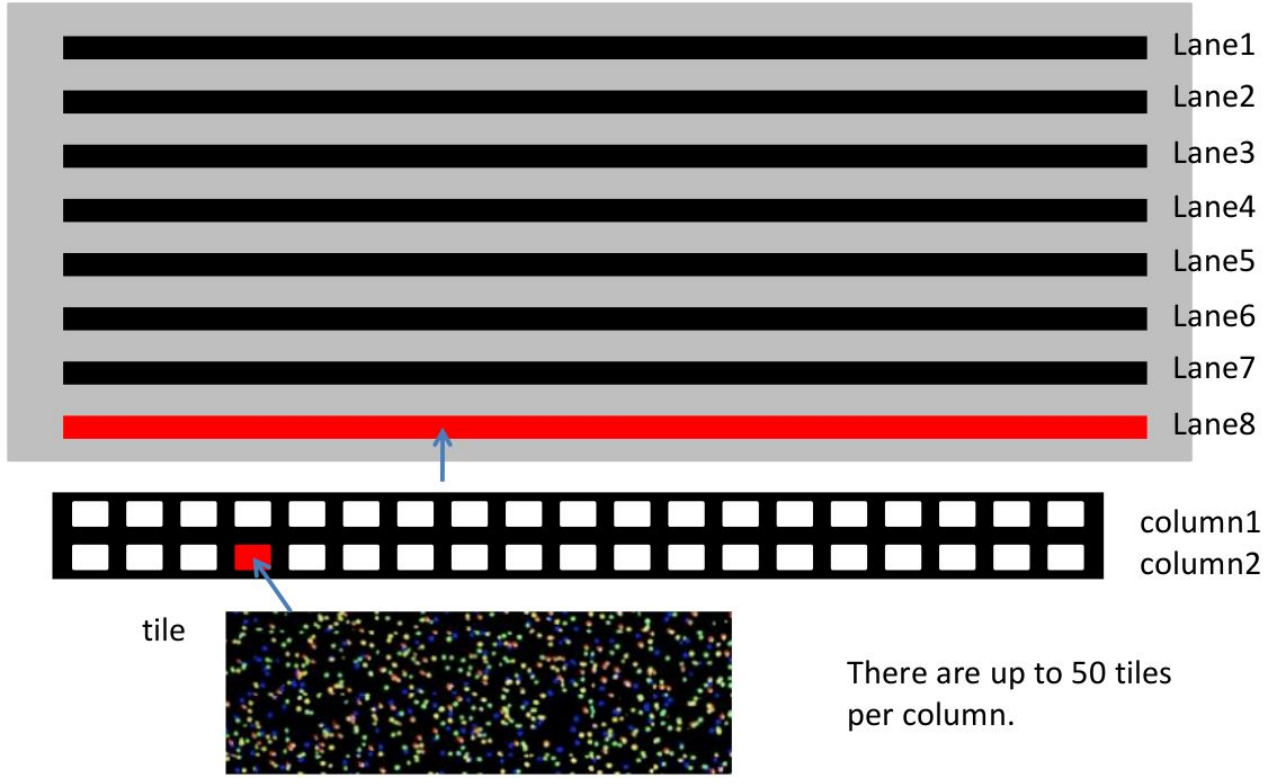
Illumina - sequencing 3



Illumina - sequencing 4

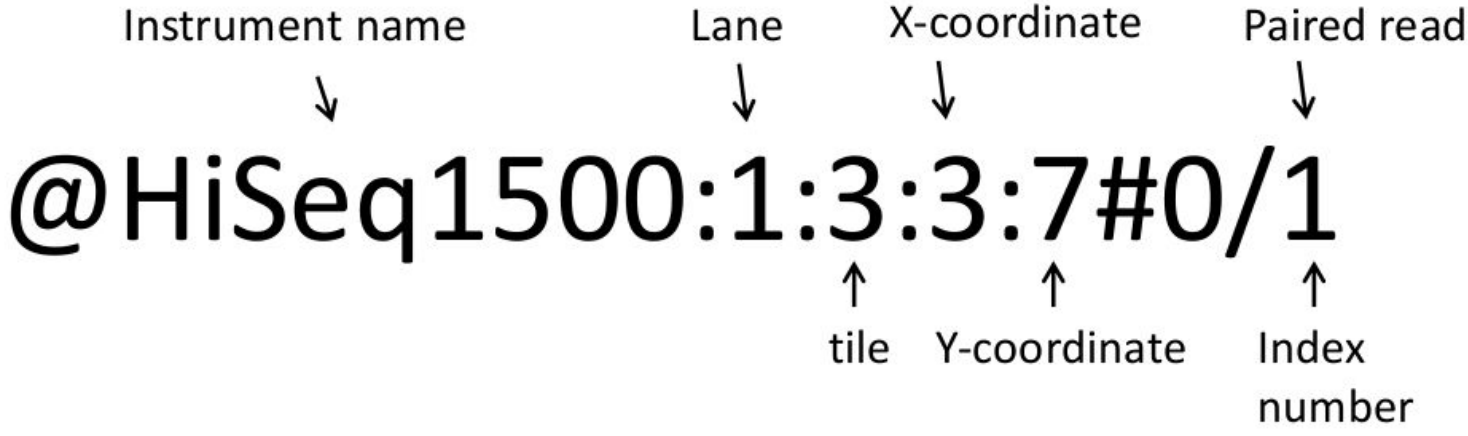


Illumina - flow cell layout

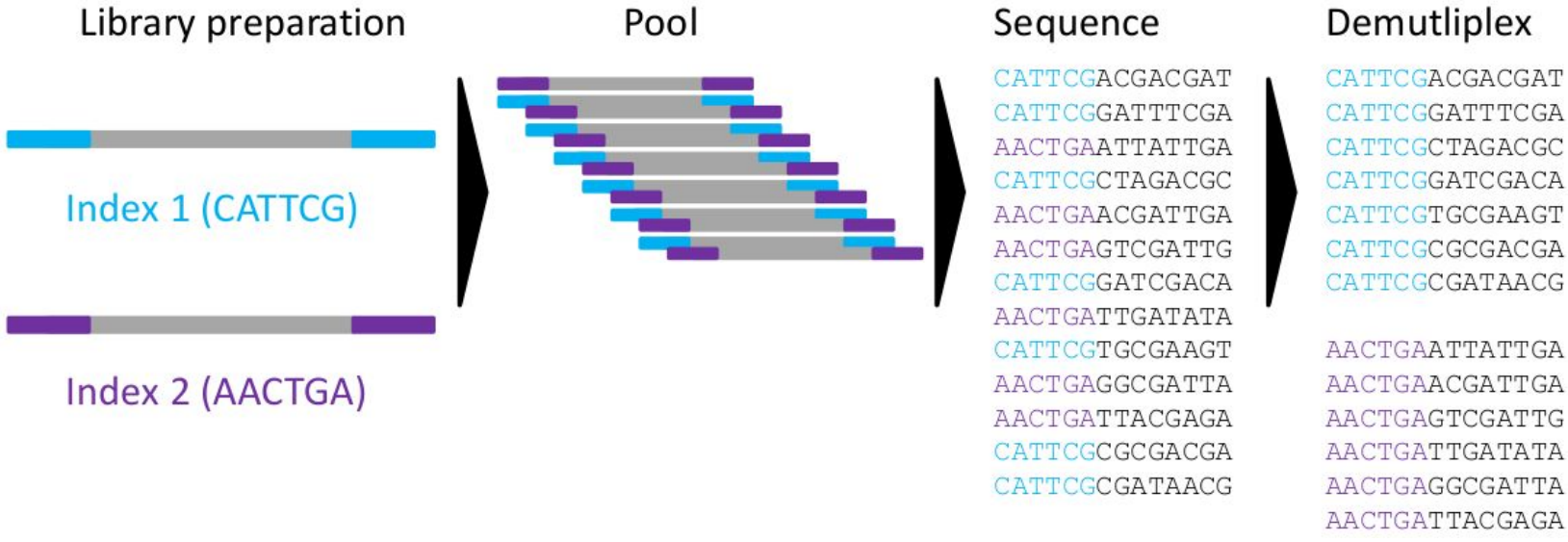


UNIVERSITÄT BONN

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

- Single end (SE):

- Paired-end (PE):

UNIVERSITÄT BONN

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:

  @HiSeq1500:1:3:3:7#0/1

  ACGTACGTACGT

  +

  “”?CB””:DC”

UNIVERSITÄT BONN

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	

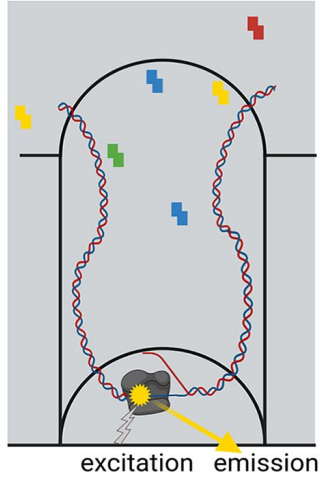
UNIVERSITÄT BONN

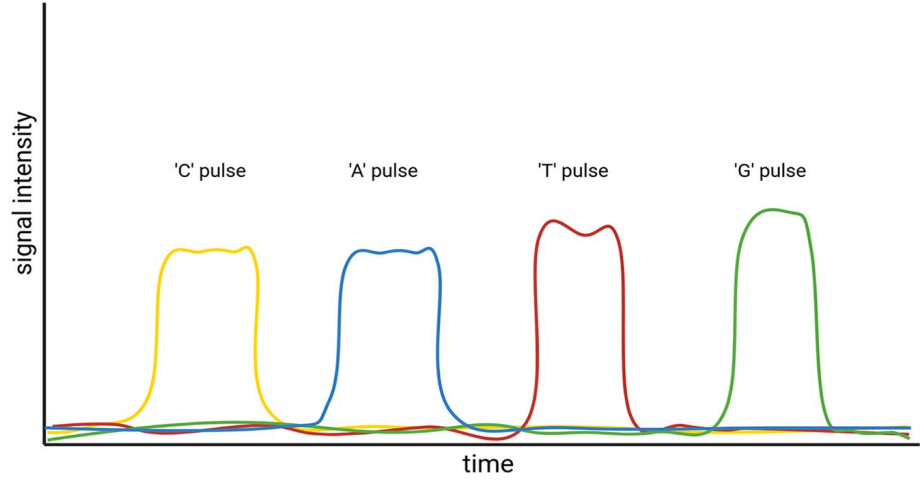
PacBio sequencing

Pacific Biosciences (PacBio)

- Polymerase located at bottom of well

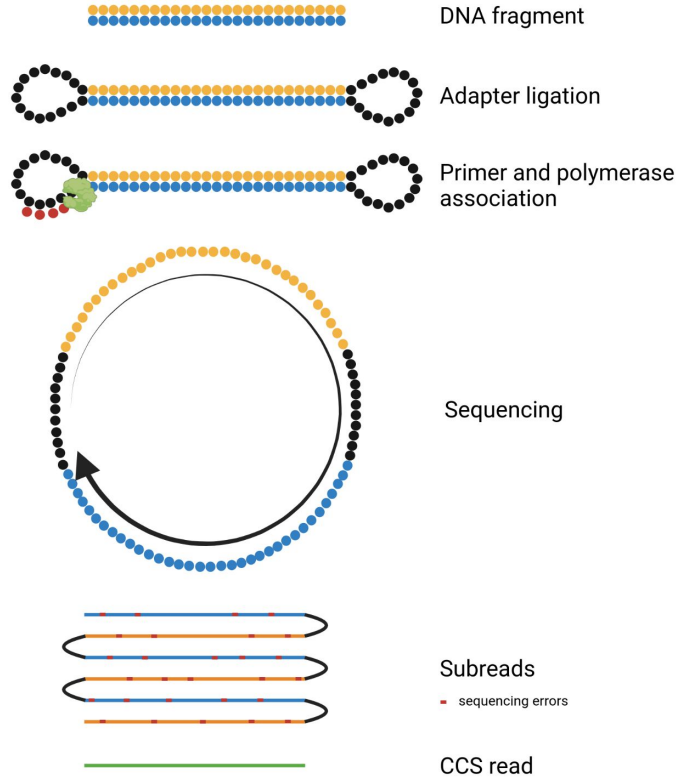
- ZMWs = Zero Mode Waveguide





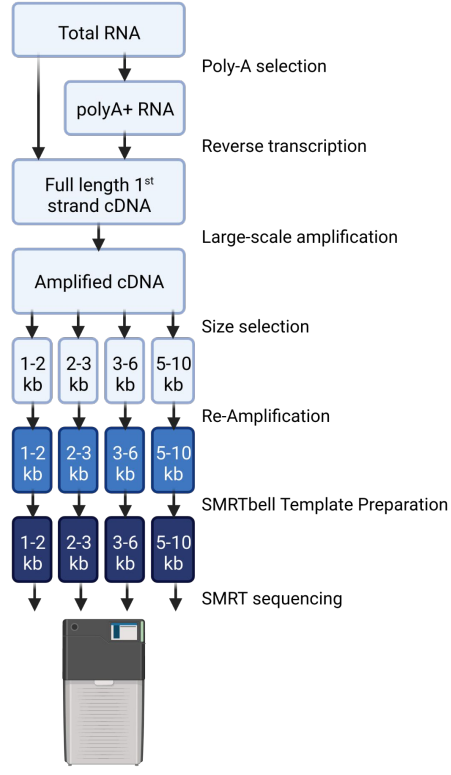
Pucker et al., 2022: 10.1017/qpb.2021.18

PacBio - HiFi


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

PacBio Iso-Seq (cDNA sequencing)

- SMRT = Single Molecule Real Time sequencing
- Full length cDNA sequencing is beneficial for gene prediction
- Iso-Seq generates several kb long reads and not only 2x300bp



UNIVERSITÄT BONN

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)

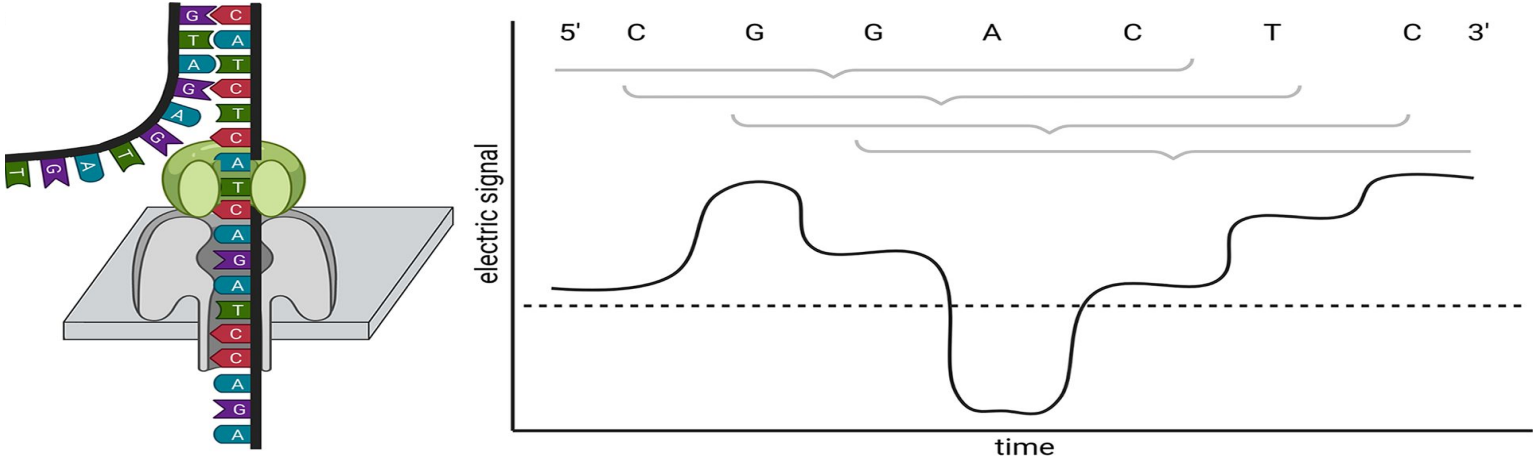
UNIVERSITÄT BONN

ONT sequencing

Oxford Nanopore Technologies (ONT)

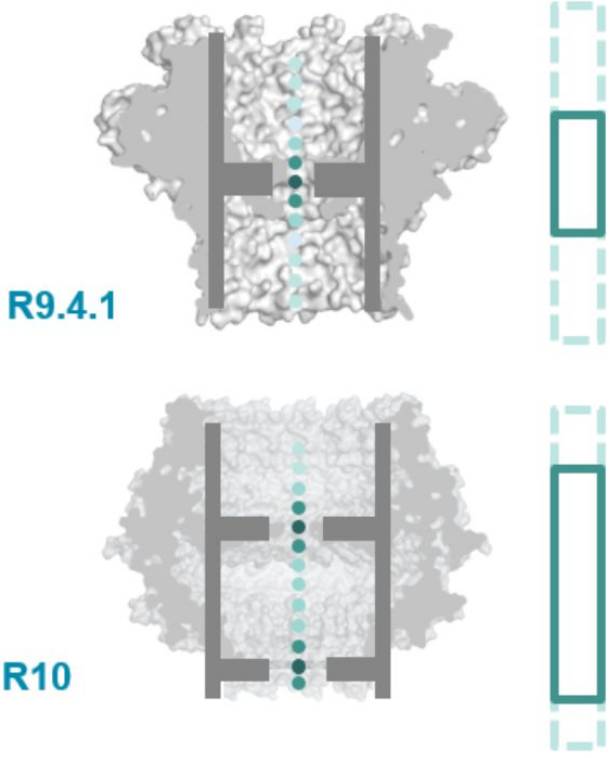
Under development since the 1980s

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



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

Nanopores



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

UNIVERSITÄT BONN

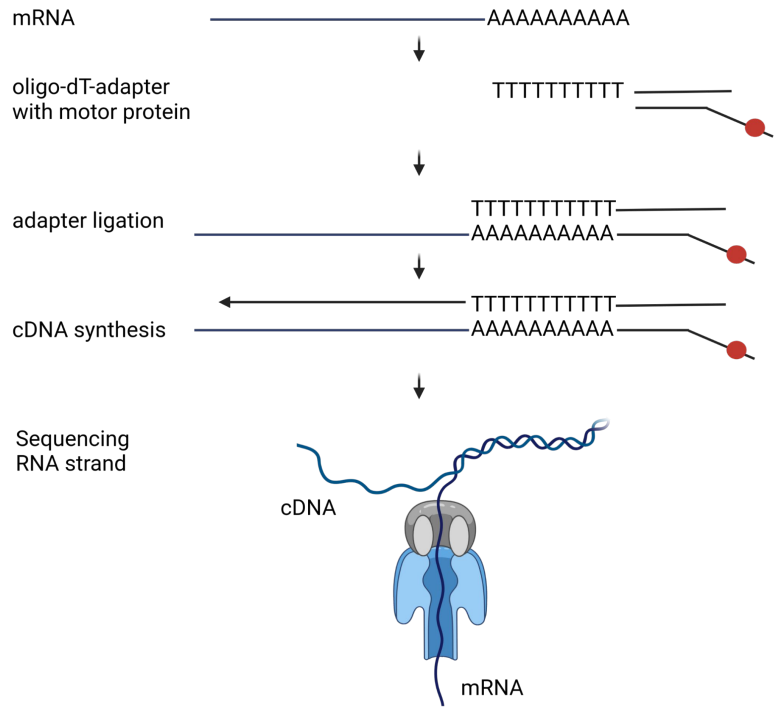
ONT sequencing workflow

		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

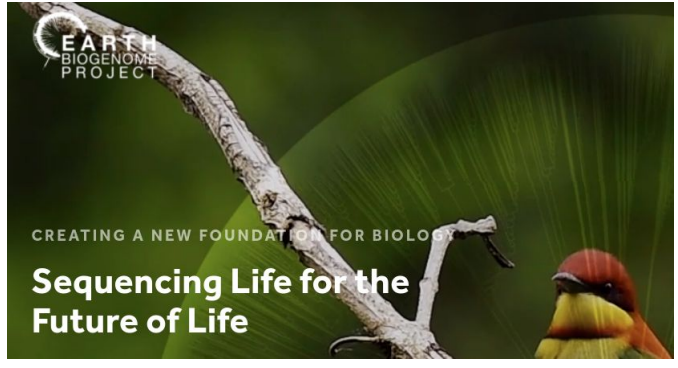
Direct RNA sequencing

- Only sequencing technology to analyze RNA directly at high throughput
- RNA sequencing requires adjusted data processing
- Full length sequences of RNAs are generated



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

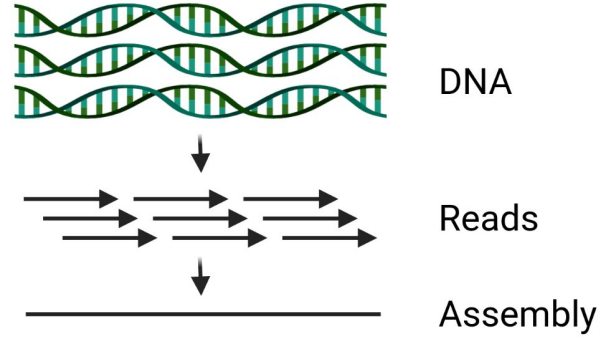
- Sequencing orphan crop genomes
- Portable and affordable sequencers
- Genome projects conducted by individual labs (not just sequencing centers)

UNIVERSITÄT BONN

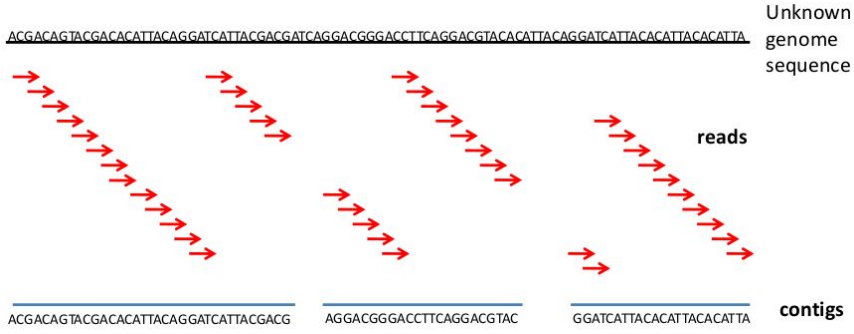
Genome sequence assembly

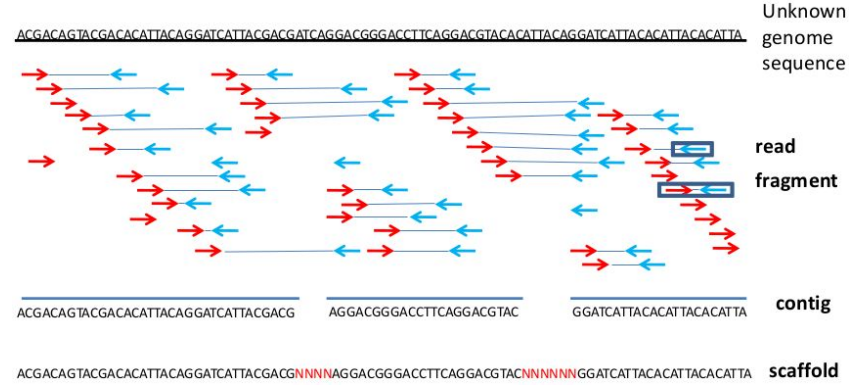
The assembly problem

- 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



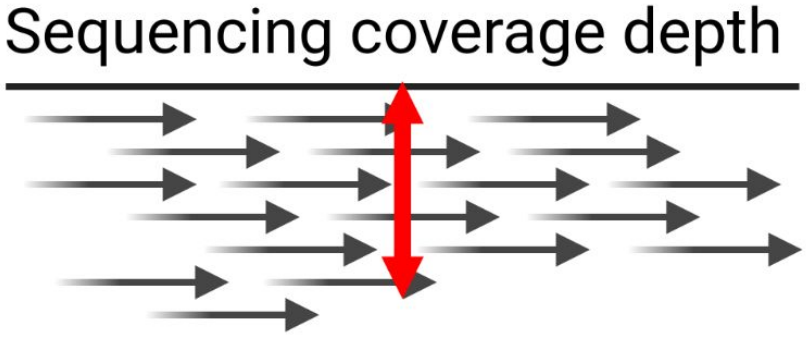
Contigs, scaffolds, pseudochromosomes

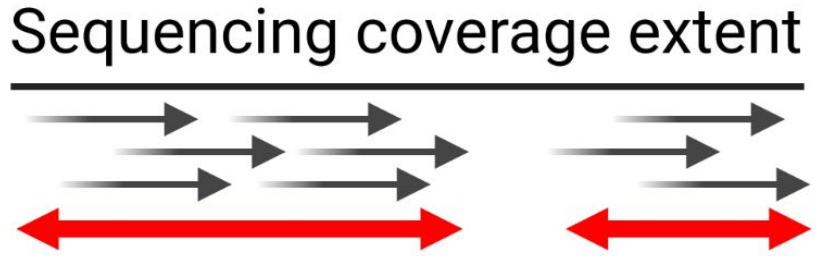




Pseudochromosomes = scaffolds representing an entire chromosome
Gaps = regions between contigs that are represented by Ns

Sequencing coverage depth vs. coverage extent





UNIVERSITÄT BONN

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

UNIVERSITÄT BONN

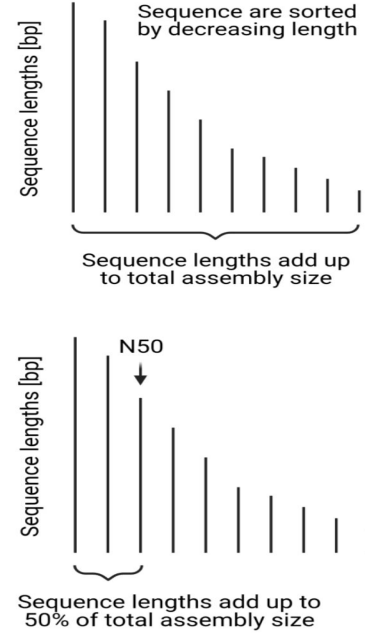
Assembly evaluation

- 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

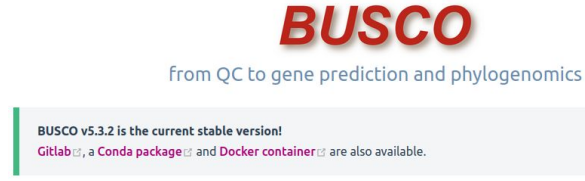


Pucker et al., 2019: 10.1371/journal.pone.0216233

UNIVERSITÄT BONN

Evaluation - completeness

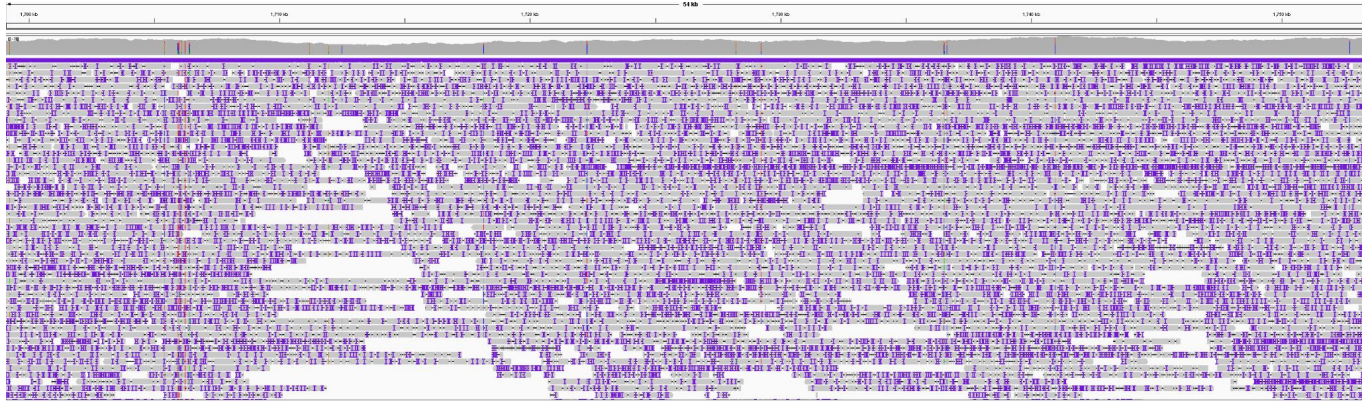
- Check assembly completeness: inspection for presence of conserved genes
- BUSCO = Benchmarking Universal Single-Copy Orthologue
- BUSCO genes are used to assess assembly completeness



UNIVERSITÄT BONN

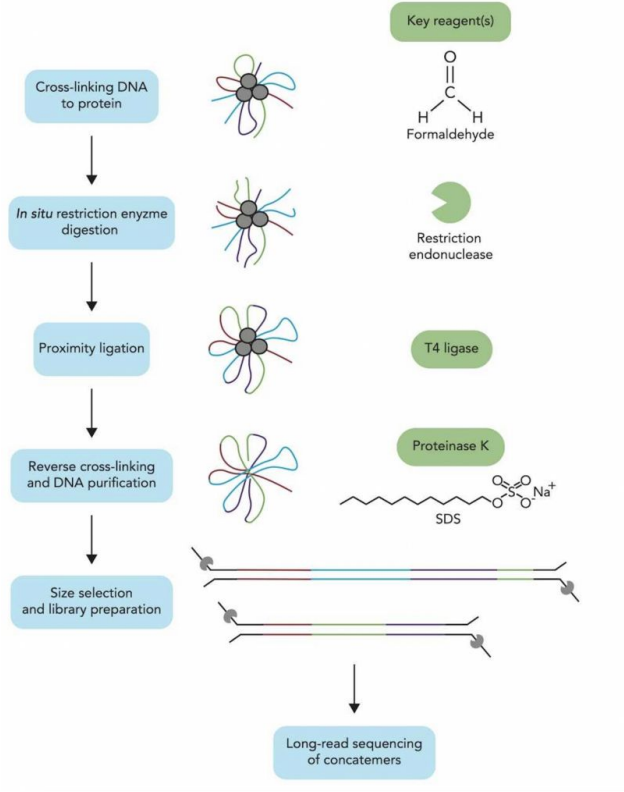
Evaluation - assembly correctness

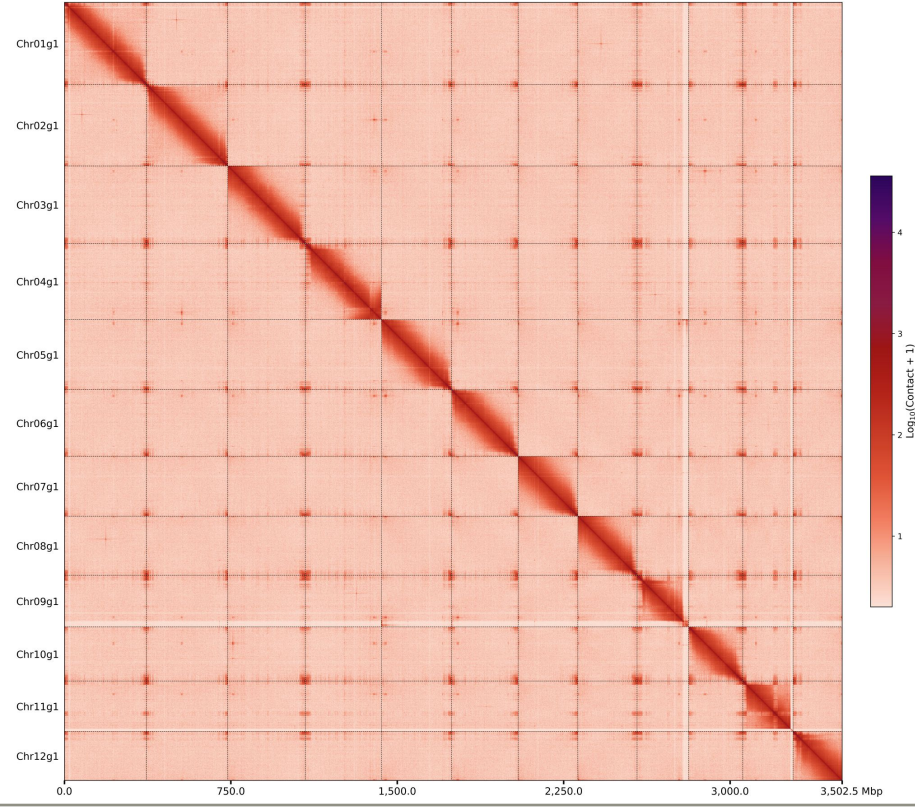
- Check assembly correctness: analyses of read mappings
- Integrative Genomics Viewer (IGV) can visualize read mappings
- Tools: REAPR, SQUAT



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

Scaffolding with Pore-C data





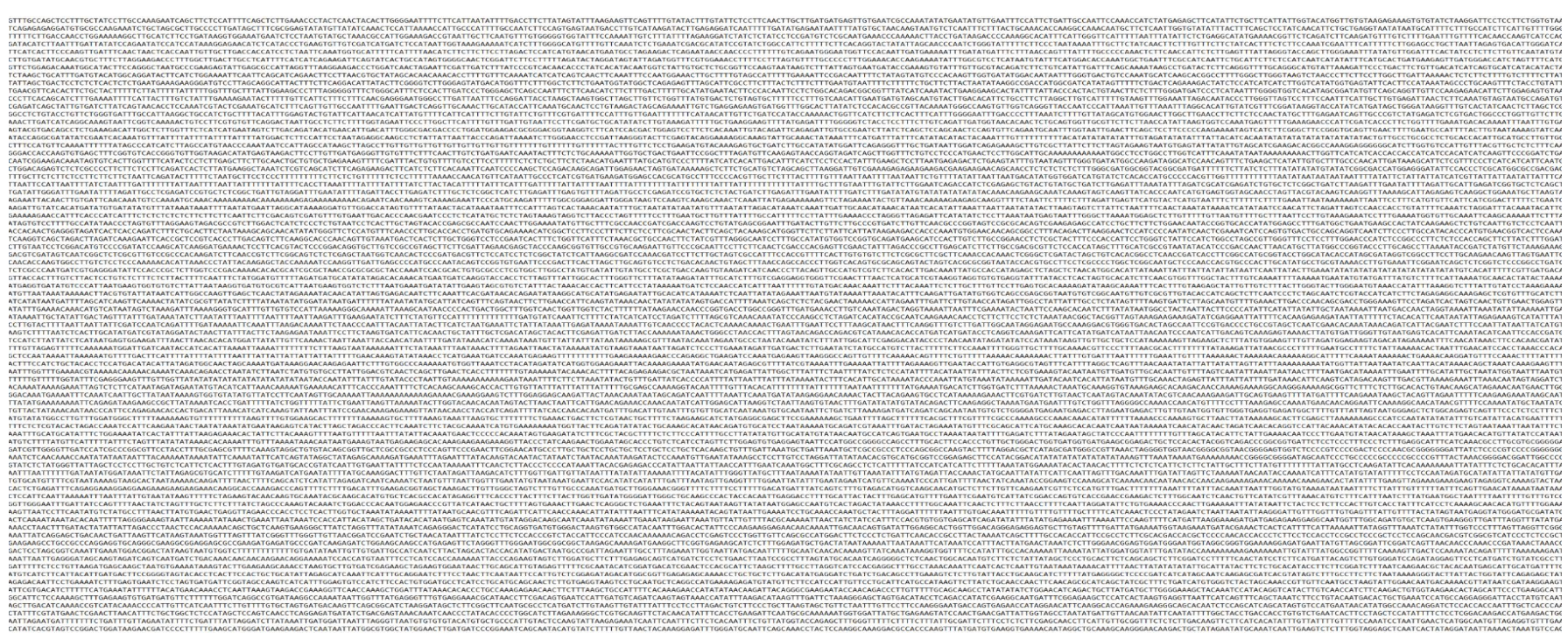
Nowak *et al.*, 2025: 10.1101/2024.06.15.599162
Evashepparddixon, CC BY-SA 4.0

UNIVERSITÄT BONN

Structural annotation

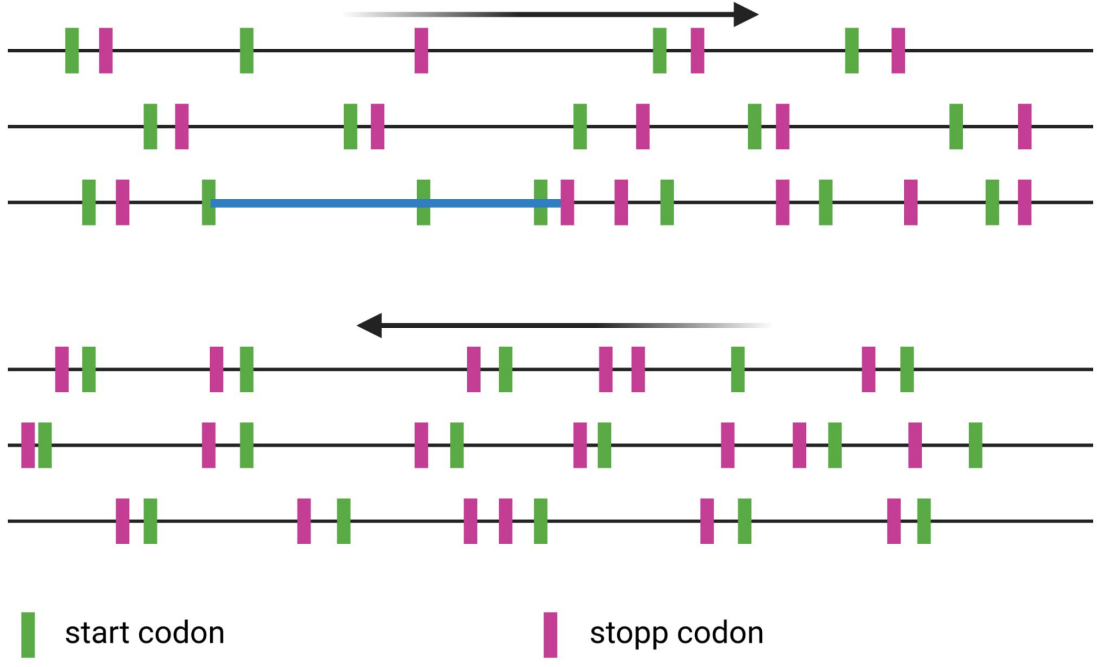
UNIVERSITÄT BONN

Finding genes in a genome sequence



Siadjeu & Pucker et al., 2020: 10.3390/genes11030274

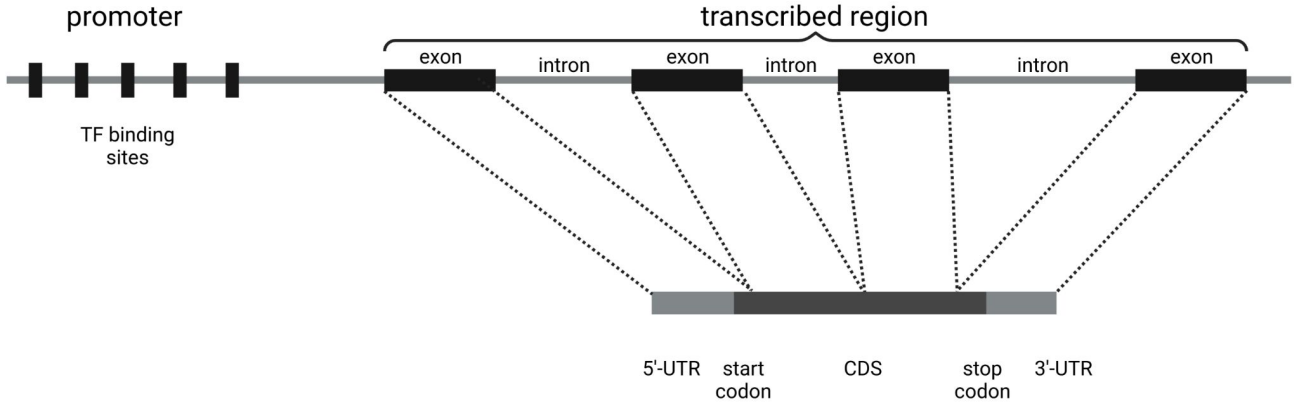
Finding ORFs



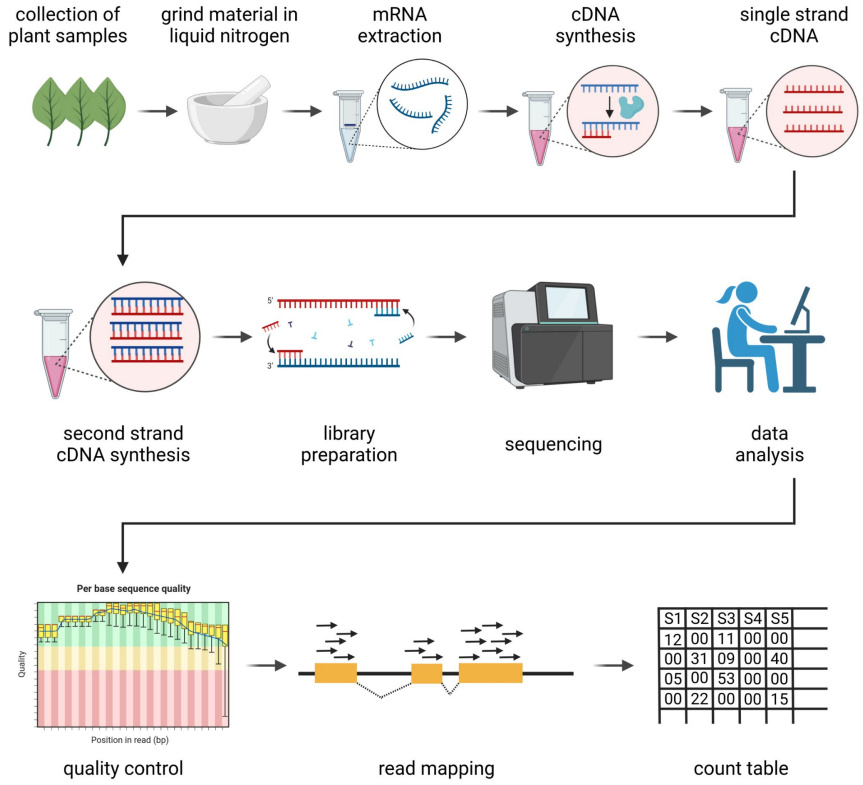
Plant gene structure

CDS = (Protein) Coding Sequence

ORF = Open Reading Frame

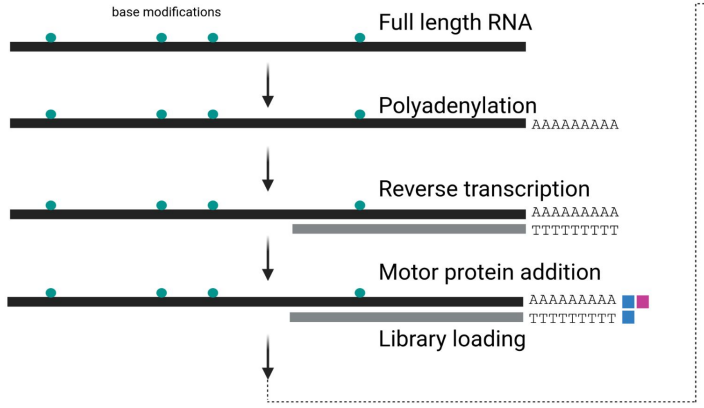


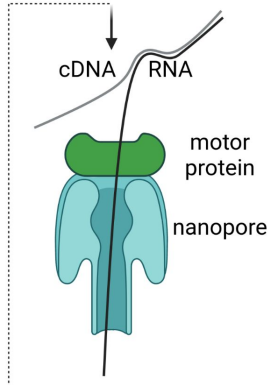
RNA-seq


collection of plant samples
grind material in liquid nitrogen
mRNA extraction
cDNA synthesis
single strand cDNA
second strand cDNA synthesis
library preparation
sequencing
data analysis
Per base sequence quality
Quality
Position in read (bp)
S1 S2 S3 S4 S5
12 00 11 00 00
00 31 09 00 40
05 00 53 00 00
00 22 00 00 15
quality control
read mapping
count table

Full length transcript sequences

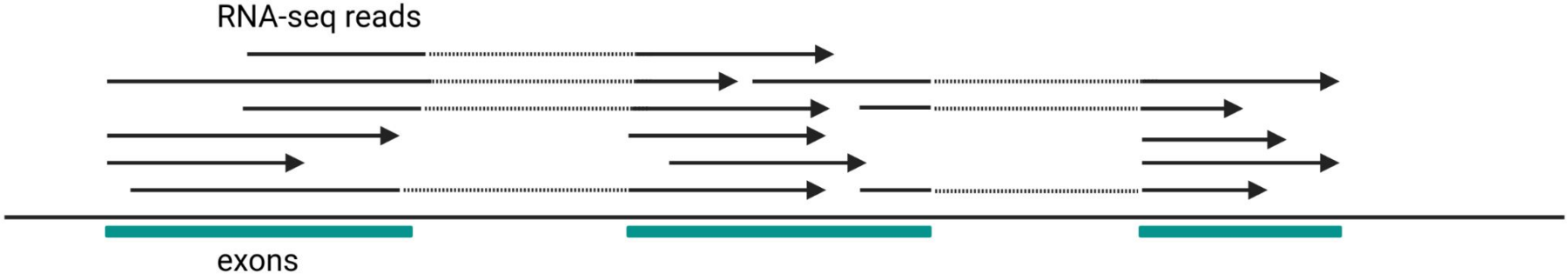
- Capturing full length transcripts for RNA-seq (Illumina)
- Full length cDNA sequencing (PacBio, ONT)
- Direct RNA sequencing (ONT)





RNA-seq hints

- Aligned RNA-seq reads indicate exon positions
- Splitting of reads indicates intron positions
- CDS can be identified as ORF within the covered regions

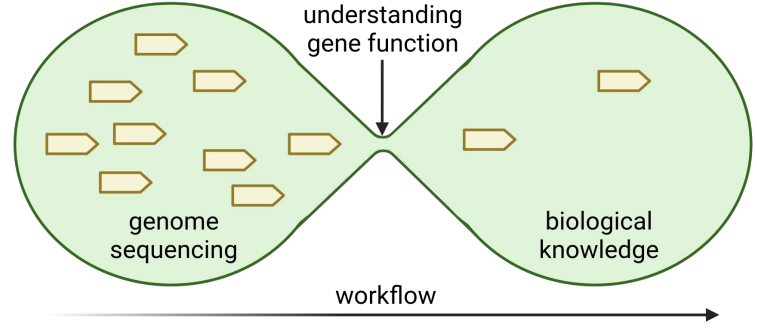


UNIVERSITÄT BONN

Functional annotation

Structural and functional annotation

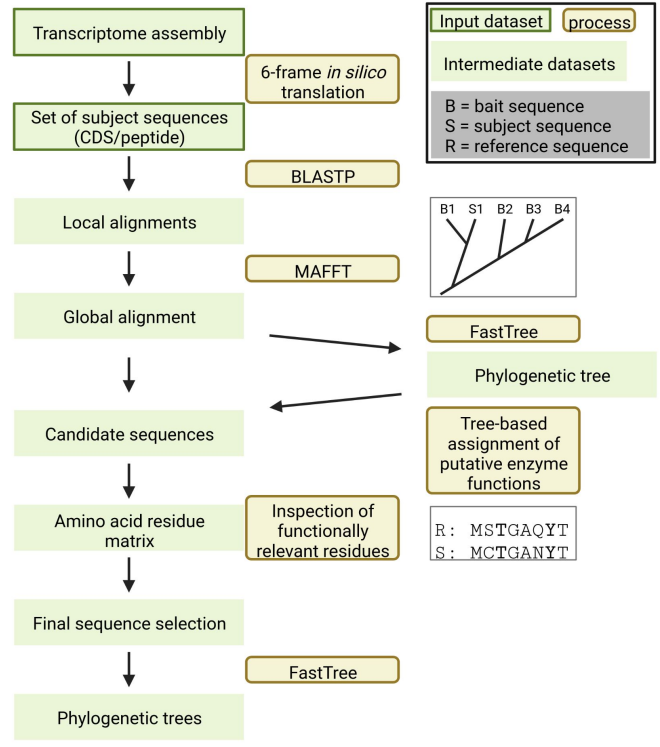
- Knock-out experiments are not scaleable to all species and genes
- Transfer of knowledge across plant species required
- Functional annotation is the new challenge
  - BLAST as basic analysis option
  - KIPEs: pathway annotation
  - Annotators for gene families



Pucker, 2024: 10.20944/preprints202402.0645.v1

Pathway annotation

- KIPEs = Knowledge-based Identification of Pathway Enzymes
- Functionally relevant amino acid residues known
- Automatic screening of large data sets
- Systematic comparisons of flavonoid biosynthesis across species
- Identification of targets for characterization

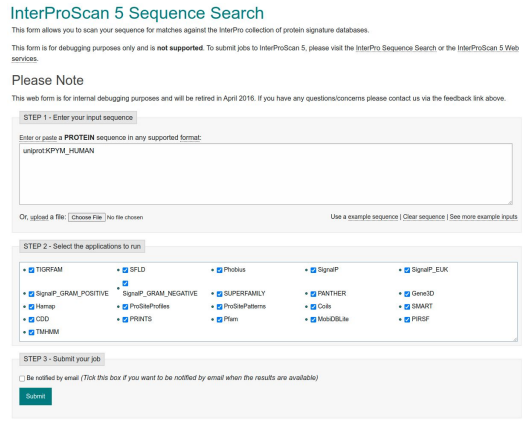


Pucker, Reiher, Schilbert, 2020: 10.3390/plants9091103
Rempel, Choudhary, Pucker, 2023: 10.1101/2022.06.30.498365

UNIVERSITÄT BONN

InterProScan5

- Comparison vs. multiple databases (KEGG, GO, PANTHER, ….)
- Efficient tool for generating an overall annotation



Results for job iprscan5-I20220320-102557-0980-87120812-p2m

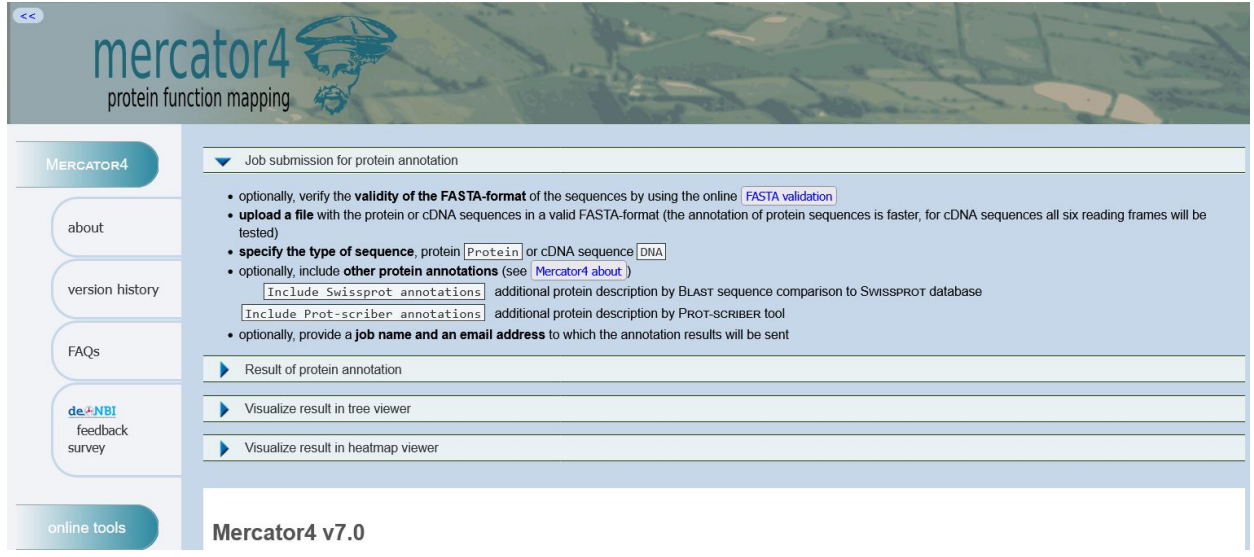
Tool Output | Submission Details

Download in XML format | Download in TSV format | Download in GFF3 format | Download in SVG format | Download HTML tarball file | Download in JSON format

sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	Gene3D	G3DSA:3.40.47.10	-	1	241	4.8E-101	T	20-03-2022	
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	SUPERFAMILY	SSF53901	Thiolase-like	241	393	2.98E-51	T	20-03-20	
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	SUPERFAMILY	SSF53901	Thiolase-like	10	237	2.38E-78	T	20-03-20	
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	Pfam	PF02797	Chalcone and stilbene synthases, C-terminal domain		244	394	1.5E-71		
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	PANTHER	PTHR11877:SF81	BNAA02G30320D PROTEIN	6	394	0.0	T	20-03-2022	
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	Gene3D	G3DSA:3.40.47.10	-	242	395	9.7E-62	T	20-03-2022	IPR01603
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	PANTHER	PTHR11877	HYDROXYMETHYLGLUTARYL-COA SYNTHASE		6	394	0.0	T	
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	CDD	cd00831 CHS_like		21	390	0.0	T	20-03-2022	-
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	PIRSF	PIRSF000451	PKS_III	7	394	0.0	T	20-03-2022	IPR011141
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	Pfam	PF00195	Chalcone and stilbene synthases, N-terminal domain		10	233	2.9E-119		
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	ProSitePatterns	PS00441	Chalcone and stilbene synthases active site.		161	177	-		

http://www.ebi.ac.uk/Tools/services/web/toolform.ebi?tool=iprscan5&sequence=uniprot:KPYM_HUMAN
Jones et al., 2014: 10.1093/bioinformatics/btu031

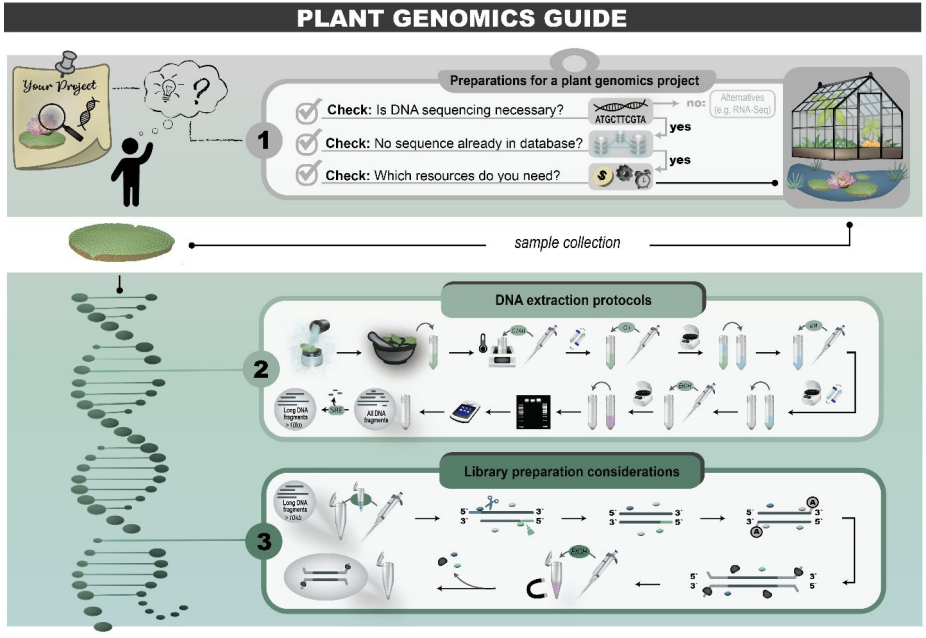
- Upload of FASTA file containing polypeptide sequences
- Annotation through hierarchical sequence clustering

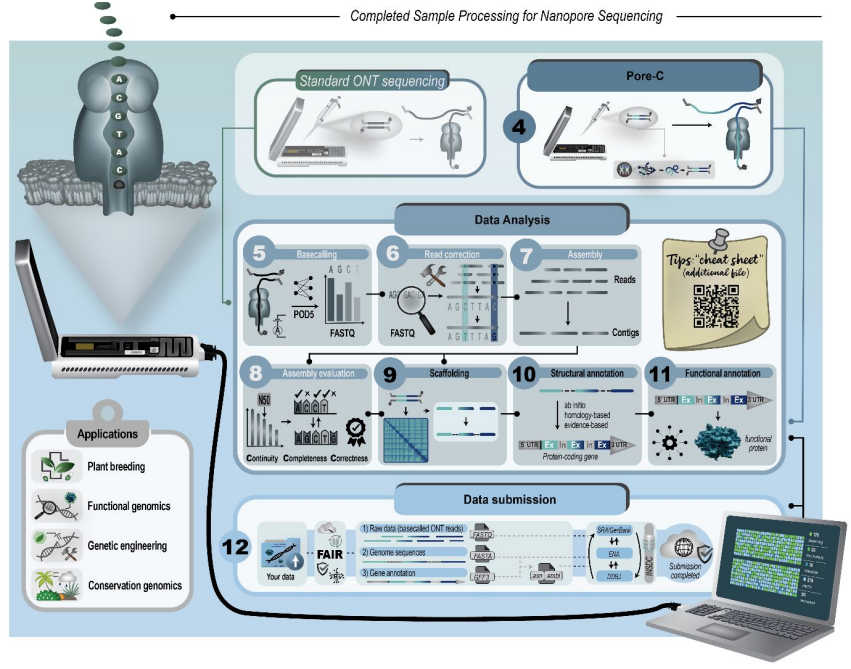


https://www.plabipd.de/mercator_main.html

UNIVERSITÄT BONN

Plant Genomics Workflow





Supplementary file with all commands!

de Oliveira et al., 2025: 10.20944/preprints202508.1176.v2

UNIVERSITÄT BONN

Summary

- History of DNA
- Sequencing methods: Sanger, Illumina, PacBio, ONT
- Genome sequence assembly
- Structural annotation
- Functional annotation

UNIVERSITÄT BONN

Time for questions!

Questions

1. Which sequencing technologies are suitable for plant genomics?
2. How strong is the correlation between plant genome size and number of genes?
3. What is the phred score?
4. What is the structure of FASTQ and FASTA?
5. What is needed to get a good structural annotation?
6. What are the current challenges in plant genomics?

Literature

- de Oliveira, J. A. V. S.; Choudhary, N.; Meckoni, S. N.; Nowak, M. S.; Hagedorn, M.; Pucker, B. (2025). Cookbook for Plant Genome Sequences. doi: 10.20944/preprints202508.1176.v2.

- Wolff, K.; Friedhoff, R.; Schwarzer, F.; Pucker, B. (2023). Data Literacy in Genome Research. Journal of Integrative Bioinformatics, 2023, pp. 20230033. doi: 10.1515/jib-2023-0033.

- Pucker B, Irisarri I, de Vries J and Xu B (2022). Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. Quantitative Plant Biology, 3, E5. doi: 10.1017/qpb.2021.18.