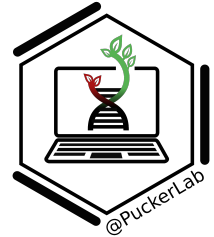
UNIVERSITÄT BONN



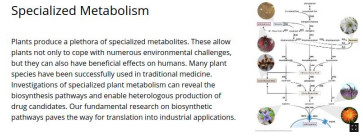
Prof. Dr. Boas Pucker

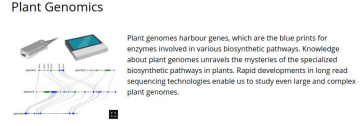
Long Read Genomics

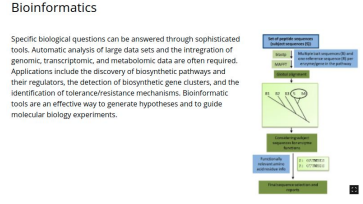
- Sequencing Technologies

Boas Pucker - CV

- Biochemistry at HHU Düsseldorf
- (Systems) Biology at Bielefeld University
- Doctoral student (CeBiTec, Bielefeld University)
  - Genomics & Bioinformatics; synthetic biology (iGEM)
- Post doc (Ruhr-University Bochum)
- Post doc (Department of Plant Sciences, Cambridge, UK)
- Plant Biotechnology & Bioinformatics, TU Braunschweig (2021-2025)
  - Specialized plant metabolites, applied bioinformatics
- University of Bonn (since 2025): Plant Biotechnology & Bioinformatics



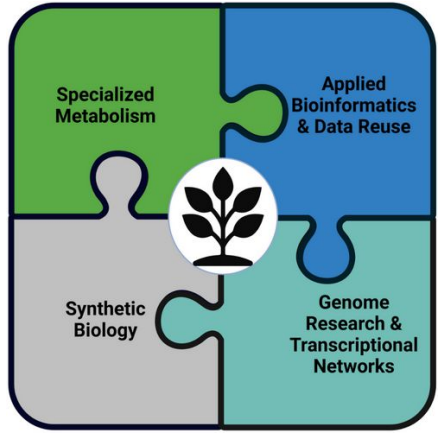




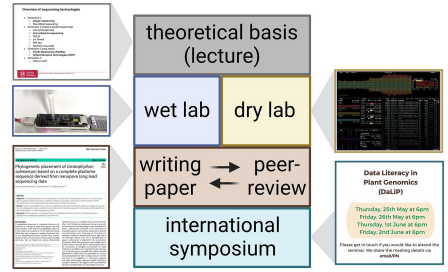
UNIVERSITÄT BONN

Plant Biotechnology & Bioinformatics

- Flavonoids: Anthocyanins, Flavonos, Flavones, Proanthocyanidins
- Withanolides
- Carotenoids
- …



Tools: KIPES, NAVIP, MGSE, MYB/bHLH_annotator, DupyliCate…



Wolff et al., 2023: 10.1515/jib-2023-0033
Dassow et al., 2025: 10.1101/2025.01.31.635922

UNIVERSITÄT BONN

Availability of slides

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

- Questions: Feel free to ask at any time

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

UNIVERSITÄT BONN

Table of content

UNIVERSITÄT BONN

Discovery of DNA

- Isolation of DNA by Friedrich Miescher in 1869
- Working in the castle in Tübingen
- Used a kitchen as lab

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

What is the genome of a plant?

UNIVERSITÄT BONN

Major types of DNA in plants

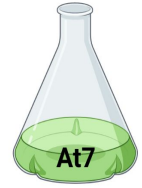
- gDNA from the nucleus
- mtDNA from the mitochondria (chondrome)
- cpDNA from the chloroplast (plastome)
- pDNA (plasmids, only in biotechnological applications)

UNIVERSITÄT BONN

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)



Busche et al., 2019: 10.1534/g3.119.400847
Siadjeu & Pucker et al., 2020: 10.3390/genes11030274
Sielemann et al., 2023: 10.1186/s12864-023-09823-2
Pucker et al., 2019: 10.3390/genes10090671

UNIVERSITÄT BONN

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

UNIVERSITÄT BONN

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

UNIVERSITÄT BONN

DNA extraction methods

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

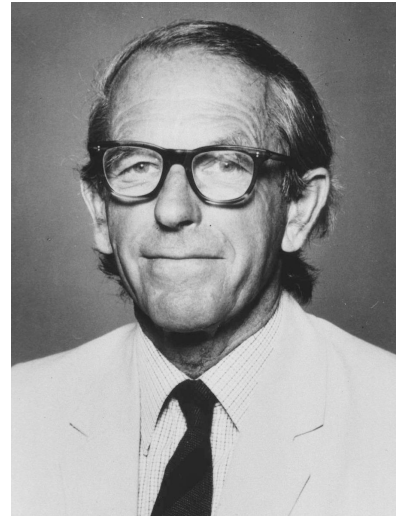
- Plasmid DNA
  - TELT: cheap and good quality for small plasmids
  - Alkaline lysis: cheap and good quality
  - Standard plasmid isolation kit: high quality but expensive

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)

UNIVERSITÄT BONN

Do you know other sequencing methods?

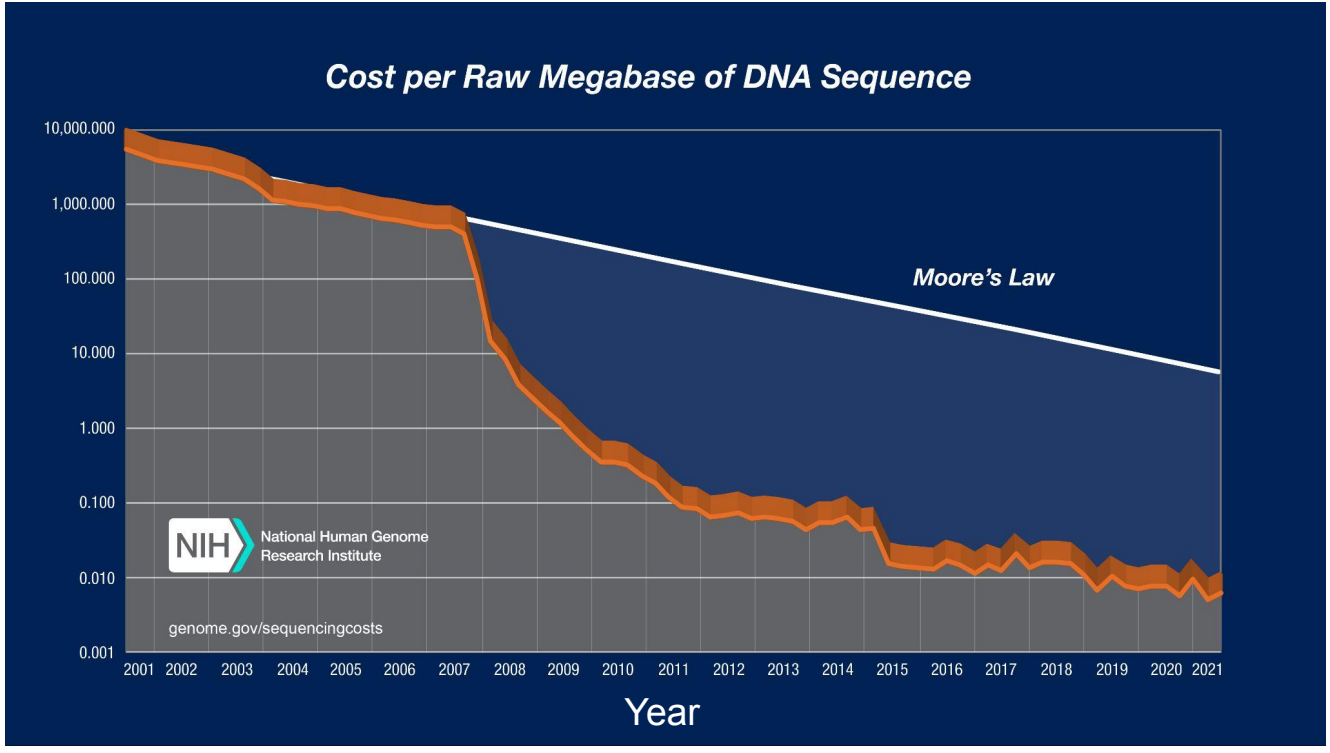
UNIVERSITÄT BONN

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?

UNIVERSITÄT BONN

Development of sequencing capacity

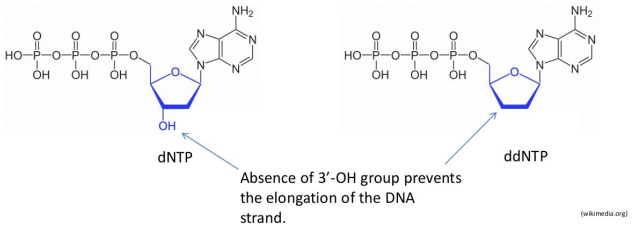


https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

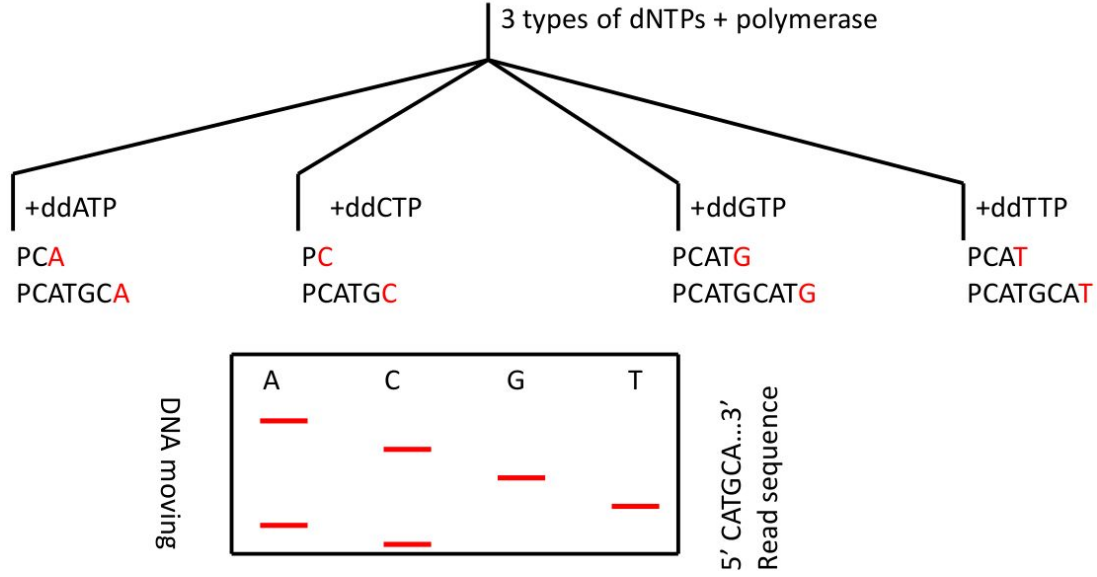
UNIVERSITÄT BONN

Sanger sequencing

Concept of Sanger sequencing

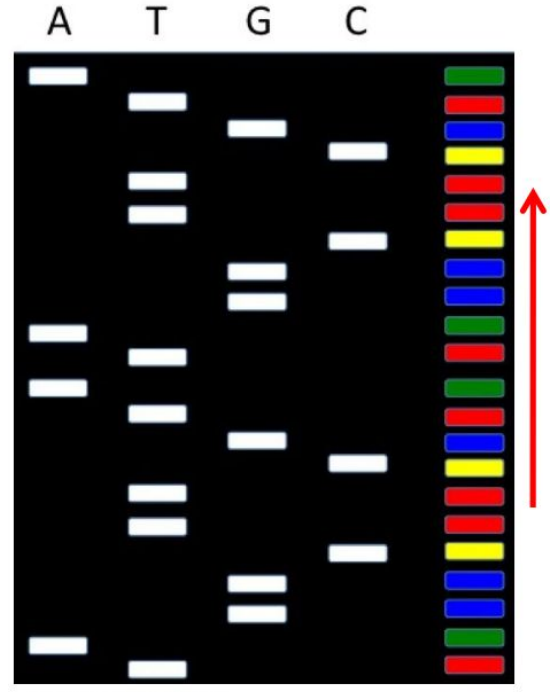


Primer (P): 5' –TGCATGGCATGATGCATG-3'
Template: 3' –ACGTACCGTACTACGTACGTACGTACGTCTAGGT-5'



Sanger sequencing - original version

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



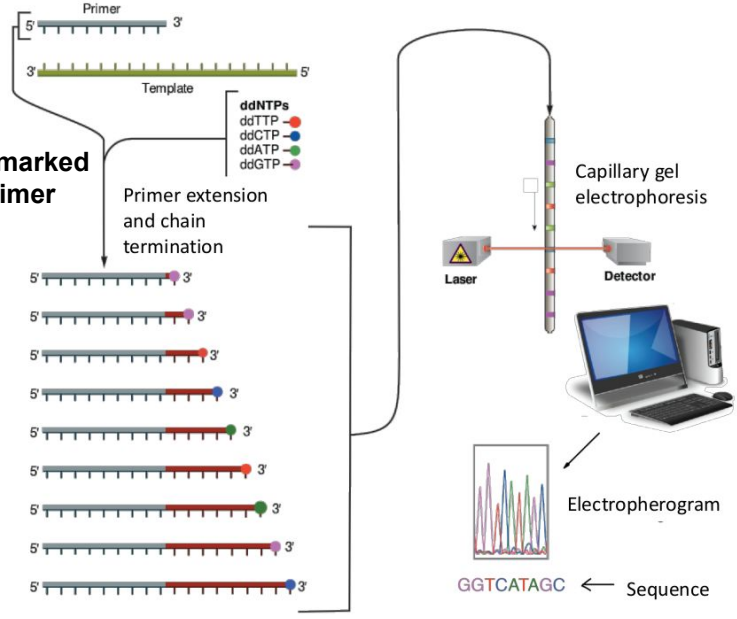
(modified from wikimedia.org)

Sanger sequencing - today

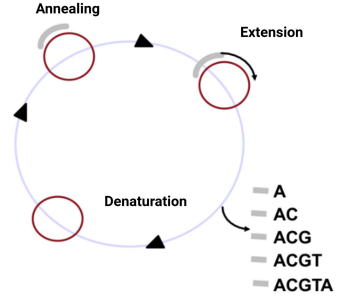
Only one reaction!

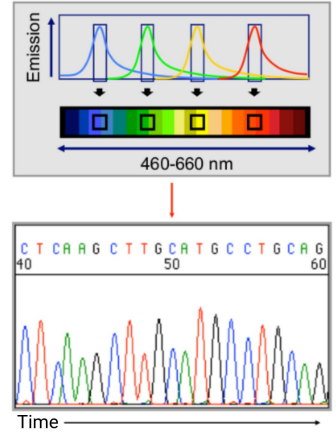
ddNTPs are marked instead of primer

Low input required due to cycle sequencing



Capillary avoids interference of adjacent lanes on a gel!





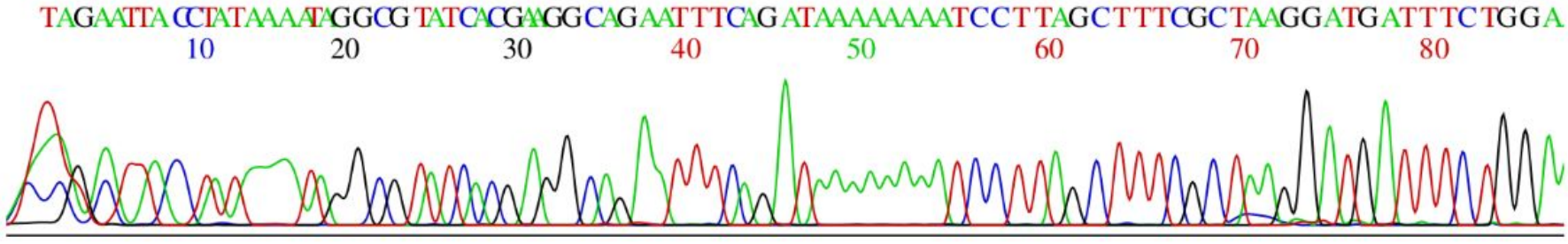
Figures modified from wikipedia

UNIVERSITÄT BONN

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
```

Phred-Score

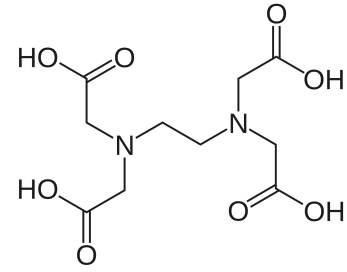
- Negative logarithm of the error probability for given position in read

- 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%

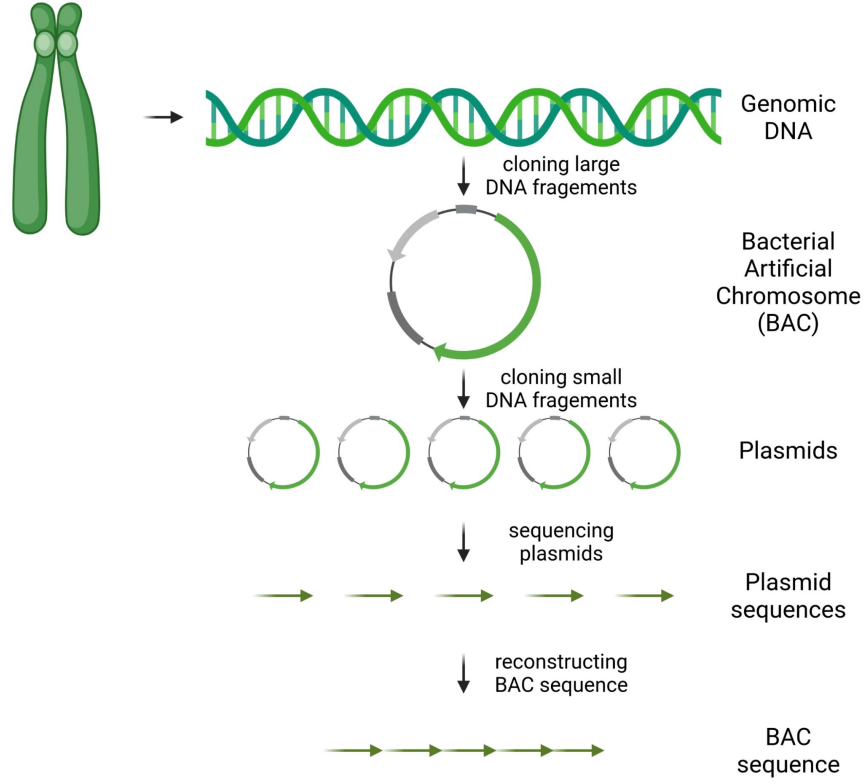
Sanger sequencing - practical consideration

- 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



UNIVERSITÄT BONN

BAC-based sequencing strategy



UNIVERSITÄT BONN

What is a contig/scaffold?

UNIVERSITÄT BONN

Contigs and scaffolds

- Contig = continuous sequence

```
ACGATACAGTAACCTACCGGATTAGACT
```

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

```
ACGATACAGTAACCTACCGGATTAGACTNNNNNNNACGACGGATACGTACAGTNNNNNNNNATTAGACCA
```

UNIVERSITÄT BONN

Sequencing the *Arabidopsis thaliana* genome

2000 — 2005 — 2010 — 2015 — 2020

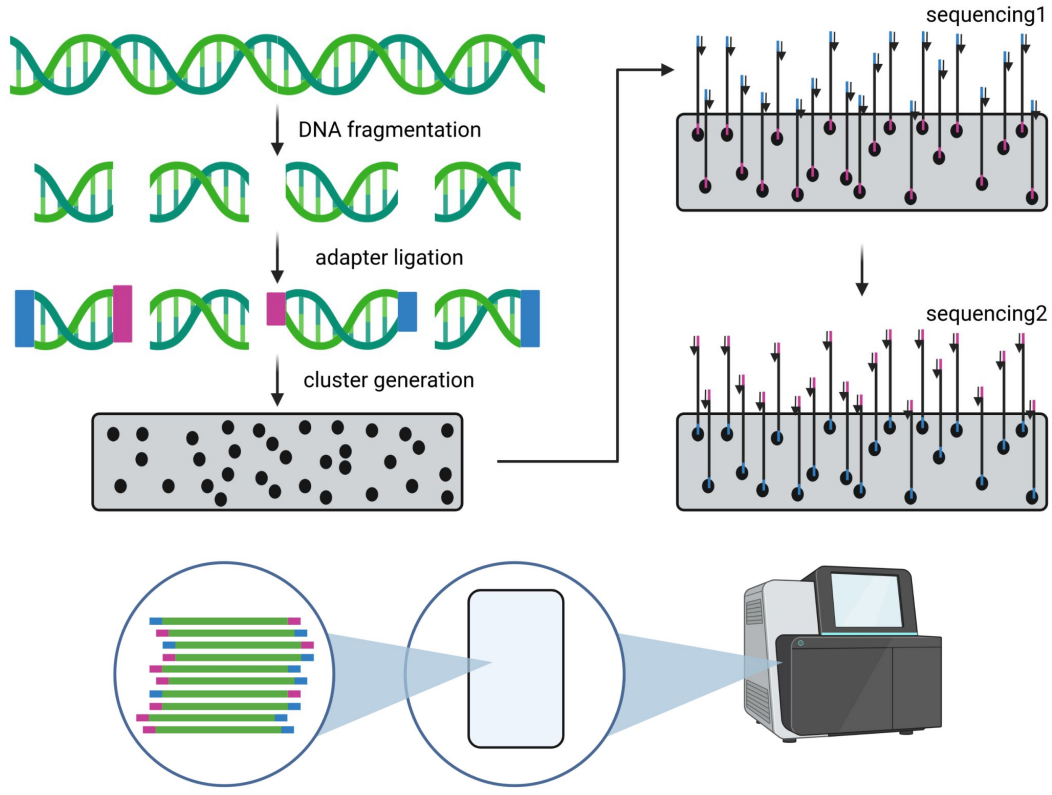
2000	2005	2010	2015	2020
Col-0 genome sequence release	Annual improvements of reference sequence	Re-sequencing projects	Ler and Nd-1 genome sequence release	Many more genome sequences released
The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. 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

UNIVERSITÄT BONN

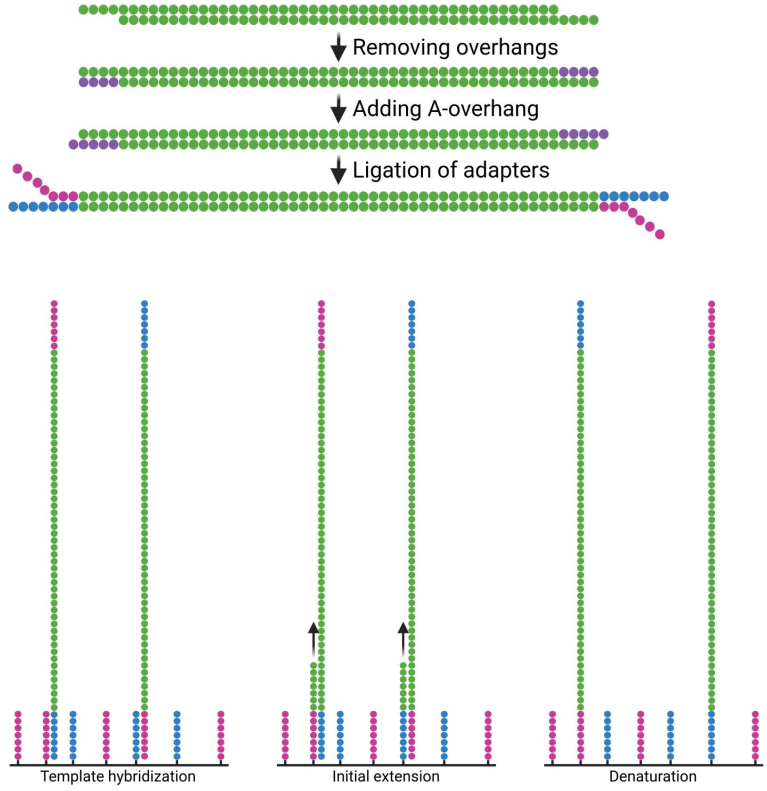
Illumina sequencing

UNIVERSITÄT BONN

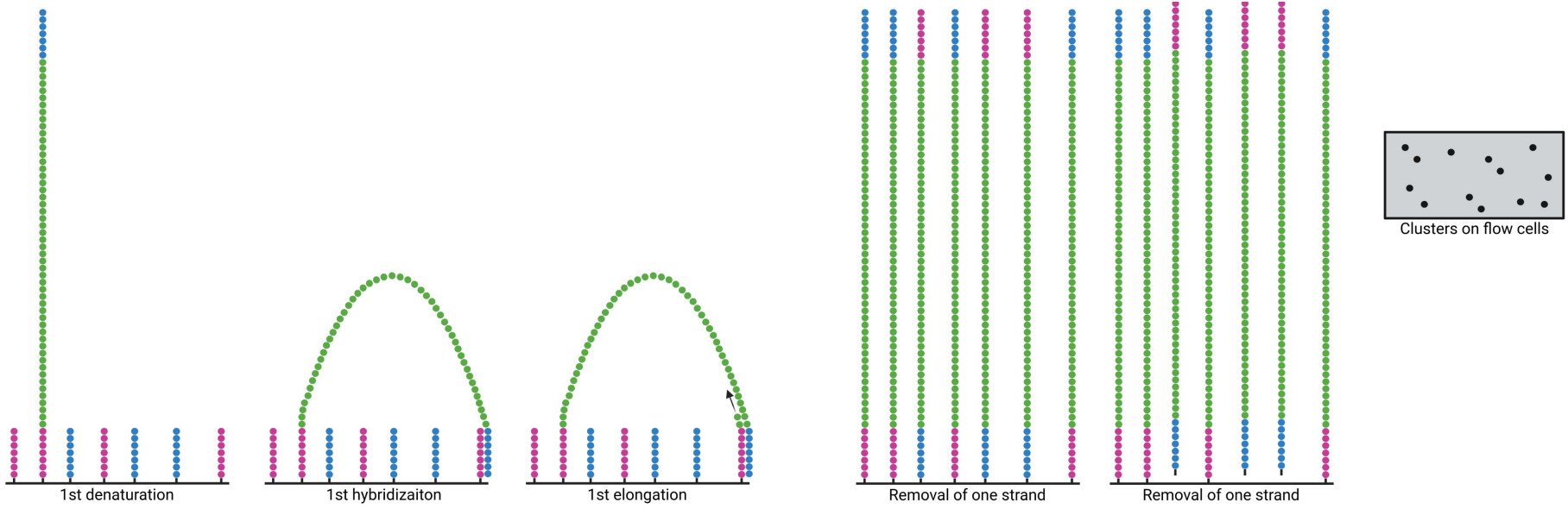
Illumina - sequencing overview



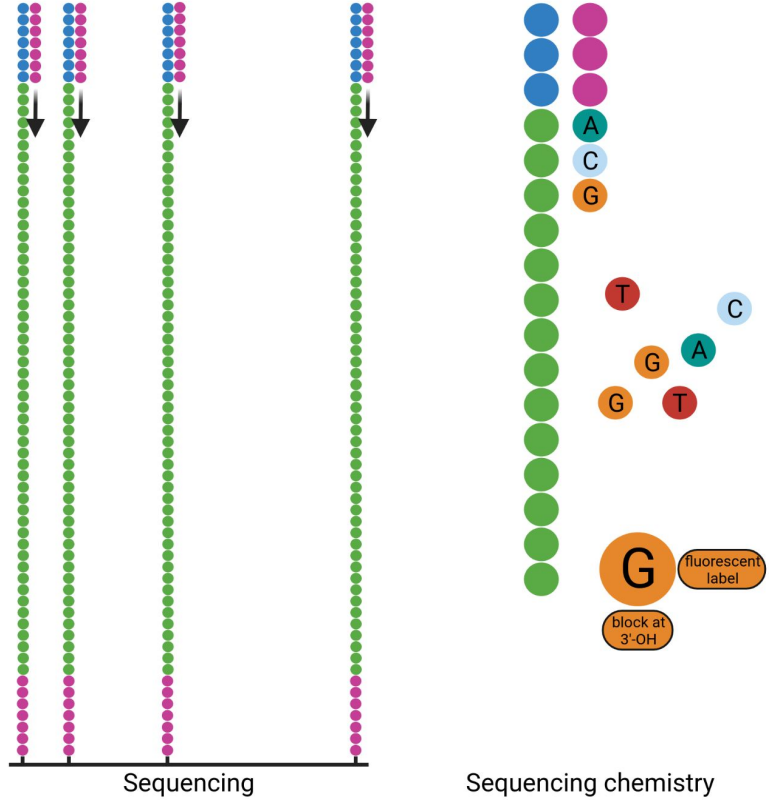
Illumina - sequencing 2



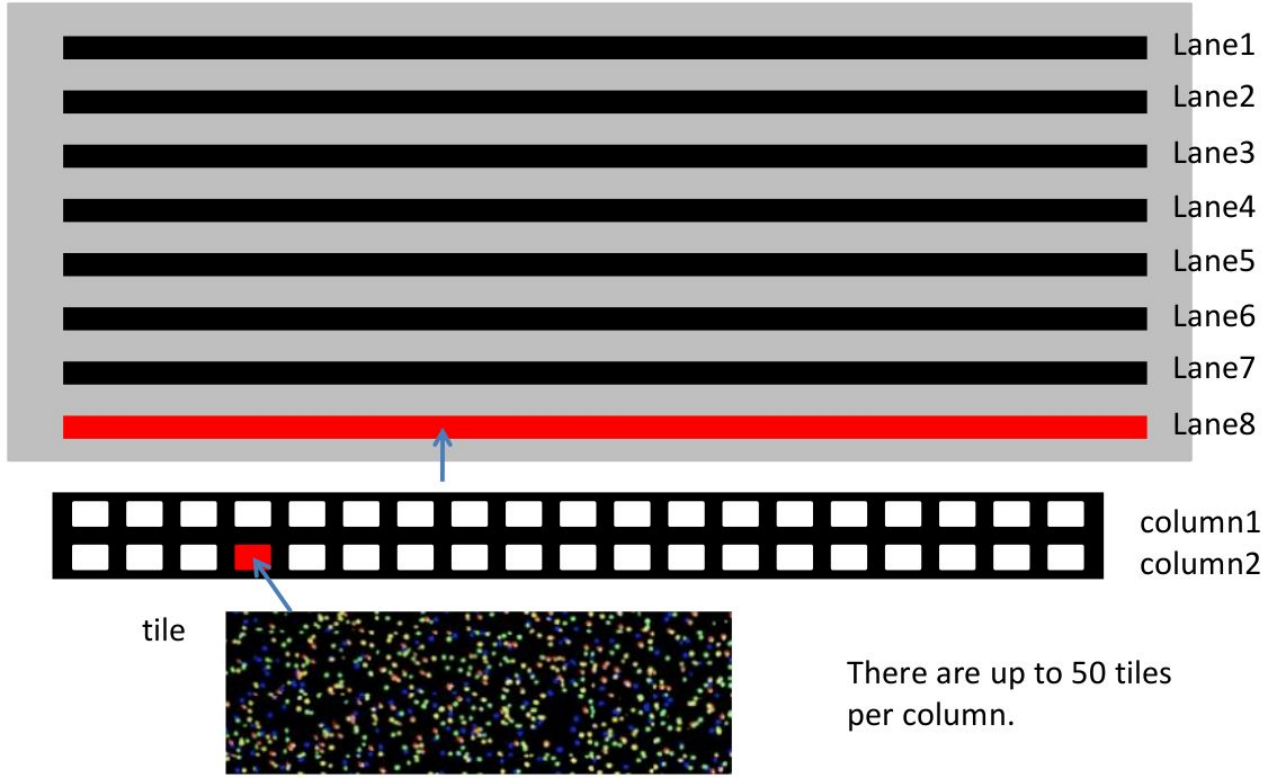
Illumina - sequencing 3



Illumina - sequencing 4

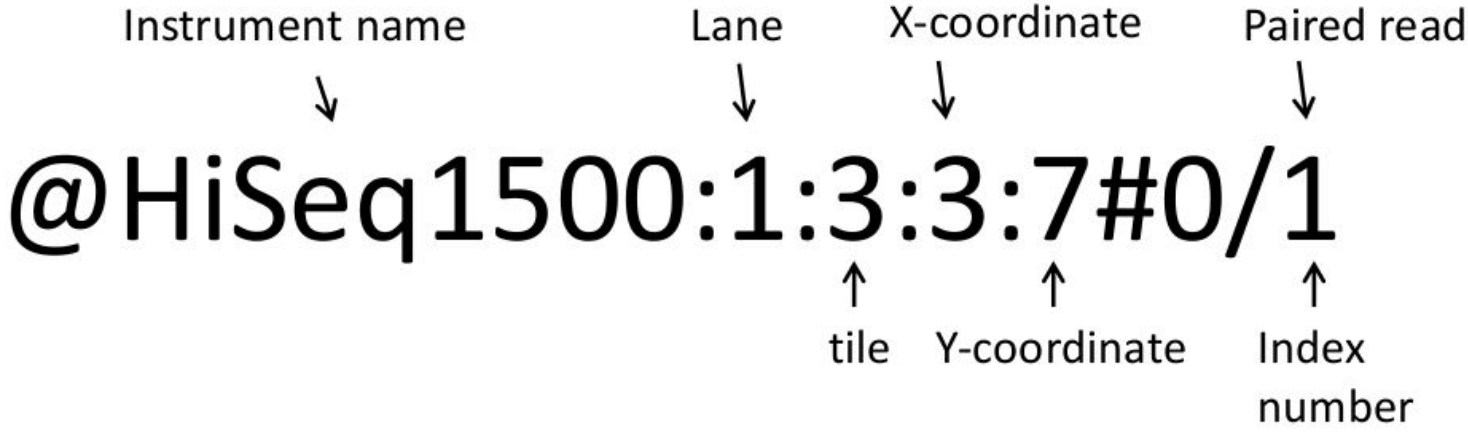


Illumina - flow cell layout

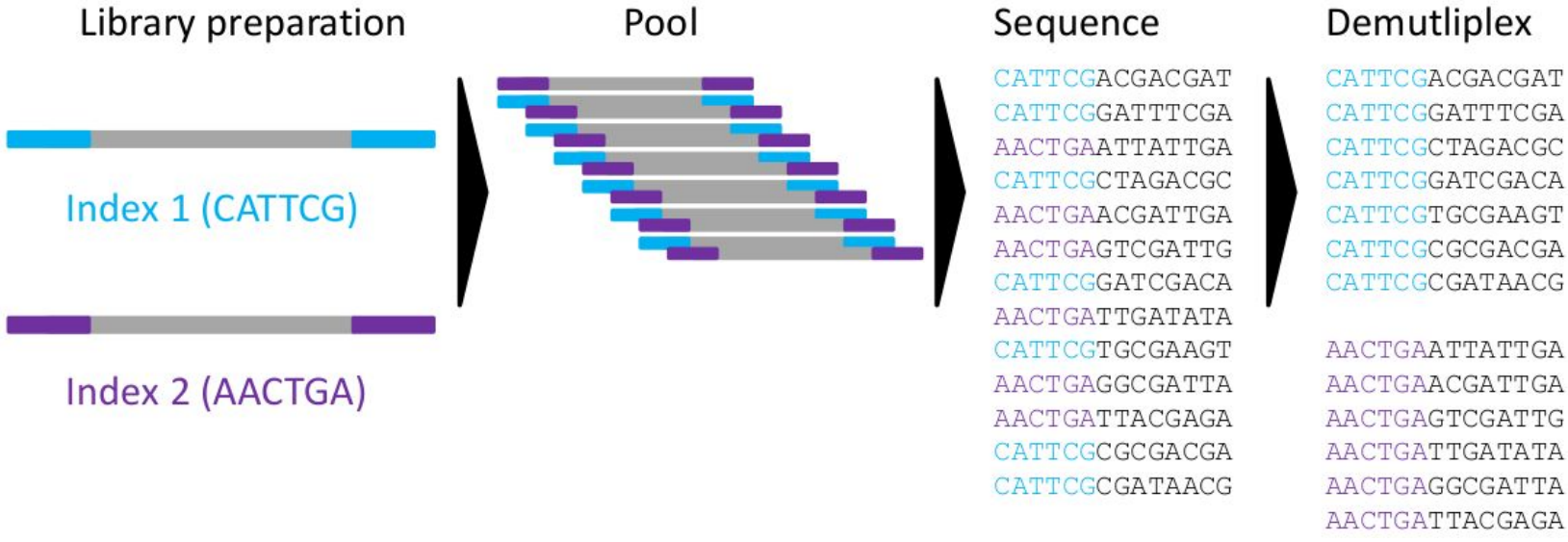


UNIVERSITÄT BONN

Illumina - Read ID nomenclature



Illumina - multiplexing



Illumina - sequencing modi

- Type:
  - SE = single end
  - PE = paired-end
  - MP = mate pair

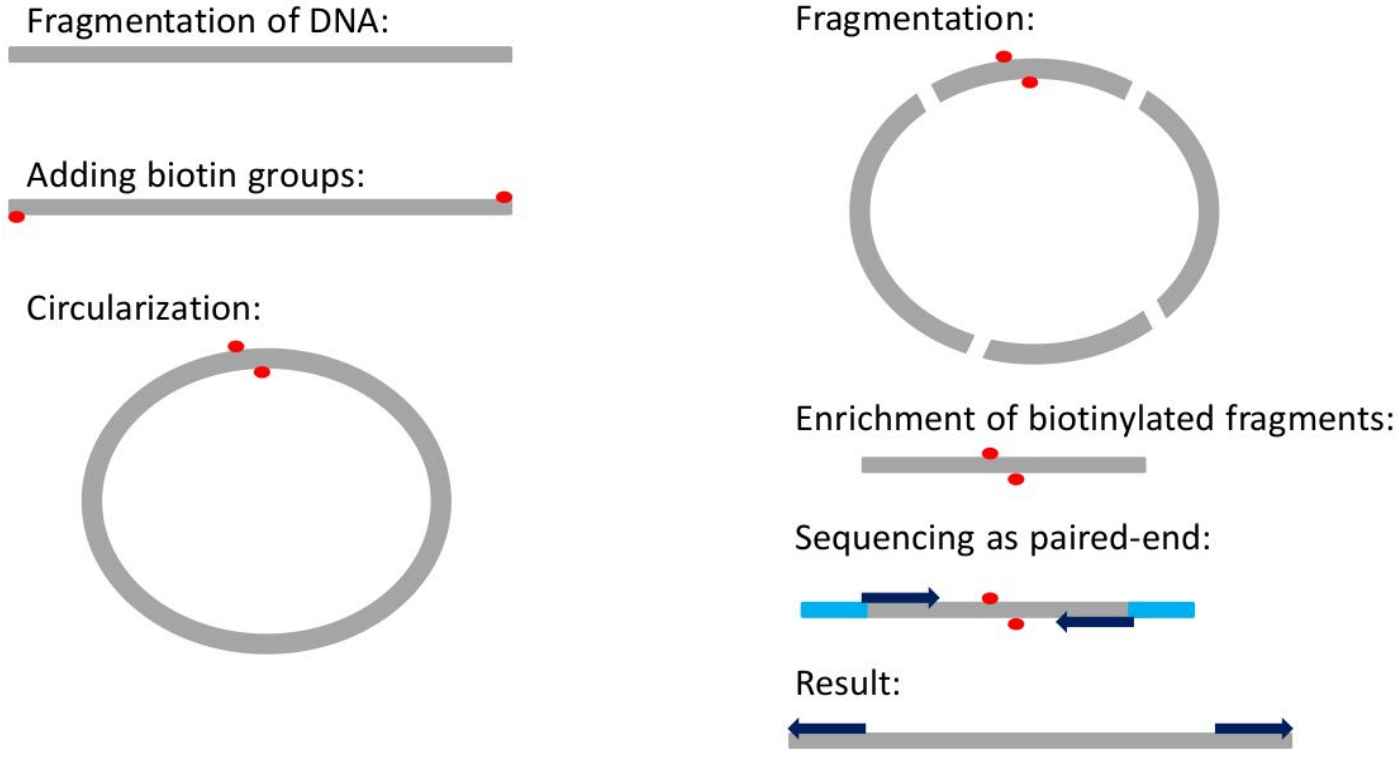
- Read length:
  - 32nt, 50nt, 75nt, 100nt, 150nt, 250nt, 300nt

- Examples:
  - 2x250nt PE, 2x100nt MP, 1x100nt SE

- Single end (SE):

- Paired-end (PE):

Illumina - sequencing modi (mate pair)



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:

  @seq1

  ACGTACGTACGT

  +

  “”?CB””:DC”

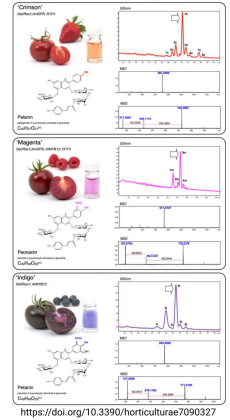
UNIVERSITÄT BONN

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)

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)



https://doi.org/10.3390/horticulturae7090327

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

Arabidopsis thaliana

Oryza sativa

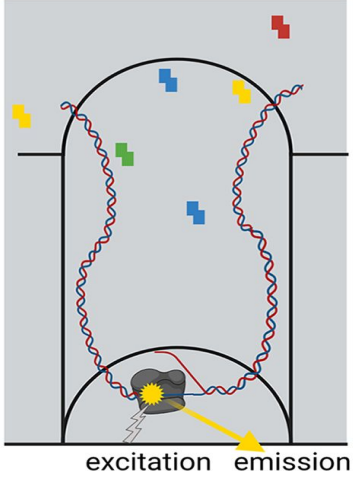
UNIVERSITÄT BONN

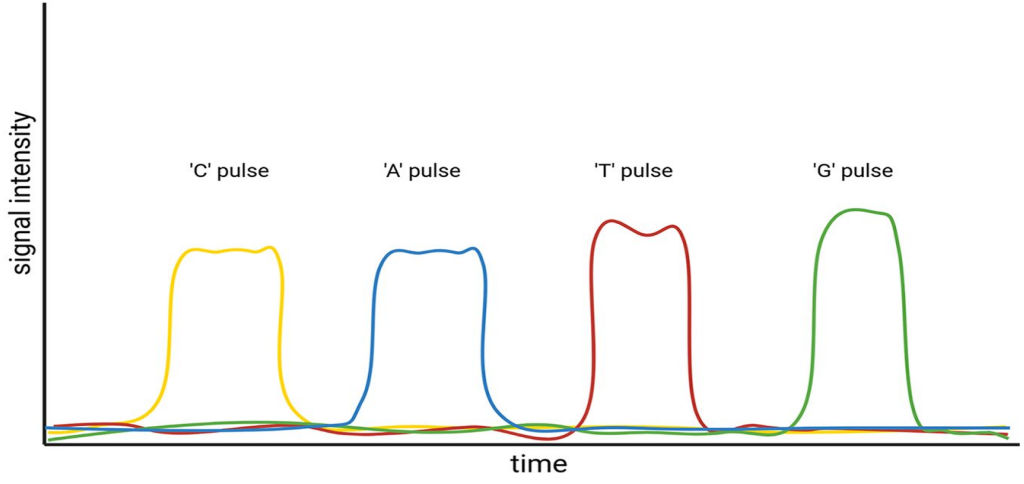
PacBio

Pacific Biosciences (PacBio)

- Polymerase located at bottom of well

- ZMWs = Zero Mode Waveguide

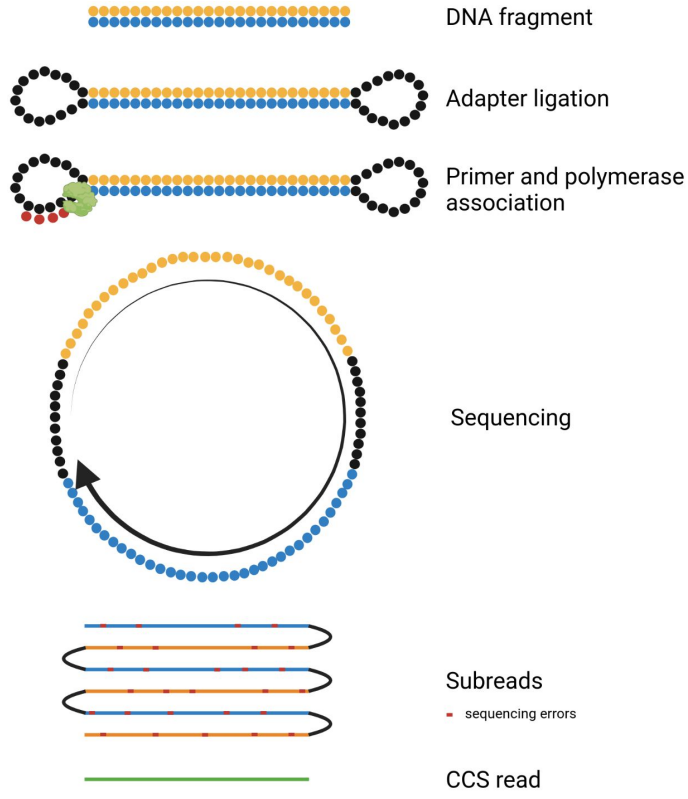




Pucker et al., 2022: 10.1017/qpb.2021.18

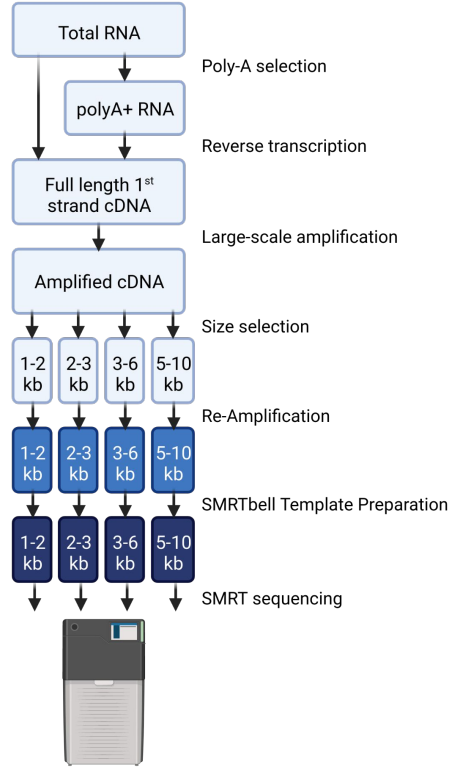
UNIVERSITÄT BONN

PacBio - HiFi



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



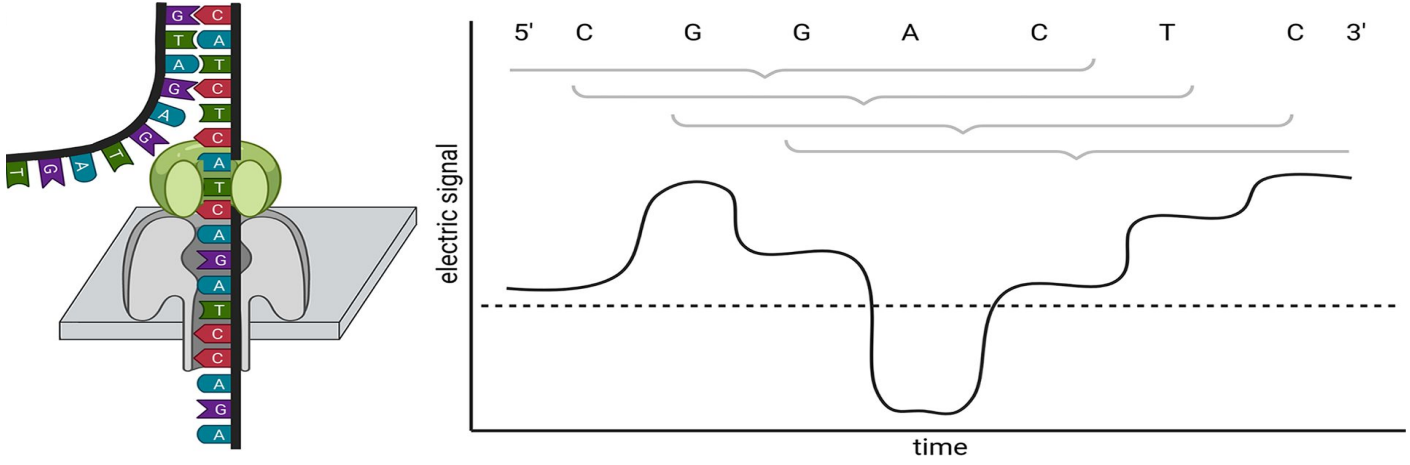
UNIVERSITÄT BONN

ONT

Oxford Nanopore Technologies (ONT)

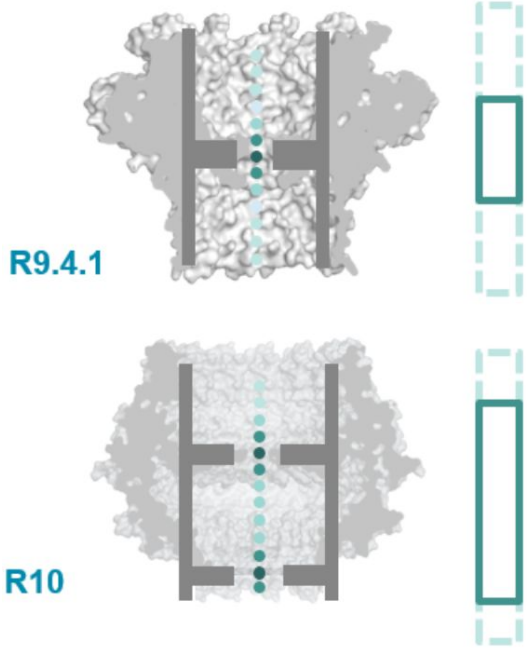
Under development since the 1980s

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



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

Nanopores



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

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

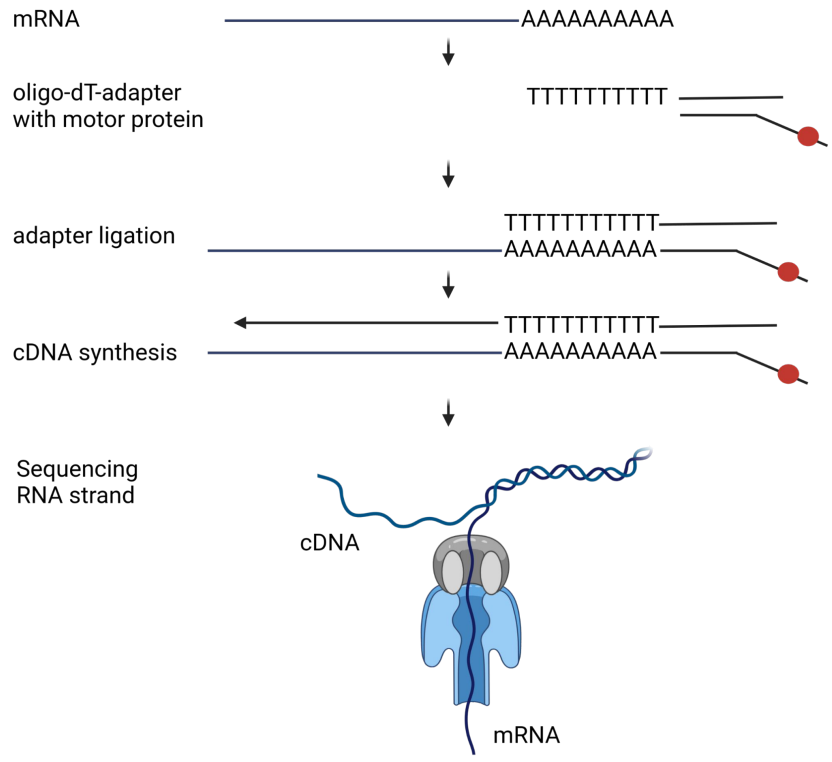
UNIVERSITÄT BONN

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 (usually >>1 Mbp)
- Centromeric and other large repetitive regions remain a challenge

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

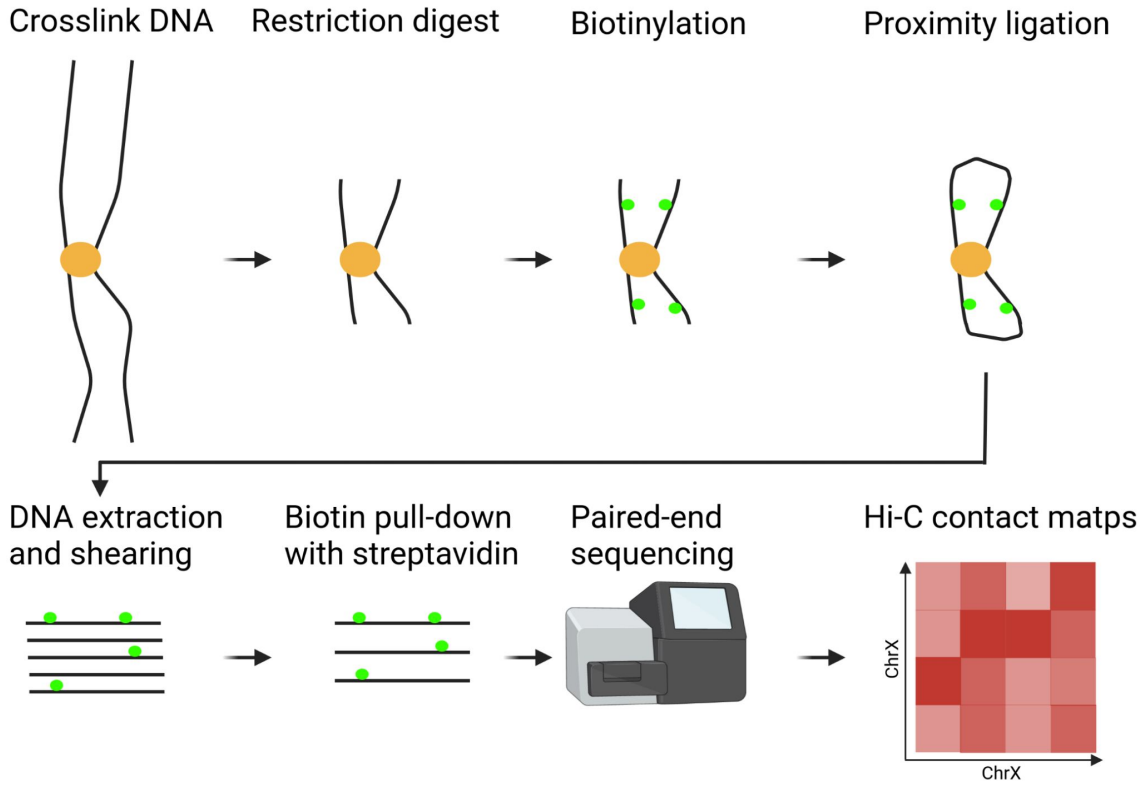


UNIVERSITÄT BONN

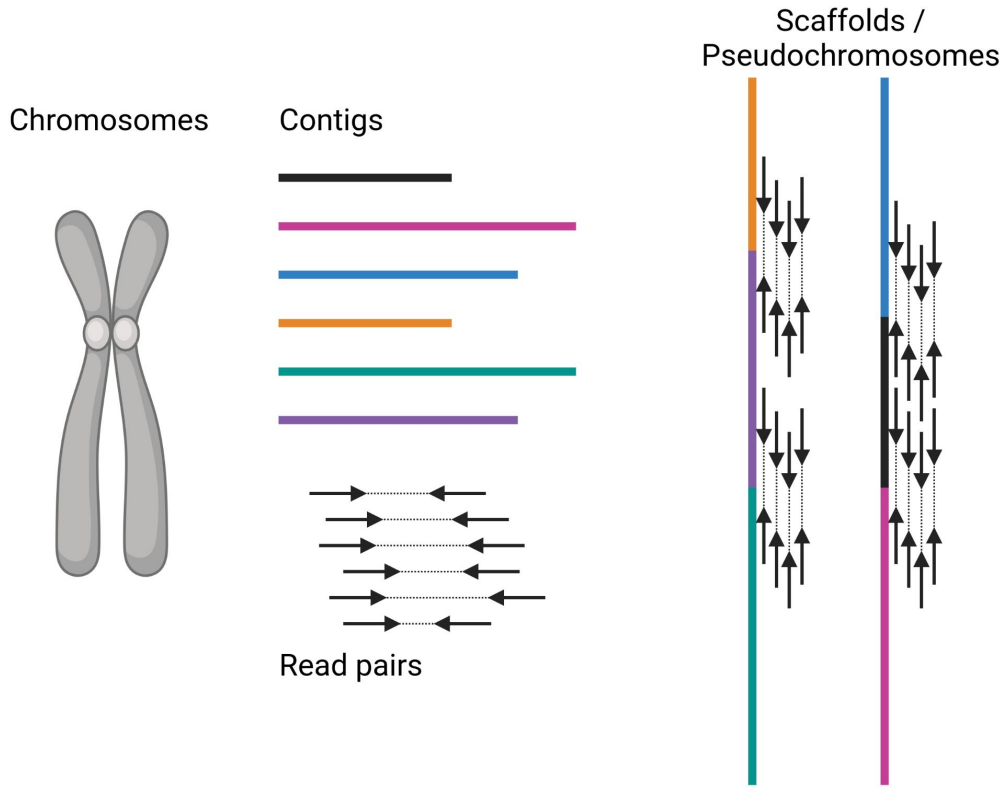
Trends in plant genomics

UNIVERSITÄT BONN

Hi-C (1)



Hi-C (2)

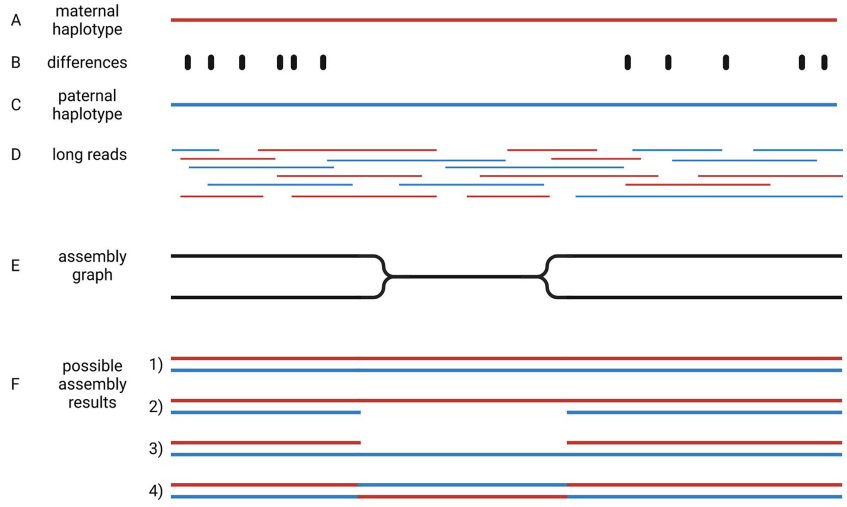

Chromosomes
Contigs
Read pairs
Scaffolds /
Pseudochromosomes

UNIVERSITÄT BONN

Haplophases

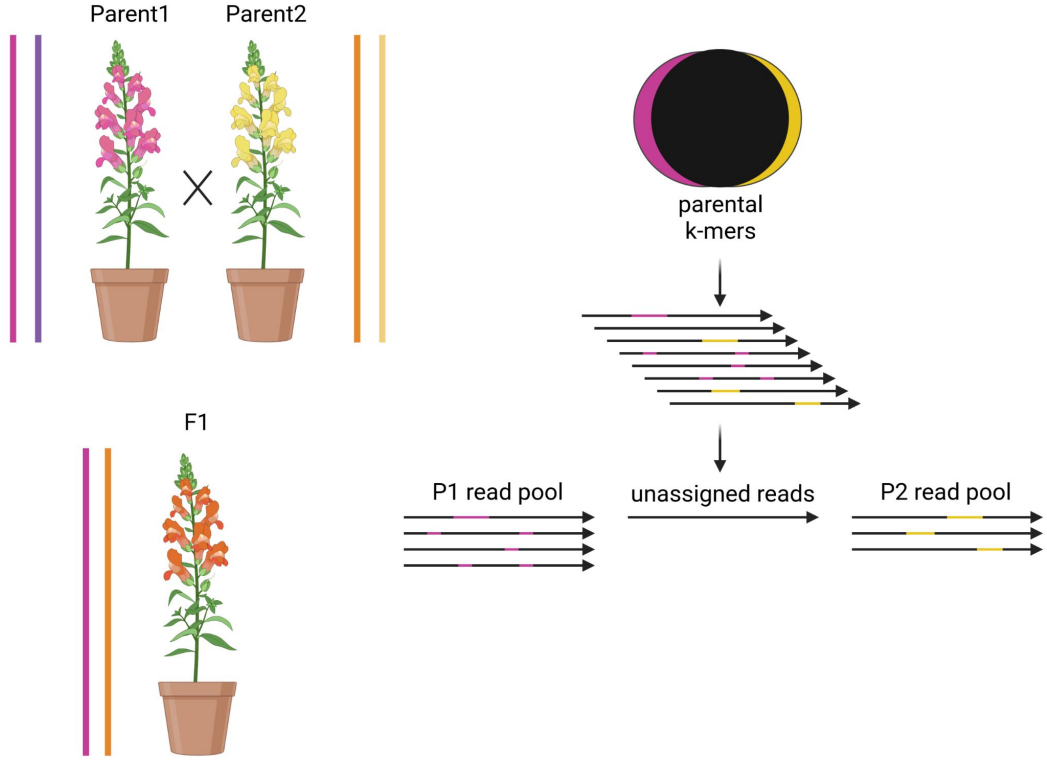
- Haplotype = combination of alleles

- Haplophase = representation of a haplotype



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

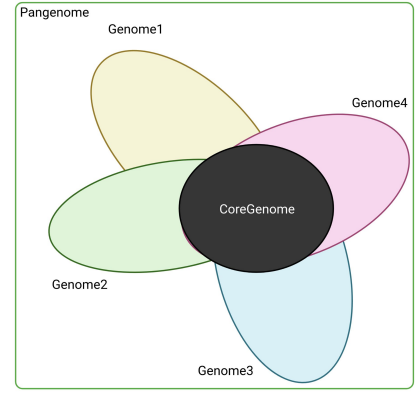
TrioBinning

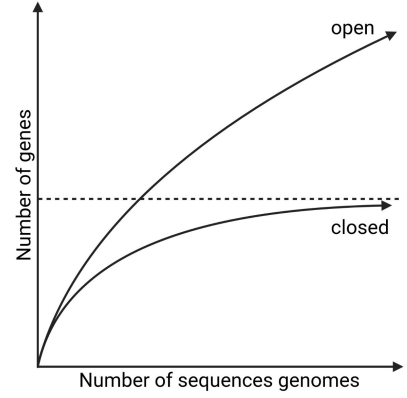


Koren et al., 2018: 10.1038/nbt.4277

Pangenomics

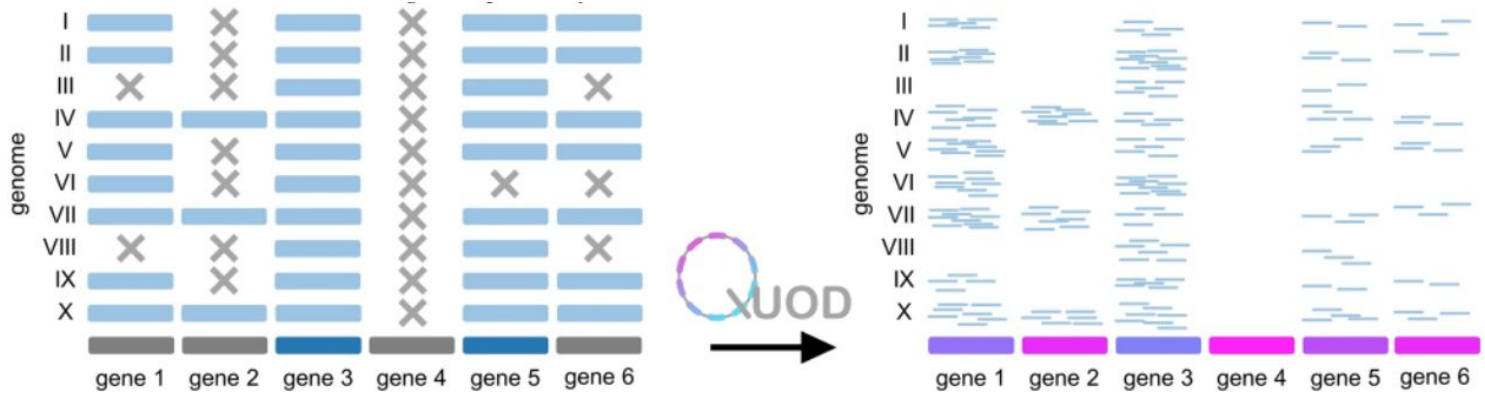
- Pangenome = combination of all genomes in a taxon
- Core genome = present in all species of a taxon
- Shell genome = present in many species of a taxon
- Cloud genome = only present in individual species





Pangenomics

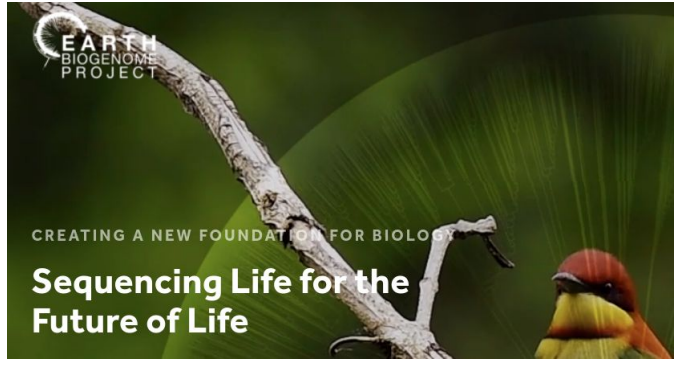
- Core genes = genes are present in all species of a taxon
- Conditionally dispensable genes = genes are present in many, but not in all species
- Dispensable genes = genes only present in a few species of a taxon



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/

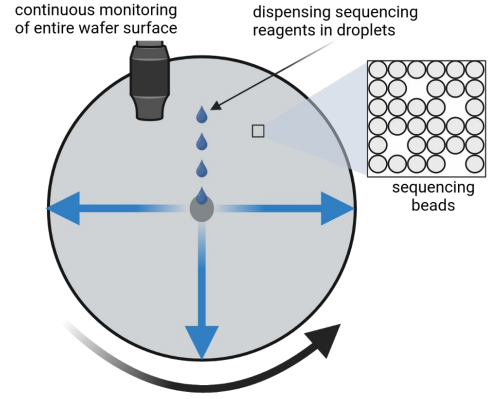
UNIVERSITÄT BONN

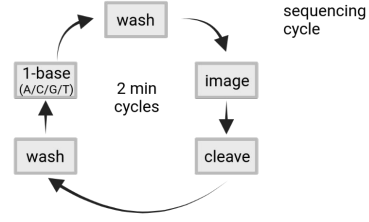
Democratization of genomics

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

New seq technology

- About 300 nt read length possible
- Base quality: Q15-Q40
- Homopolymers pose an issue (up to 4 possible)
- More natural than labeled nucleotides supplied per sequencing cycle





https://doi.org/10.1101/2022.05.29.493900

UNIVERSITÄT BONN

Roche SBX

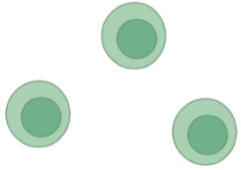
- SBX = Sequencing by expansion
- Xpandomer = large spacer inserted between bases (high signal to noise)
- X-NTPs are bases in Xpandomer (generated by replicating the DNA template strand)
- TCE = translocation control element -> controls nanopore sequencing
- Xpandomer is analyzed by nanopore in membrane
- measurement of one reporter at a time (high current pulse to move by one reporter)
- Details:

  https://sequencing.roche.com/global/en/article-listing/sequencing-platform-technologies.html

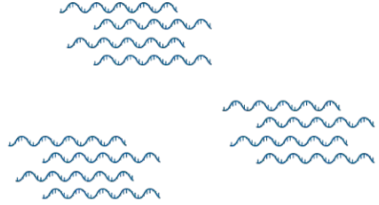
https://doi.org/10.1101/2025.02.19.639056

UNIVERSITÄT BONN

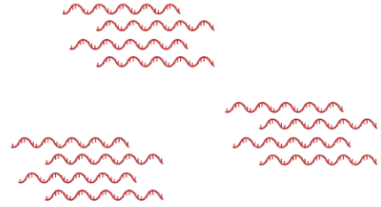
Single cell transcriptomics



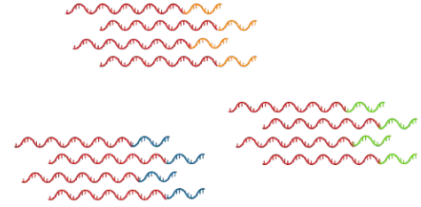
separating cells



RNA extraction

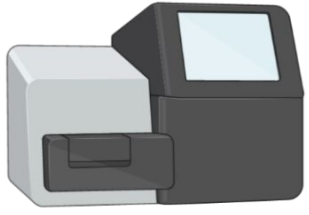


cDNA conversion

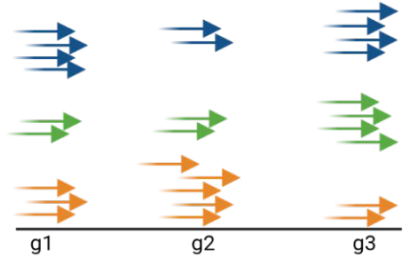


barcoding (UMI)

UMI = Unique Molecular Identifiers



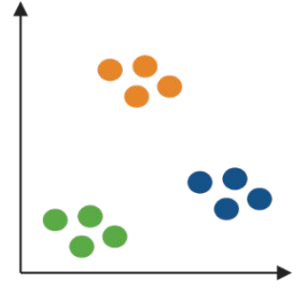
sequencing



read mapping

	s1	s2	s3
g1	0.6	0.3	0.6
g2	0.8	0.8	1.6
g3	0.8	1.2	0.6

normalization

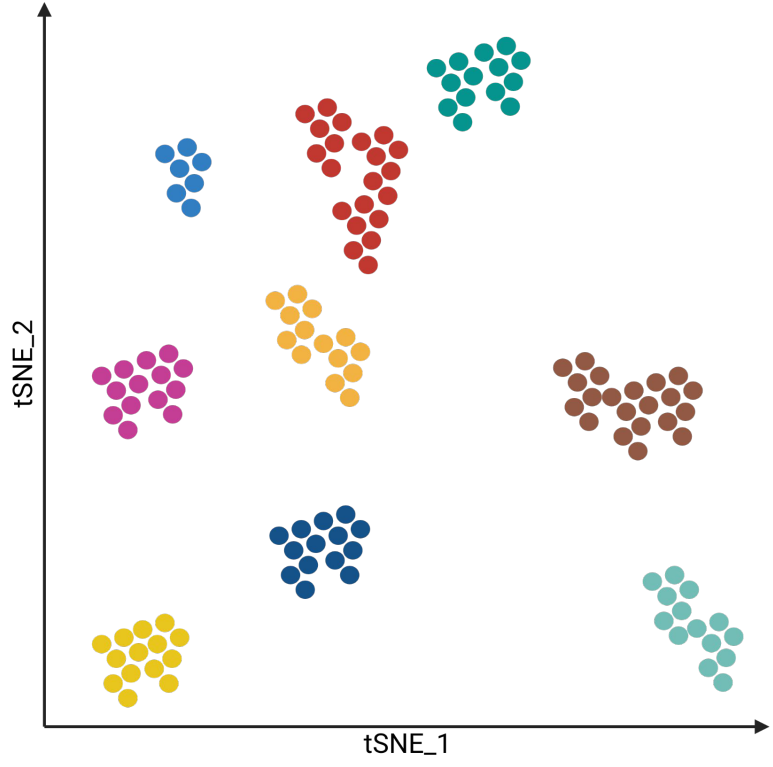


clustering

UNIVERSITÄT BONN

Molecular relationships among individual cells

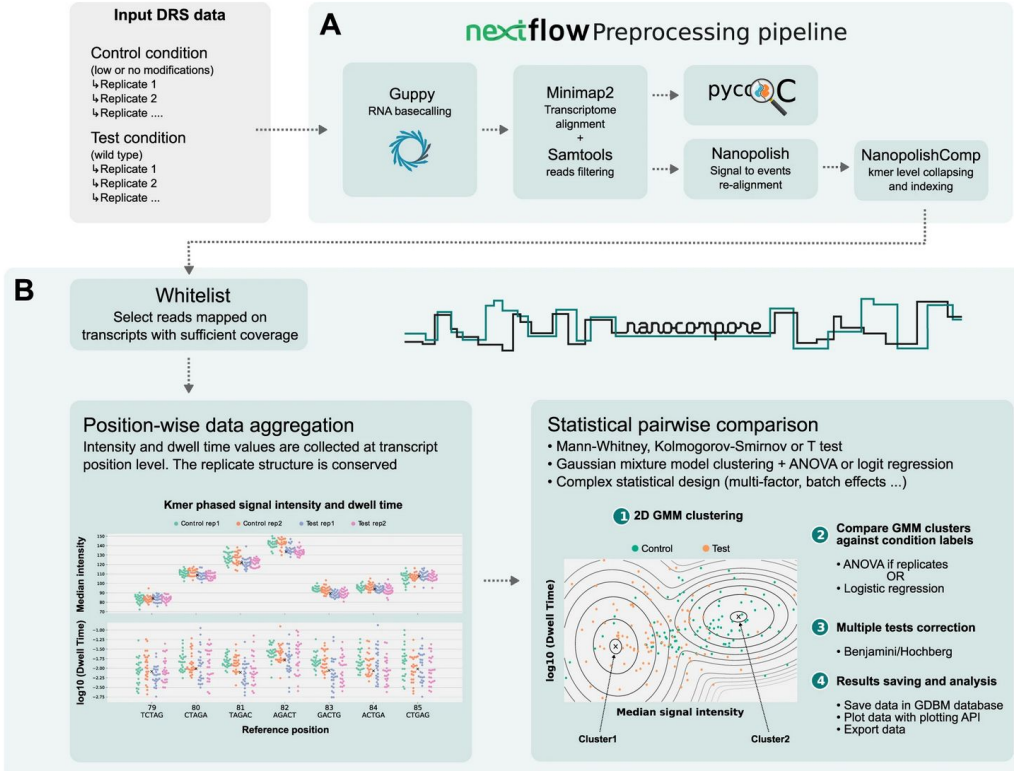
t-distributed stochastic neighbor embedding (tSNE)



Inspired by Ryu et al., 2019: 10.1104/pp.18.01482

UNIVERSITÄT BONN

RNA modification analysis



https://doi.org/10.1038/s41467-021-27393-3

UNIVERSITÄT BONN

Cloud computing

UNIVERSITÄT BONN

ELIXIR / de.NBI cloud

- ELIXIR: European organisation for bioinformatics (infrastructure)
- de.NBI: German Network for Bioinformatics Infrastructure
- de.NBI offers compute resources, training courses, and services

UNIVERSITÄT BONN

de.NBI cloud

- Virtual machine (VM) for data analysis
- Accounts are required for access (ORCID for login)
- Addition to project required for access

UNIVERSITÄT BONN

Requesting projects (de.NBI)

- What information is required?
  - Brief project description with resource justification
  - Information about involved researchers

- Which resources can be requested?
  - Large memory machines (4TB)
  - Machines with powerful GPUs (A40, A100)
  - Compute cluster connecting multiple nodes (slurm via BiBiGrid)

- Who can request resources?
  - Research groups at German universities

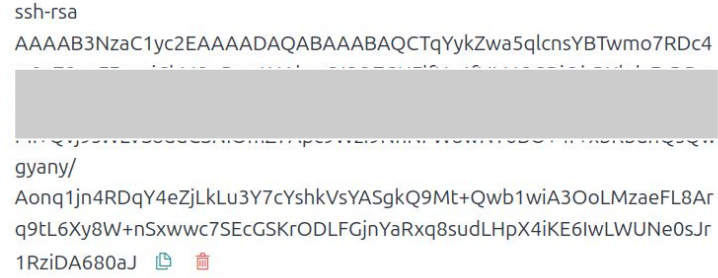
https://github.com/BiBiServ/bibigrid

Access with key pair

- Access to virtual machines (VMs) is controlled with key file

- User generates pair of private and public key
  - Private key = locally stored file
  - Public key = corresponding file stored in de.NBI portal and VMs

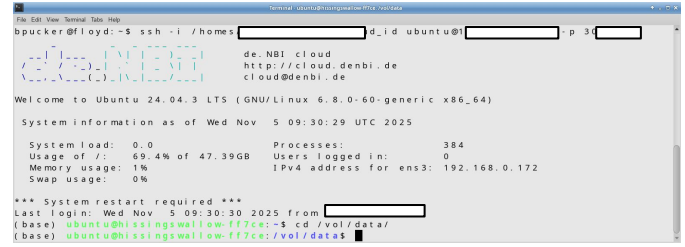
- Connection via ssh (secure shell)



UNIVERSITÄT BONN

Starting a simple VM

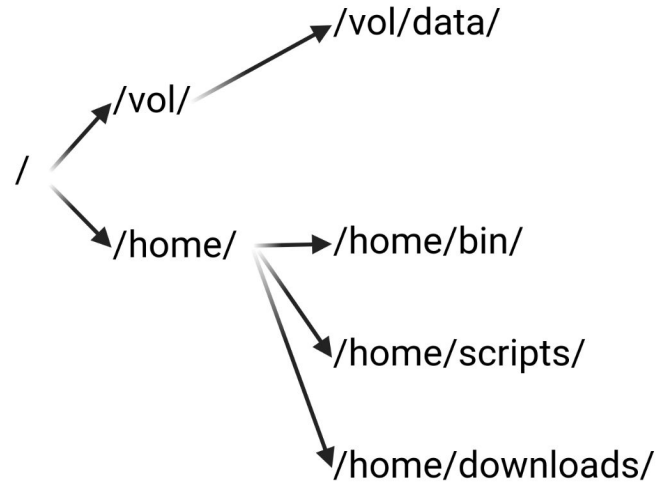
- VM is computer running in the cloud
- Operation system is Ubuntu (linux system)
- Users have admin privileges and ability to install software
- Excellent opportunity for students (easy to start and delete)
- VM is running constantly thus long running jobs can be computed there



UNIVERSITÄT BONN

File management on linux

- Linux (Ubuntu) is operation system of choice for bioinformatics
- Hierarchical structure of directory (‘/’ is basis)
- Separation of tool installation and data sets
- File naming conventions:
  - Never use spaces in file or directory names
  - Include dates in file names (year-month-day)
- Indicating commands with prefix ‘$’



UNIVERSITÄT BONN

Permissions

- Files can have different permissions:
  - Read (r)
  - Write (w)
  - Execute (x)

- Users have full permissions to edit their own files

- Downloaded files are usually not executable without adjustment

- chmod XXX <FILE_NAME> … can be run to change file permissions

UNIVERSITÄT BONN

Transferring files

- Filezilla: graphical user interface for file transfer protocols
  - https://filezilla-project.org/
- Scp (secure copy): command line file transfer method
- Wget: command line file transfer method
  - https://ftp.gnu.org/old-gnu/Manuals/wget-1.8.1/html_mono/wget.html
- Rsync: sophisticated file transfer method that avoids redundant transfers
  - https://wiki.ubuntuusers.de/rsync/

UNIVERSITÄT BONN

Transferring files: scp / rsyn

```
scp -P 12345 -i /path/to/my/private_key -r \
/my/local/path/file.txt \
ubuntu@123.45.67.89:/vol/data/file.tx

rsync --append-verify -avzhP \
-e "ssh -i /path/to/my/private_key -p 12345" \
/my/local/path/victoria.overlaps.paf \
ubuntu@123.45.67.89:/vol/data/victoria.overlaps.paf
```

UNIVERSITÄT BONN

Running jobs through CLI

- CLI = command line interface
- All commands need to be entered into a terminal
- Recommendation: prepare your commands in a README (serves as documentation)
- Copy & paste on Ubuntu: mark text and click with middle mouse button to paste

```
/vol/data/dorado-0.8.3-linux-x64/bin/dorado \
basecaller sup \
/vol/data/R160_0cb0204a/ \
--modified-bases 5mCG_5hmCG \
> /vol/data/R160_0cb0204a.mod.bam \
2> /vol/data/R160_0cb0204a.dorado.mod.err.txt && \
/vol/data/dorado-0.8.3-linux-x64/bin/dorado \
basecaller sup \
/vol/data/R159_a5ce4e6e/ \
--modified-bases 5mCG_5hmCG \
> /vol/data/R159_a5ce4e6e.mod.bam \
2> /vol/data/R159_a5ce4e6e.dorado.mod.err.txt && \
/vol/data/dorado-0.8.3-linux-x64/bin/dorado \
basecaller sup \
/vol/data/R158_5dda675f/ \
--modified-bases 5mCG_5hmCG \
> /vol/data/R158_5dda675f.mod.bam \
2> /vol/data/R158_5dda675f.dorado.mod.err.txt && \
/vol/data/dorado-0.8.3-linux-x64/bin/dorado \
basecaller sup \
/vol/data/R157_747b0957/ \
--modified-bases 5mCG_5hmCG \
> /vol/data/R157_747b0957.mod.bam \
2> /vol/data/R157_747b0957.dorado.mod.err.txt && \
/vol/data/dorado-0.8.3-linux-x64/bin/dorado \
basecaller sup \
/vol/data/R156_160e7a35/ \
--modified-bases 5mCG_5hmCG \
> /vol/data/R156_160e7a35.mod.bam \
2> /vol/data/R156_160e7a35.dorado.mod.err.txt && \
/vol/data/dorado-0.8.3-linux-x64/bin/dorado \
basecaller sup \
/vol/data/R155_a339f972/ \
--modified-bases 5mCG_5hmCG \
> /vol/data/R155_a339f972.mod.bam \
2> /vol/data/R155_a339f972.dorado.mod.err.txt &
```

UNIVERSITÄT BONN

Running Python scripts

- Run script or open argument infos:
  - $ python <SCRIPT_NAME>

- Run script with arguments:
  - `$ python <SCRIPT_NAME> --argument1_name <ARGUMENT1> --argument2_name— <ARGUMENT2>`

- Scripts show help message if started with insufficient arguments

```
python3 ./KIPEs3.py --baits ./flavonoid_baits/ --positions ./flavonoid_residues/ --out ./ --subject ./croton_red.fasta --seqtype pep --scoreratio 0.3 --simcut 40.0 --minsim 0.4 --minres 0.0 --minreg 0.0 --possibilities 3 --cpus 1
```

UNIVERSITÄT BONN

Running other tools

- Show help message:
  - $ <NAME_OF_TOOL>
  - $ <NAME_OF_TOOL> -h
  - $ <NAME_OF_TOOL> -- help

- Providing arguments is different for each tool:

- Most tools show help message if provided with insufficient/wrong arguments

UNIVERSITÄT BONN

Job management

- Submitted jobs are killed when the terminal is closed
- Sending jobs “to the back” will allow you to disconnect from VM
- Running jobs in the back by appending “&” to command
- tmux: more sophisticated job management (https://github.com/tmux)

UNIVERSITÄT BONN

How to access files

- head: show only first lines of a file
- tail: show only last lines of a file
- cat: show the entire file content
- nano: text editor that allows you to modify text files in VM

UNIVERSITÄT BONN

How to redirect output

- Communication with job through STDOUT and STDERR
- STDERR: standard error; often used for progress updates in bioinformatics
- STDOUT: standard output; often used for result output in bioinformatics

UNIVERSITÄT BONN

Chaining commands

- Commands can be connected
- Output of one tool can serve as input for next tool
- Pipe (“|”) is used to direct input into command
- Combine multiple tasks with “&&”; only continue if previous job successful

UNIVERSITÄT BONN

de.NBI training course

Plant Genome Sequence Assembly and Annotation

- Bonn

Educators:
Prof. Dr. Boas Pucker, Katharina Wolff, Nancy Choudhary, Shakunthala Natarajan, Samuel Nestor Meckoni, Julie A. V. S. De Oliveira (Associated Partner)

Date:
08.12.2025-10.12.2025

Location:
Kirschallee 1, 53115 Bonn, Germany

Contents:
Unlock the secrets of plant genomes with this intensive, hands-on course in "Plant Genome Sequence Assembly and Annotation". Designed for doctoral students and postdoctoral researcher in genomics and bioinformatics, this course offers practical training in processing and assembling Oxford Nanopore Technologies (ONT) long-read sequencing data. Participants will learn to perform quality control, assemble high-quality plant genome sequences using tools like HERRO, Shasta, and NextDenovo2, and carry out structural genome sequence annotation with HISAT2/STAR, GeMoMa, and BUSCO. The course emphasizes real-world applications, providing participants with the skills to interpret, validate, and refine genome assemblies and gene models specific to plant systems. Whether you are starting your first genome project or looking to enhance your bioinformatics toolkit, this course will give you the expertise to tackle complex plant genomic data with confidence.

Learning goals:
How to get from ONT long reads to a genome sequence with corresponding gene models

Prerequisites:
Basic experience with command line

Keywords:
genome sequence, assembly, long reads, gene models, plant genomics

Tools:
HERRO, Shasta, NextDenovo2, HISAT2/STAR, BUSCO, GeMoMa

Contact:
Prof. Dr. Boas Pucker, pucker@uni-bonn.de

Training Courses 2025
Training Archive by Date
Training Archive by de.NBI Units
Online Training & Media Library
TeSS

The de.NBI Training platform

To effectively coordinate training courses, de.NBI has established the Working Group "Training and Education". The WG is composed of training experts from each de.NBI unit.

If you need more information on a course, would like to suggest course topics that de.NBI might want to cover in the future, or would like to provide feedback about our training program, please contact us!

Write email

UNIVERSITÄT BONN

Summary

- Sequencing technologies:
  - Sanger
  - Illumina
  - PacBio
  - ONT
- Trends in plant genomics
- Cloud computing

UNIVERSITÄT BONN

Time for questions!

UNIVERSITÄT BONN

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.