

Prof. Dr. Boas Pucker

Long Read Genomics

- Plant Genome Sequencing with ONT

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](mailto:pucker[a]uni-bonn.de)

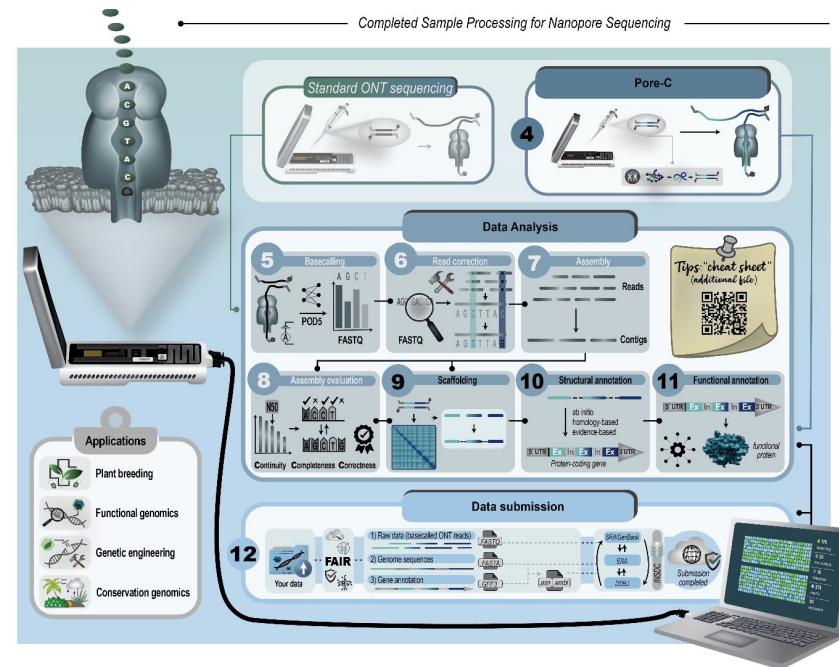
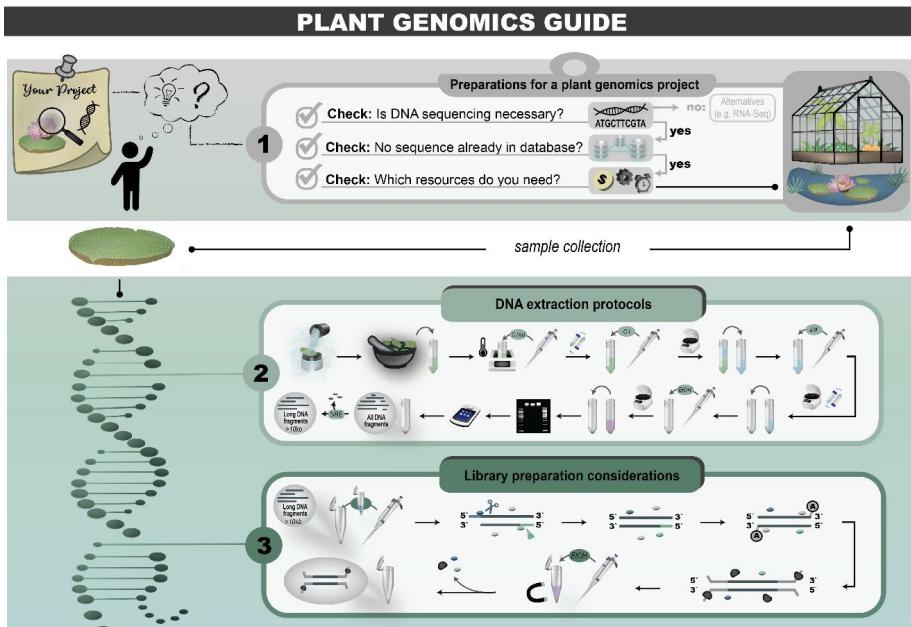


My figures and content can be re-used in accordance with CC BY 4.0, but this might not apply to all images/logos.

Repetition

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

Plant Genomics Workflow



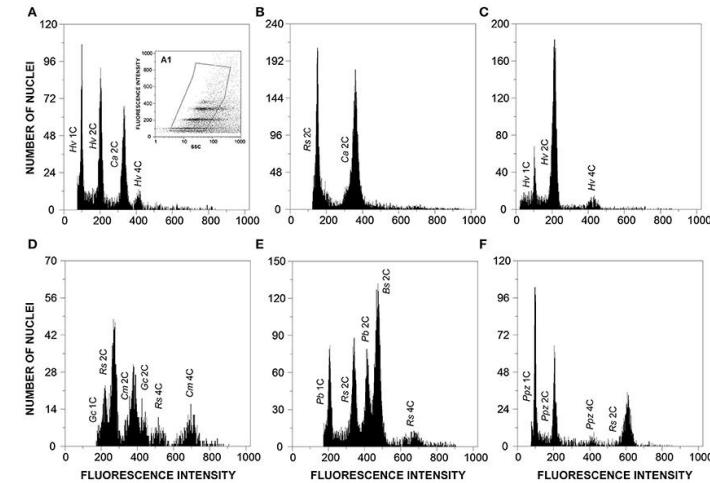
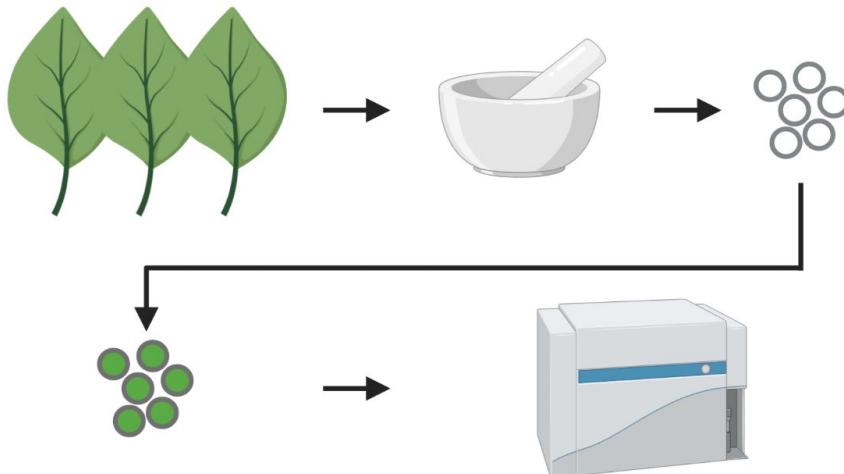
Supplementary file with all commands!

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 - 1

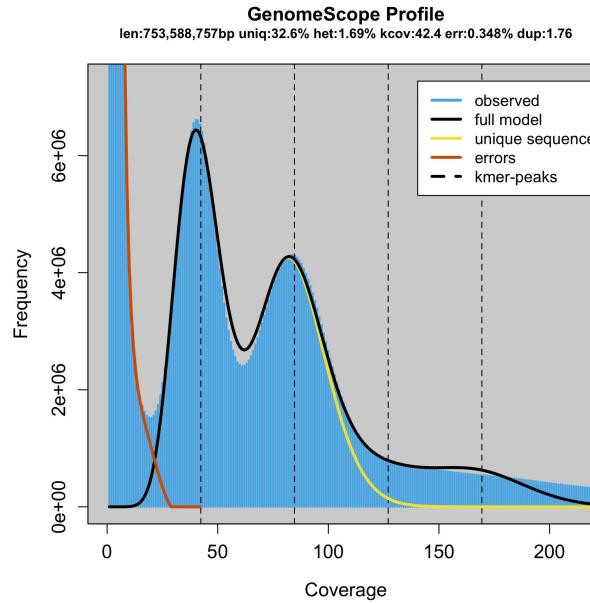
- Genome sequences of closely related species
- Flow cytometry is used to measure genome size biochemically
 - C-value
- Databases for plant genome sizes: <https://cvalues.science.kew.org/>



Genome size - 2

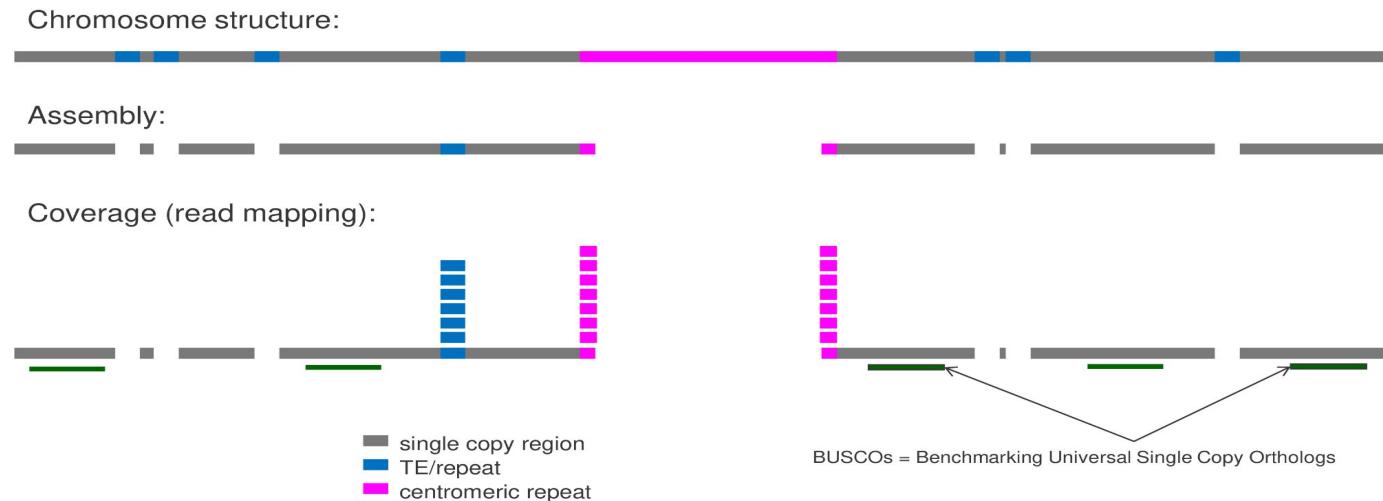
- Tools for genome size estimation based on short reads
 - K-mer-based: GenomeScope2, findGSE, gce

ACGAAGCCATAT
ACGAAGC
CGAAGCC
GAAGCCA
AAGCCAT
AGCCATAT



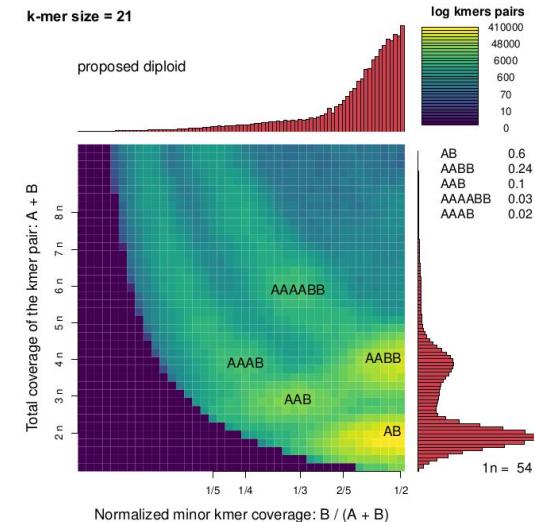
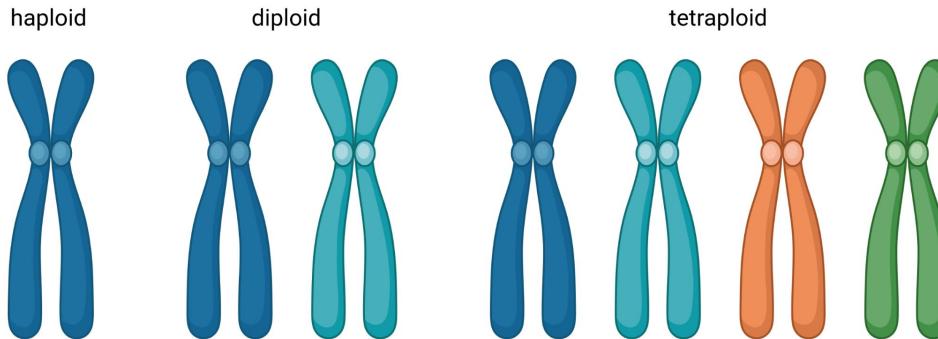
Genome size - 3

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

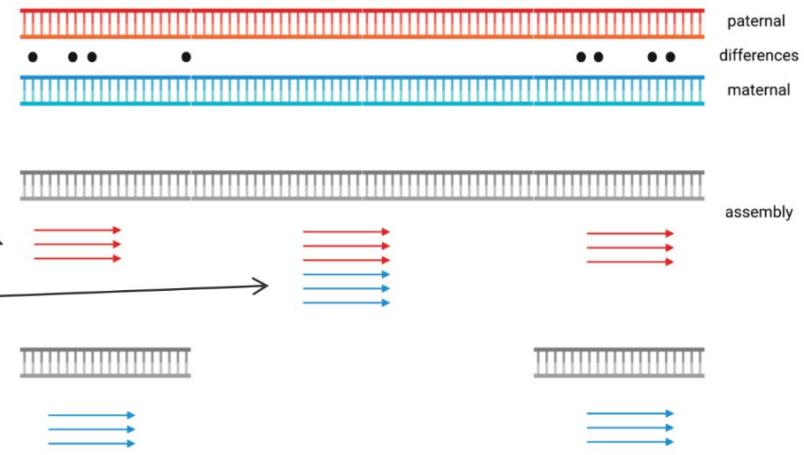
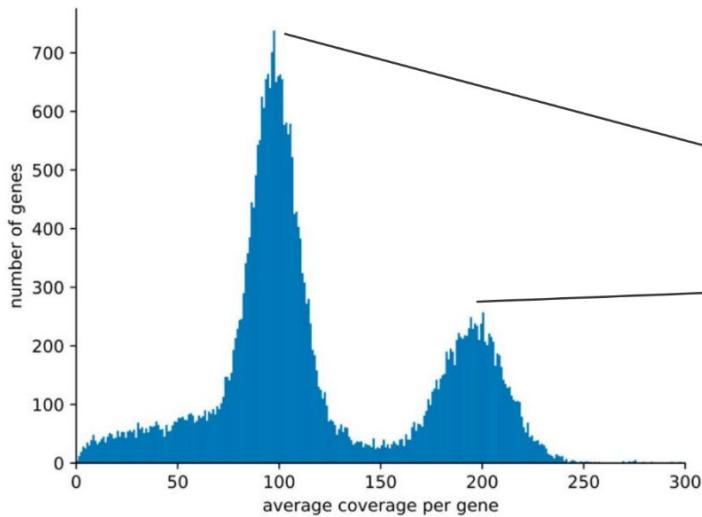


Ploidy - 1

- Ploidy = copy number of same chromosome
- Many plants are polyploid
- Some polyploid plants have close diploid relatives
- Smudgeplots to analyze polyploidy based on short reads



Ploidy - 2

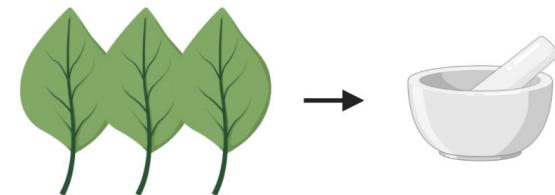


Picking an individual

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

Plant part for DNA extraction

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

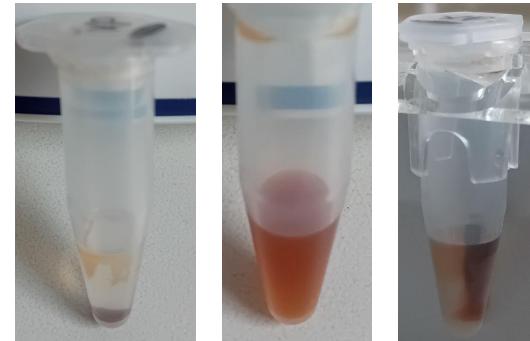


Material for DNA extraction

- Cetyltrimethylammonium-bromide (CTAB) buffer
- beta-mercaptoethanol (β -ME)
- Dichlormethane (replaces chloroform)
- TrisEDTA (TE) + RNase A
- OPTIONAL: Short Read Eliminator (SRE) kit

Material required for sequencing

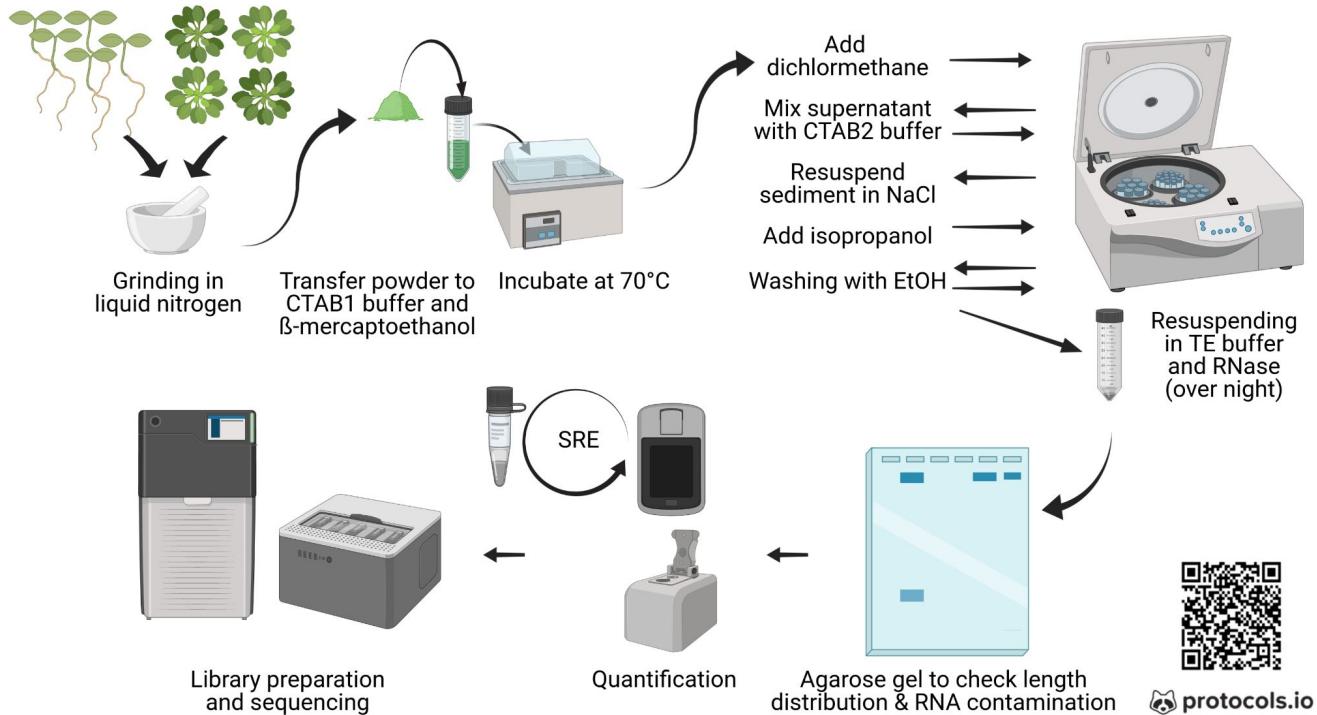
- DNA repair (companion module)
- Magnetic beads for purification
- Library preparation kit
- Sequencing kit
- Wash kit with DNase



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 \$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		FASTQ/ FASTA ↓ data submission	2h	2h	fast internet connection	

DNA extraction workflow



ONT sequencing devices

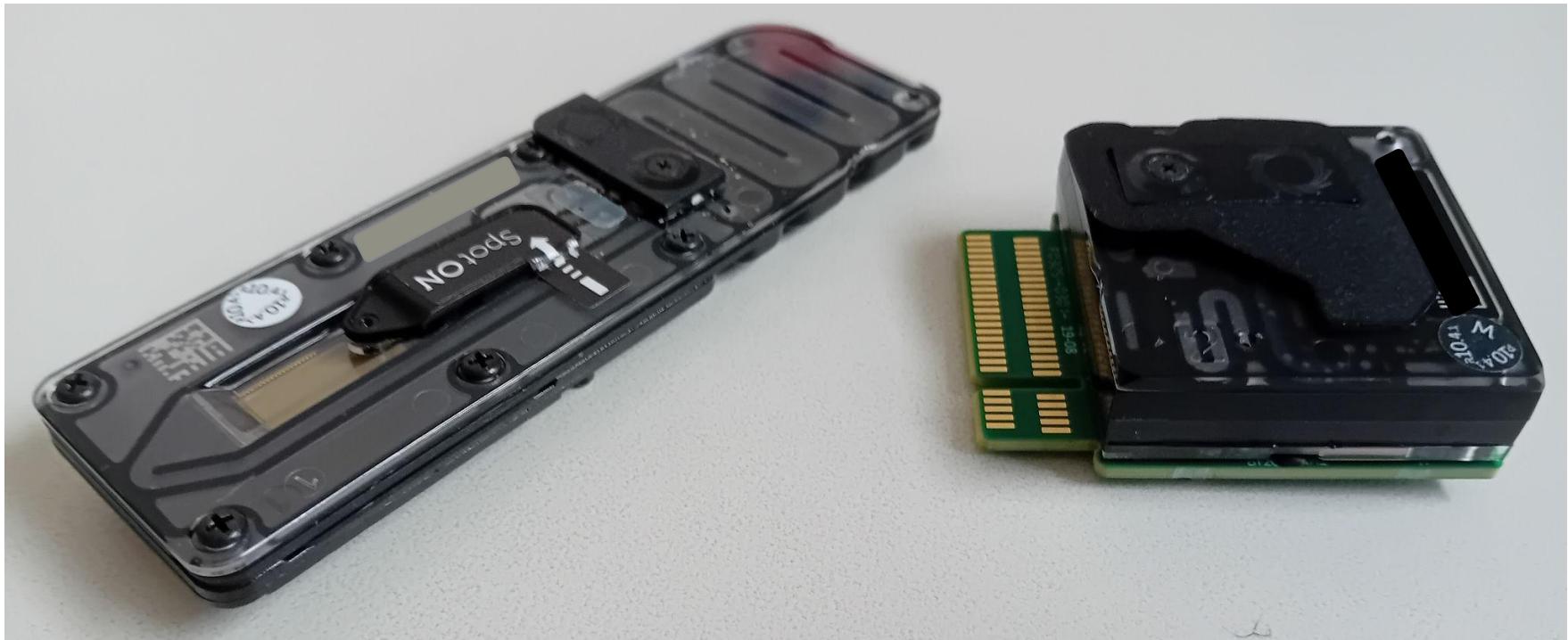


MinION

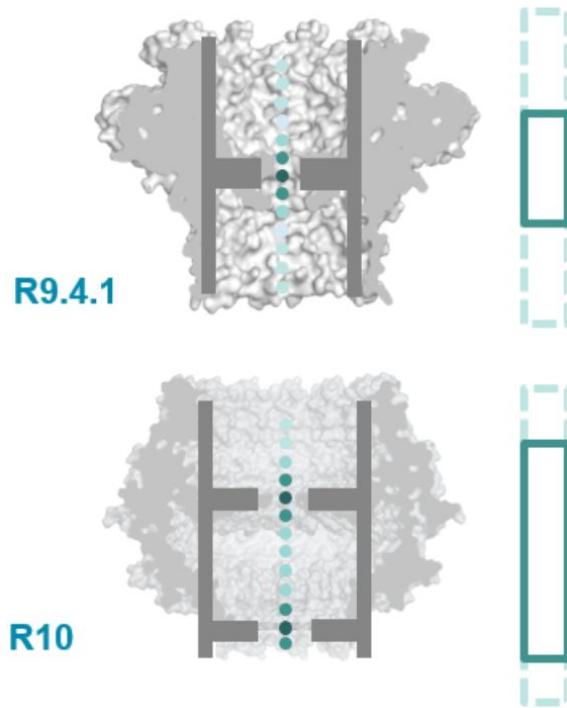


PromethION 2 Solo

Flow cells: MinION vs. PromethION

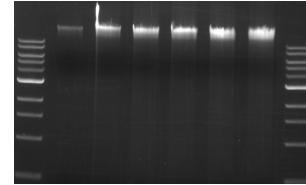


Nanopore comparison



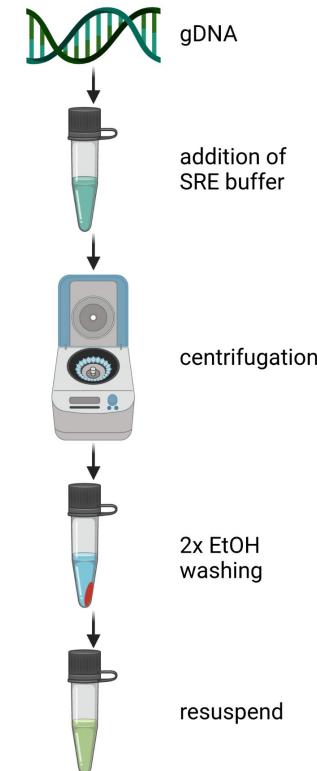
Quality control

- Agarose gel electrophoresis
- Photometric measurement via NanoDrop
- Quantification with Qubit



Short Read Eliminator (SRE)

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



DNA repair

- Repairing single strand DNA breaks
- Repairing DNA ends (3'-A overhang required for adapter ligation)



Library preparations

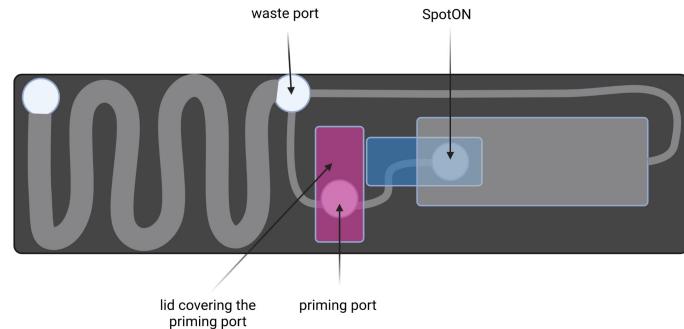
- Repaired DNA is subjected to library preparation
- Addition of adapters to DNA fragments (expensive ligase)
- Concentration of DNA can be quantified via Qubit measurement (optional)
 - Control step to ensure that library construction is working
- Purification of DNA with magnetic beads

Flow cell check

- Flow cells are delivered with storage buffer (green)
- Buffer allows technical check of flow cell (number of active pores is determined)
- Number of nanopores must be >800 (MinION) or >5000 (PromethION)
- Replacement flow cells are provided if number of pores is lower

Loading flow cell

- Removal of storage buffer
- Priming of flow cell
- Introduction of air bubbles must be avoided!!!
- Fully open ports are crucial to inject solutions (avoid force)
- Video tutorial: <https://www.youtube.com/watch?v=Pt-iaemrM88>

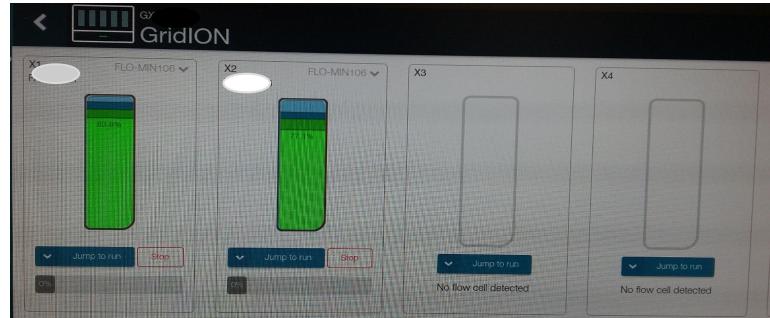


Starting sequencing

- Define the flow cell type (R10.4.1)
- Select parameters:
 - Make sure to keep POD5 files (default = off)
 - Define run time (default = 72h)
- Set the output location, run name, sample ID, ...
- Start the sequencing some minutes after loading the flow cell
 - This allows the DNA to get into contact with the nanopores

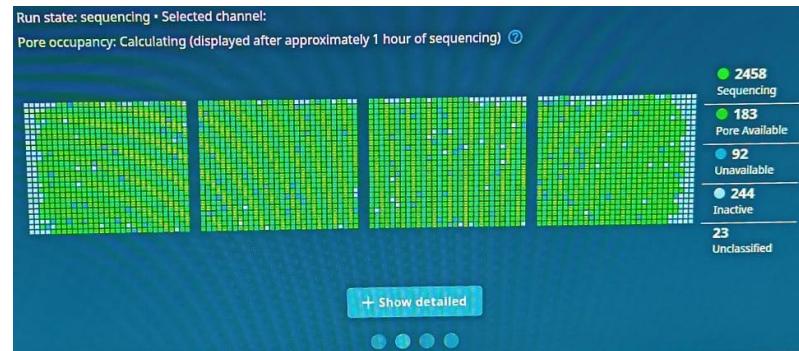
Monitor sequencing

- Number of active nanopores can be monitored in real time
- Output is estimated in real time
- Read length distribution can be assessed
- Speed of the sequencing can be monitored
- Quality of the reads is displayed



Nanopore activity

- Status of nanopores on flow cell
- Light green is good (sequencing)
- Number of sequencing nanopores change over time
- Air bubbles can destroy nanopores



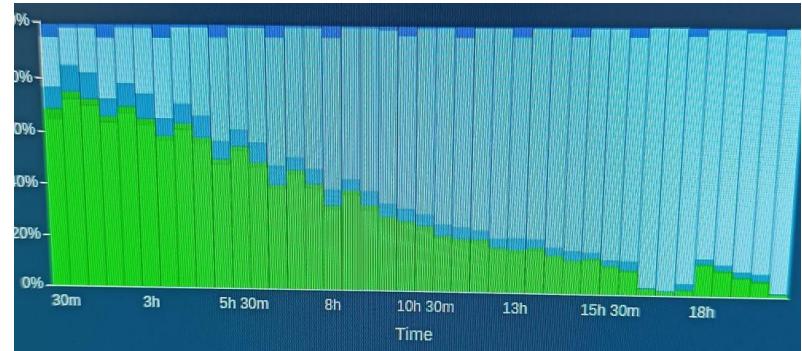
Read length distribution

- Read length distribution
- N50 summarizes length distribution



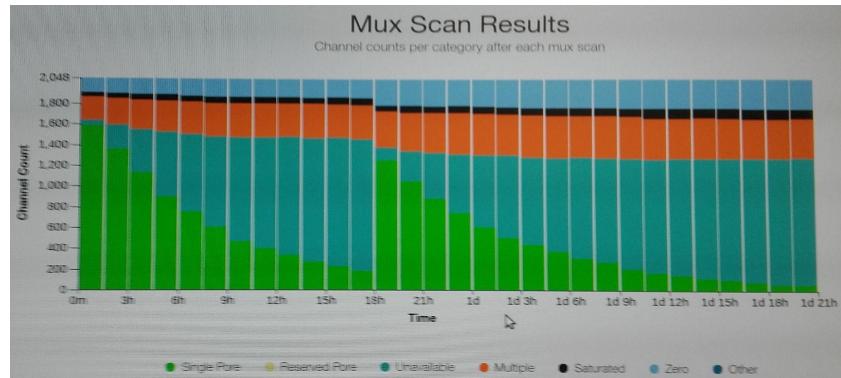
Flow cell performance over time

- Flow cell performance decreases over time
- Nanopores get blocked by DNA
- Nanopores break down due to air bubbles



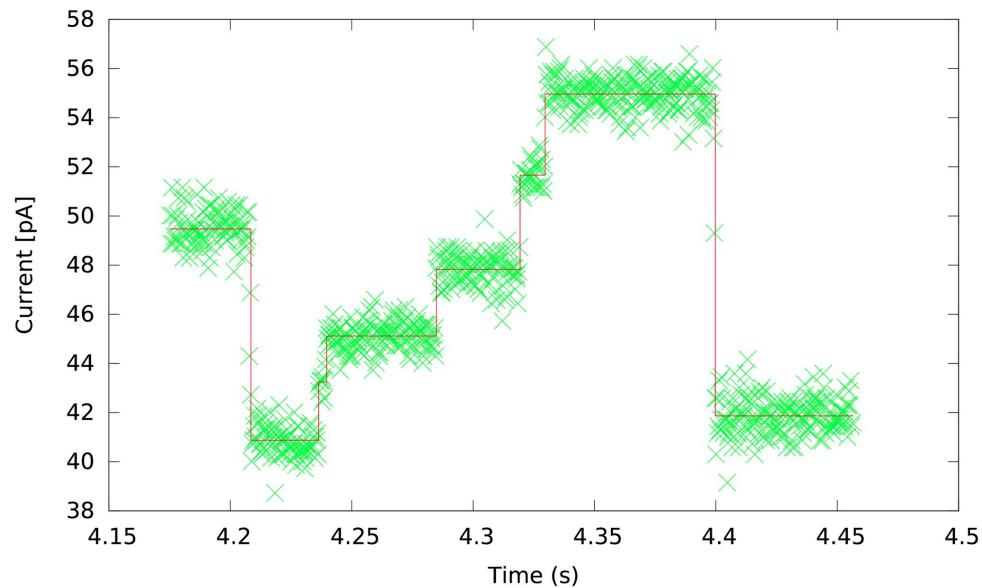
Stop & wash

- 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 (basecalling)
- Algorithmic improvement lead to higher read accuracy
- Raw sequencing data need to be stored

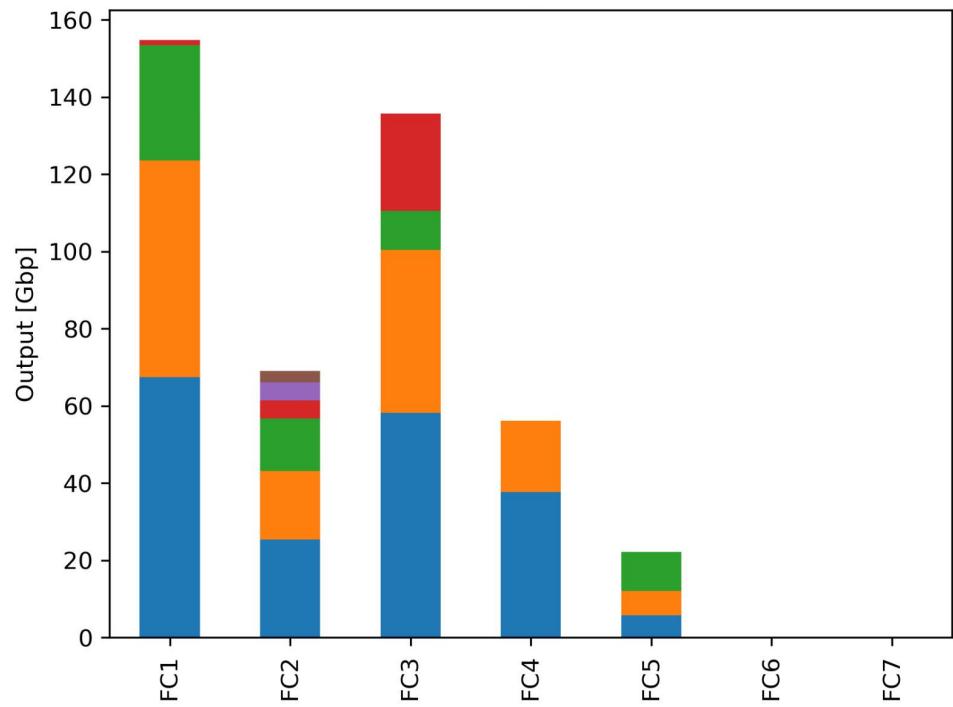


Basecalling with dorado

- Dorado allows conversion of electric signal into sequences
- Different models are available (HAC, SUP)
- Basecalling on CPUs is basically impossible
- Time estimates on GPUs: hours to days depending on dataset size

Calculating statistics

- Huge variation in output per flow cell
- Best runs: 150 Gbp
- Worst runs: 20 Gbp
- N50 for plants usually 25-65kb



Bioinformatics / Data analysis (overview)

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/>

- Virtual machine (VM) for data analysis
- Accounts are required for access (ORCID for login)
- Addition to project required for access
- User create pair of private and public keys for authentication



Compute Power for Your Project

In life sciences today, the handling, analysis and storage of enormous amounts of data is a challenging issue. For example, new sequencing and imaging technologies result in the generation of large scale genomic and image data. Hence, an appropriate IT infrastructure is crucial to perform analyses with such large datasets and to ensure secure data access and storage. In addition, it is difficult to directly compare result data that have been processed at different sites, due to a lack in standardization of workflows. The de.NBI cloud is an excellent solution to enable integrative analyses for the entire life sciences community in Germany and the efficient use of data in research and application.

To a large extent, de.NBI will close the gap of the missing computational resources for researchers in Germany. A federated de.NBI Cloud concept and infrastructure leads to the reduction in overall infrastructure and operational costs.

[CLICK HERE TO ENTER THE de.NBI CLOUD PORTAL](#)

Cloud Access Portal

de.NBI Cloud Flyer

de.NBI Cloud Poster

Cloud Training Courses

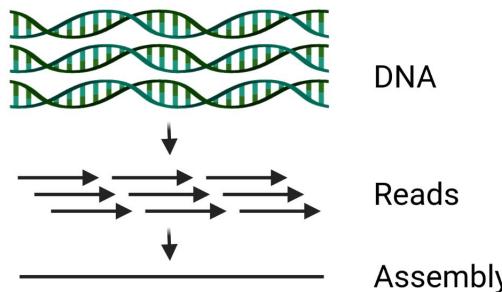
Get access to the Cloud

1.) Register for an [ELIXIR Account](#) and apply for membership in the de.NBI virtual organisation.

2.) Log in to our [de.NBI Cloud portal](#) to manage your projects and project members.

Genome sequence assembly

- Combination of overlapping reads to longer sequences
- Assembled sequences represent chromosomes
- Sequences are stored in FASTA file



```
>TRINITY_DN100016_c0_g1_i1
MPKKSSNIQQNNQNRIKRKGRPPKKKKYVQQIDSSDEDILSVRHASTRPRIISIRRNEIPMRPEIHI
>TRINITY_DN100019_c0_g1_i1
MPPKAADKKPAAKAPATASKAPEKKDAGKKTAATGDKKKRTKARKETYSSYIYKVLKVHPDTGISN
>TRINITY_DN10001_c0_g1_i2
MAKVGNPPIVDIETDGSVNEPESSEKNIEVSSSSTQAPESTNTTELLVNEKKAFSLATPAVRRVAREH
>TRINITY_DN100025_c0_g1_i1
MVENQDGCFFKPGKEFVRNSNDLEGGDFLVNLVDKISYQVVIFDGTCCPKDLCFPSIMNPIFIQHLR
>TRINITY_DN10002_c0_g1_i1
MSDELNQVFQRCREQKRPVFAVTAGYPDSEETVDILLGLEAGGADIELGIPFTDPMDGKTIQD.
>TRINITY_DN100061_c0_g1_i1
MQQVVAKLKAIITKTNVTNENSPVENSSSTSATSSIINSLHGDLSRVFDNMELESNVSNSSISSNI
```

Annotation

- Prediction of gene structures in genome sequence
 - Genes are described in text file: Generic Feature Format (GFF)

Chr1	TAIR10	chromosome	1	30427671	ID=Chr1;Name=Chr1
Chr1	TAIR10	gene	3631	5899	ID=AT1G01810;Note=protein_coding_gene;Name=AT1G01810;Parent=AT1G01810;Index=1
Chr1	TAIR10	protein	3630	5899	ID=AT1G01810_1;Protein;Name=AT1G01810_1;Derives_from=AT1G01810_1
Chr1	TAIR10	exon	3631	3913	Parent=AT1G01810_1
Chr1	TAIR10	five_prime_UTR	3631	3759	Parent=AT1G01810_1;AT1G01810_1
Chr1	TAIR10	CDS	3760	3913	0	ID=AT1G01810_1;Parent=AT1G01810_1
Chr1	TAIR10	exon	3996	4276	Parent=AT1G01810_1;AT1G01810_1
Chr1	TAIR10	CDS	3996	4276	2	Parent=AT1G01810_1;AT1G01810_1
Chr1	TAIR10	exon	4486	4605	0	Parent=AT1G01810_1;AT1G01810_1
Chr1	TAIR10	exon	4486	4605	0	Parent=AT1G01810_1;AT1G01810_1
Chr1	TAIR10	CDS	4486	4605	0	Parent=AT1G01810_1;AT1G01810_1
Chr1	TAIR10	exon	4796	5095	0	Parent=AT1G01810_1;AT1G01810_1
Chr1	TAIR10	CDS	4796	5095	0	Parent=AT1G01810_1;AT1G01810_1
Chr1	TAIR10	exon	5174	5326	0	Parent=AT1G01810_1;AT1G01810_1
Chr1	TAIR10	CDS	5174	5326	0	Parent=AT1G01810_1;AT1G01810_1
Chr1	TAIR10	exon	5439	5890	0	Parent=AT1G01810_1;AT1G01810_1
Chr1	TAIR10	CDS	5439	5890	0	Parent=AT1G01810_1;AT1G01810_1
Chr1	TAIR10	three_prime_UTR	5631	5899	0	Parent=AT1G01810_1;AT1G01810_1
Chr1	TAIR10	gene	5928	8737	ID=AT1G01820;Note=protein_coding_gene;Name=AT1G01820;Parent=AT1G01820;Index=1
Chr1	TAIR10	protein	5928	8737	ID=AT1G01820_1;Protein;Name=AT1G01820_1;Derives_from=AT1G01820_1
Chr1	TAIR10	protein	6915	8666	Parent=AT1G01820_1
Chr1	TAIR10	five_prime_UTR	8667	8737	0	Parent=AT1G01820_1;AT1G01820_1
Chr1	TAIR10	CDS	8571	8846	0	Parent=AT1G01820_1;AT1G01820_1
Chr1	TAIR10	exon	8927	8737	0	Parent=AT1G01820_1;AT1G01820_1
Chr1	TAIR10	CDS	8417	8464	0	Parent=AT1G01820_1;AT1G01820_1
Chr1	TAIR10	exon	8417	8464	0	Parent=AT1G01820_1;AT1G01820_1

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

BLAST+ 2.13.0 is here!

Starting with this release, we are including the blastn, vdb and tblastn_vdb executables in the BLAST+ distribution.

Thu, 17 Mar 2022 12:00:00 EST [More BLAST news...](#)

Web BLAST



Nucleotide BLAST
nucleotide ▶ nucleotide



blastx
translated nucleotide ▶ protein



tblastn
protein ▶ translated nucleotide



Protein BLAST
protein ▶ protein

- Assignment of functions to predicted genes
 - BLAST, InterProScan5, Pfam

Read mapping

- Alignment of reads to a genome sequence
- Mapping needs to be fast at the cost of accuracy
- Long read mapping tools: minimap2, GraphMap, NGMLR
- Manual inspection of read mappings via Integrative Genomics Viewer (IGV)

Variant calling

- Identification of sequence differences between reads and reference sequence
- Differences are listed in a specific file type: Variant Caller Format (VCF)
- ONT long reads are well suited for the identification of large structural variants

Databases for submission of sequencing data

- Sequence Read Archive



- European Nucleotide Archive



ENA
European Nucleotide Archive

Home Submit Search Rulespace About Support

Message updated 2022-05-13.

We're happy to announce that our submission services are back to normal after a brief downtime.

We recommend that you subscribe to the ENA-announce mailing list.

Support Guides
Downloading Data
Contact Us

For SARS-CoV-2 data submissions, users should contact us in advance of submission at virus-data@ebi.ac.uk for specific advice on options and to access the highest levels of support. We have also launched a Drag-and-Drop Data Submission Service (currently in beta) suitable for certain SARS-CoV-2 submissions. We are inviting submitters to try this out. Please contact us at the email above for details.

European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. More about ENA.

Access to ENA data is provided through the browser, through search tools, through large scale file download and through the API.

Submit Search Rulespace Support

Enter text search terms
Barcode Name: B000000
Enter accession
Barcode: Taconic:R008, B000000, PRZEBA02

Tweets by @ENASequences

European Nucleotide Archive (ENA) Retweeted

ELIXIR Europe

Nanopore sequencing trends

Barcode

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

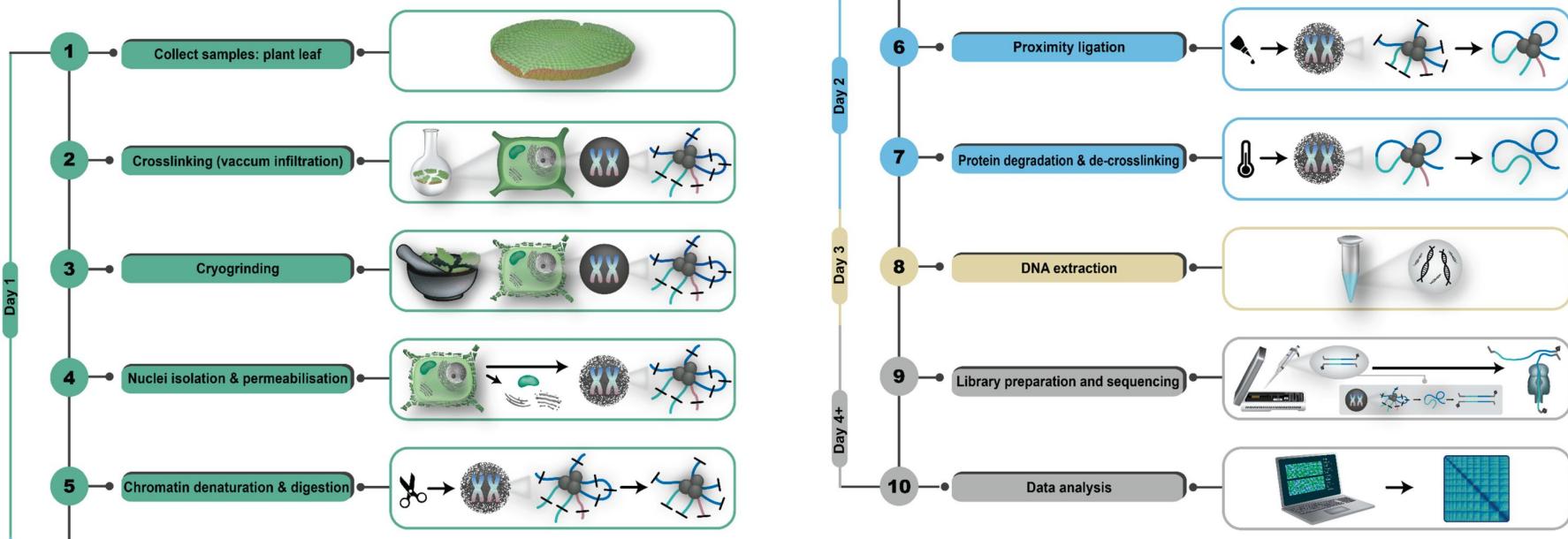
'Read Until' adaptive sampling / Targeted sequencing

- Check DNA strand based on first sequenced bases against reference:
 - strand is of interest: continue
 - strand is not of interest: reject
 - Example: only sequence bacterial DNA included in a plant sample
- Cas9 Sequencing Kit:
 - Dephosphorylation of 5'-DNA ends
 - Cas9 binds to target sequences (based on guide RNA)
 - Cleavage results in blunt ends and 5'-phosphorylation
 - 3' dA-tailing to prepare for adapter ligation
 - Adapters are preferentially ligated to Cas9 cut sites

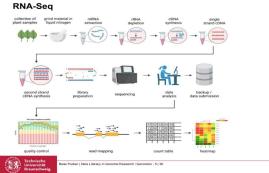
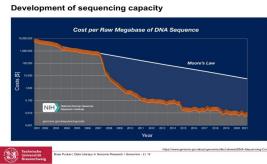
Direct RNA sequencing

- PolyA tail (3') required for primer ligation
- Generation of complementary DNA strand through reverse transcriptase recommended
- Sequencing of RNA strand allows analysis of base modifications

Pore-C (Hi-C with ONT)



Teaching sequencing in practical courses

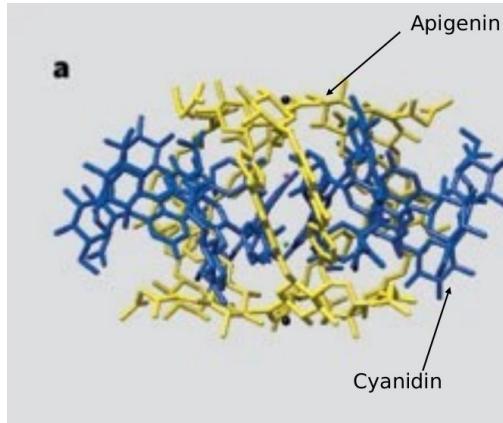


Data Literacy and
Plant Genomics

Genome sequencing projects (examples)

Centaurea cyanus

- Famous for a striking blue flower color
- Complex comprising pigments and metal
- Genome sequence reveals pigment biosynthesis genes

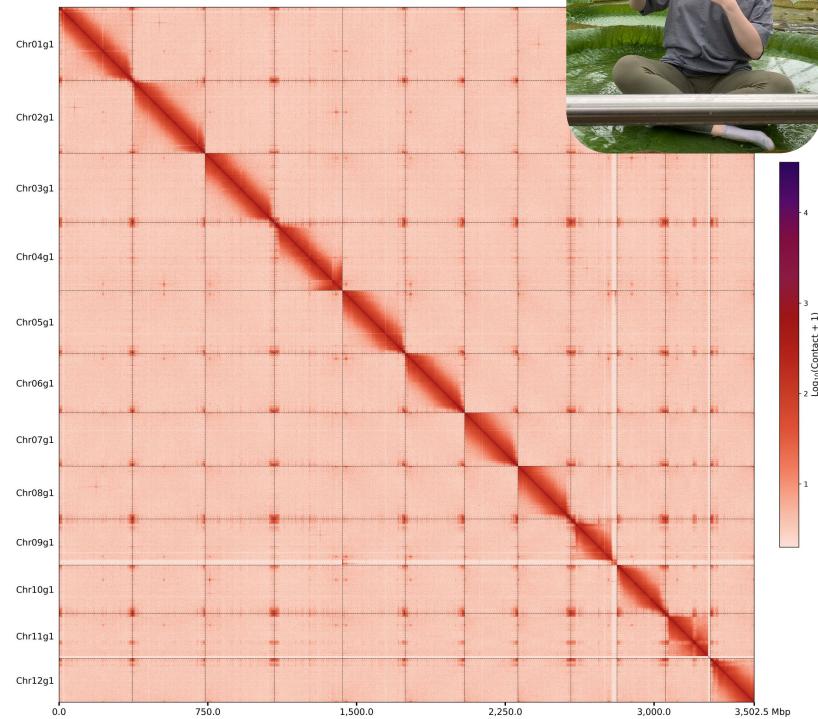
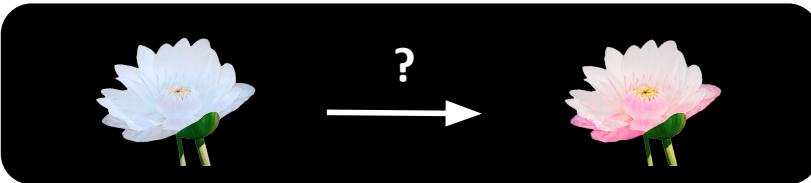


Shiono, M., Matsugaki, N. & Takeda, K. Structure of the blue cornflower pigment. *Nature* **436**, 791 (2005). <https://doi.org/10.1038/436791a>



Victoria cruziana

- Water lily with giant floating leaves
- Flowers at night and floral color transitions from white to pink
- Pore-C resulted in an almost perfect genome sequence



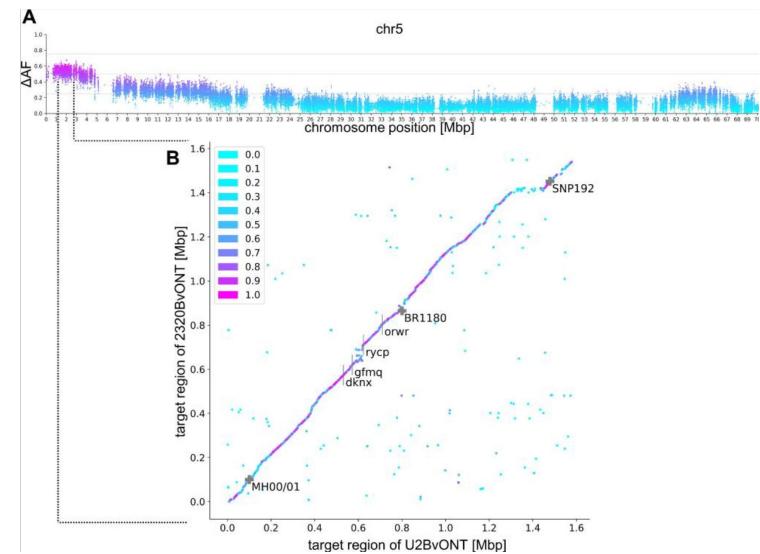
Aquilegia vulgaris

- Popular ornamental plant
- Huge variety of different flower colors
- Genome sequence provides access to pigment biosynthesis genes



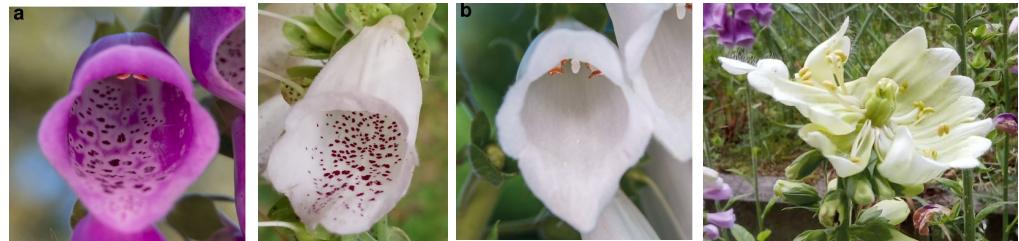
Beta vulgaris

- Important crop in northern hemisphere
- Accounts for about 20% of sugar production
- Nematodes cause serious yield losses



Digitalis purpurea

- Ornamental plant with magenta flowers
- Flower color and shape variation
- Genome sequence reveals underlying genes



Summary

- Planning the sequencing experiment
- DNA extraction, quality control, library preparation
- Nanopore sequencing & real time monitoring
- Data analysis (overview)
- Nanopore sequencing applications
- Genome sequencing projects (examples)

Time for questions!

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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1017/qpb.2021.18).