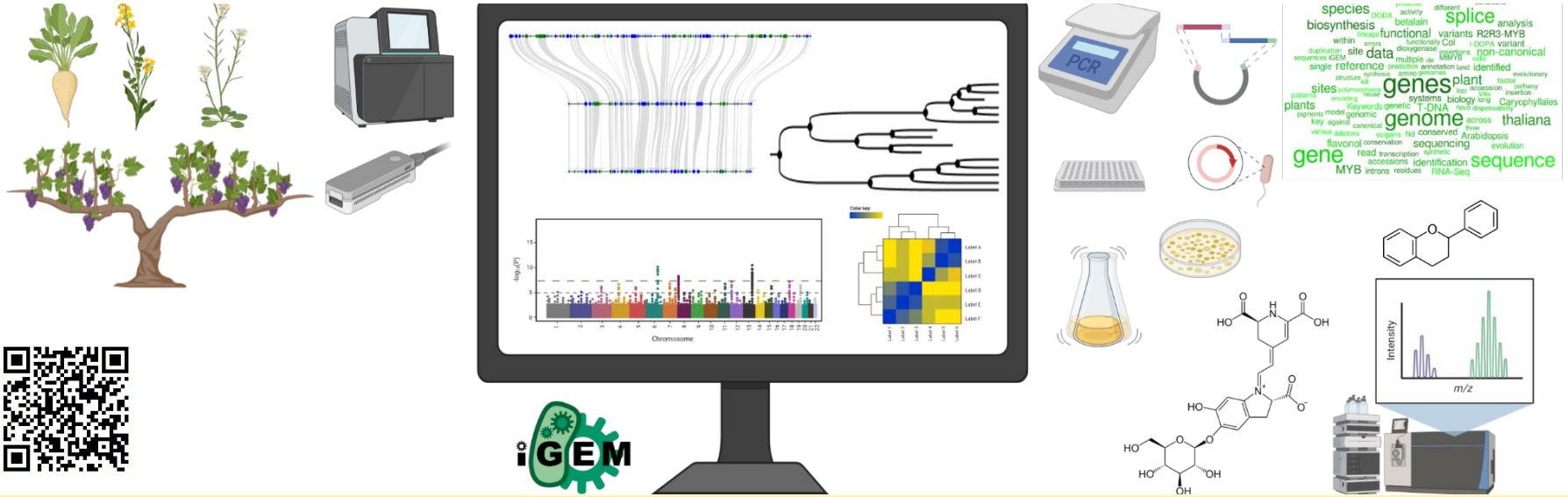
Technische Universität Braunschweig



ONT: from DNA extraction to sequencing

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

Availability of slides

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

- Questions: Feel free to ask at any time

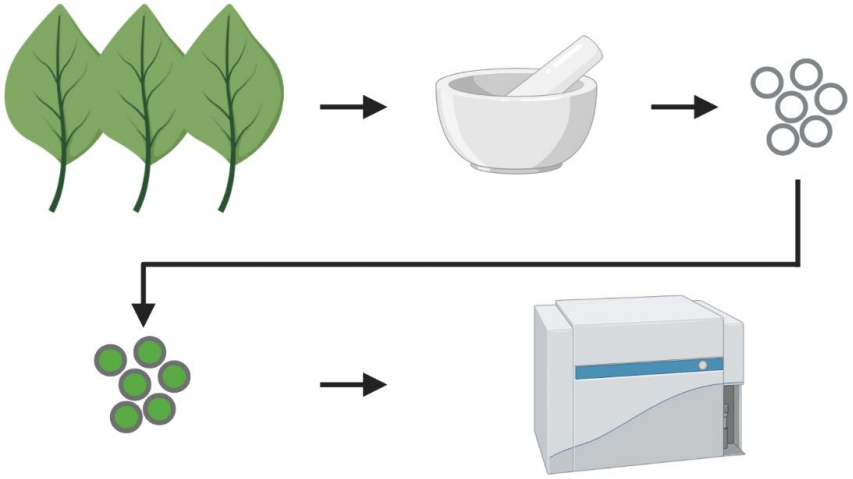
- Feedback, comments, or questions: b.pucker[a]tu-braunschweig.de

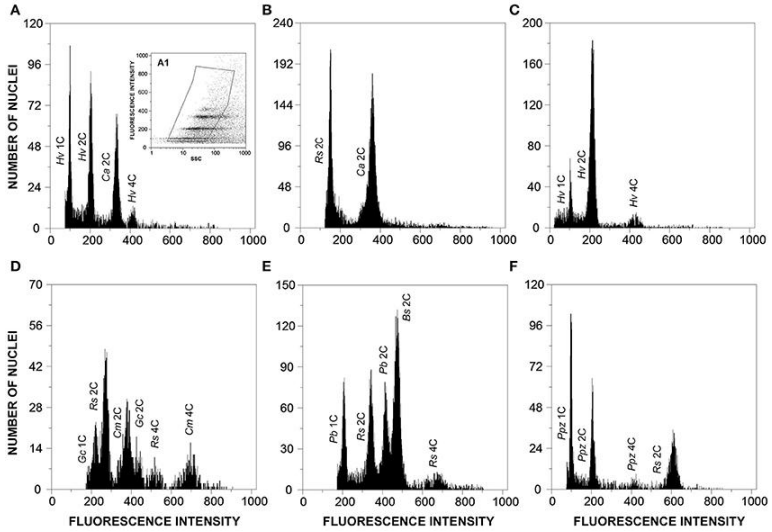
Considerations & preparations

- 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/



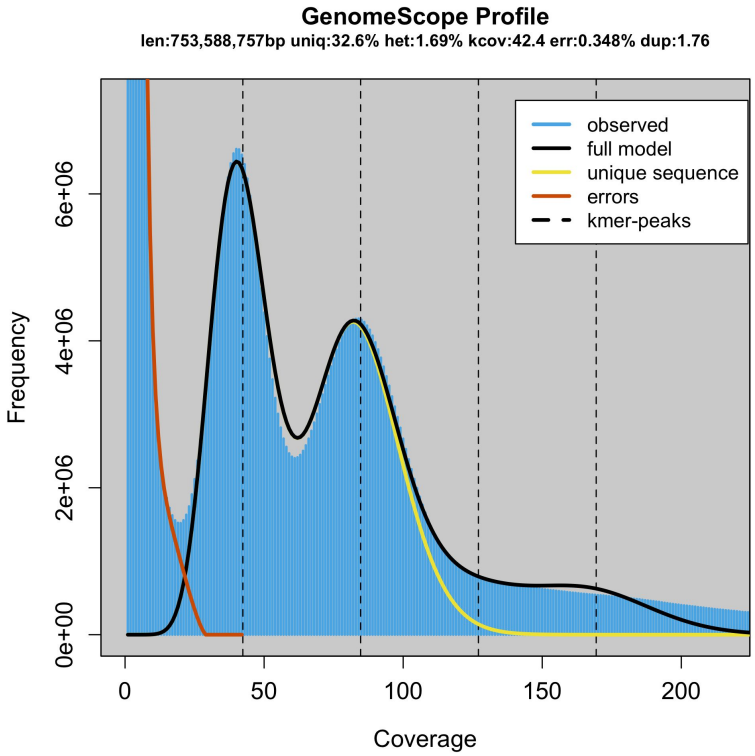


https://doi.org/10.3389/fpls.2014.00422

Genome size - 2

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

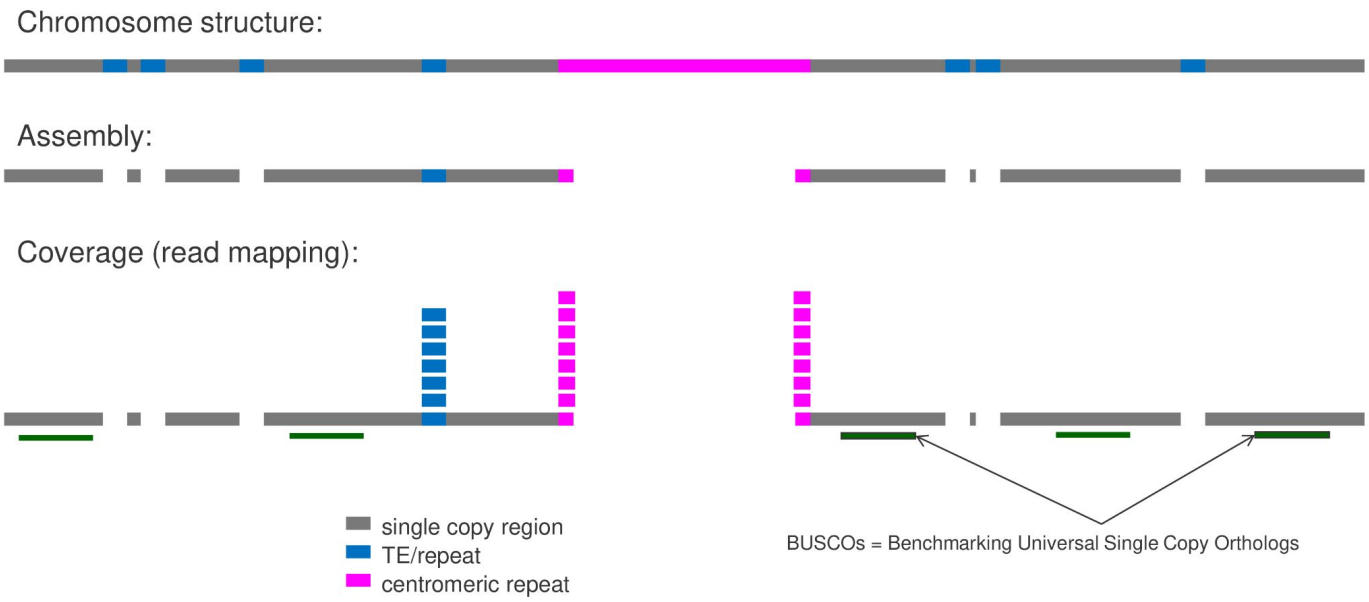
```
ACGAAGCCATAT

ACGAAGC

  CGAAGCC

    GAAGCCA

      AAGCCAT

        AGCCATAT
```



https://github.com/schatzlab/genomescope

Genome size - 3

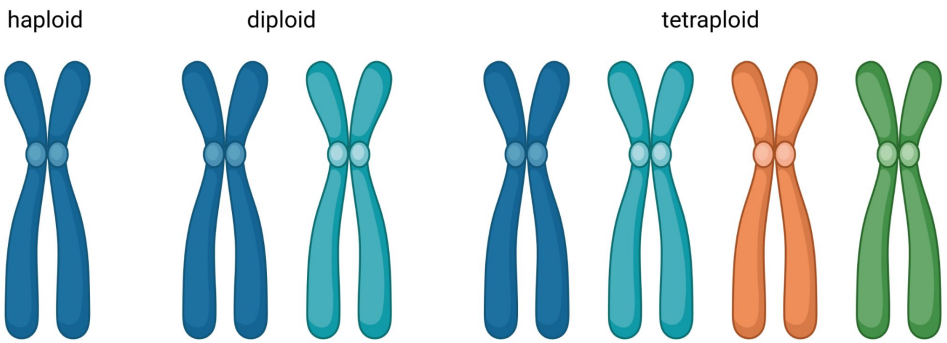
- Tools for genome size estimation based on short/long reads
  - Mapping-based: MGSE, Gnodes

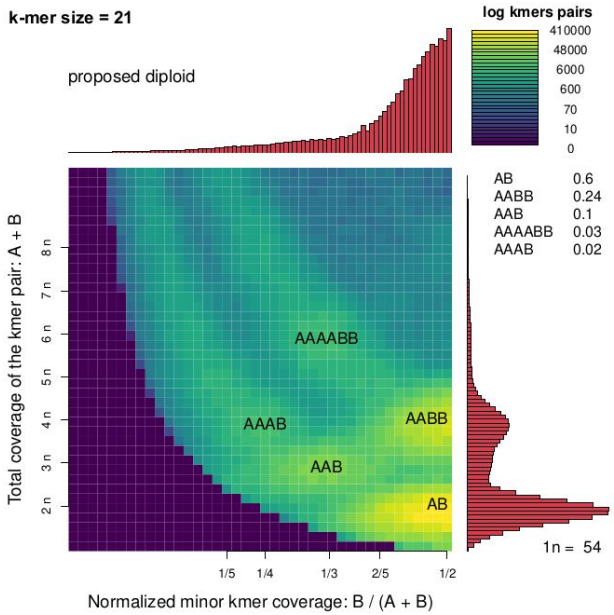


Pucker, 2019: 10.1101/607390

Ploidy - 1

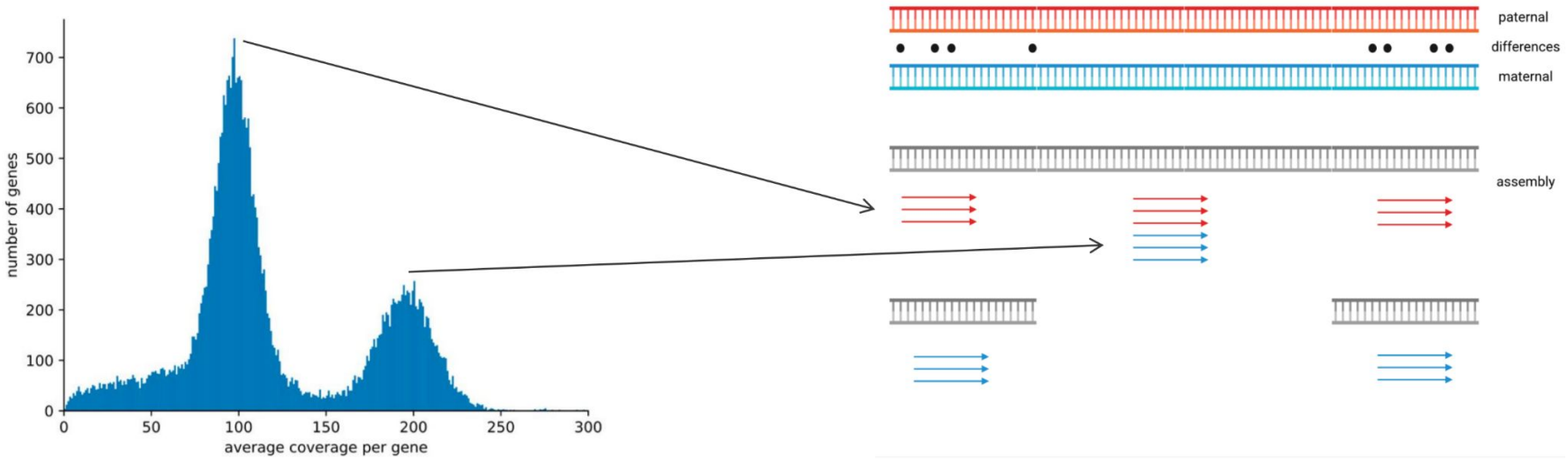
- Ploidy = copy number of same chromosome
- Many plants are polyploid
- Some polyploid plants have close diploid relatives
- Smudgeplots to analyze polyploidy based on short reads





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

Ploidy - 2



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

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-mercaptoehtanol (ß-ME)
- Dichlormethane (replaces chloroform)
- TrisEDTA (TE)
- RNase A
- Short Read Eliminator (SRE) kit

Material required for sequencing

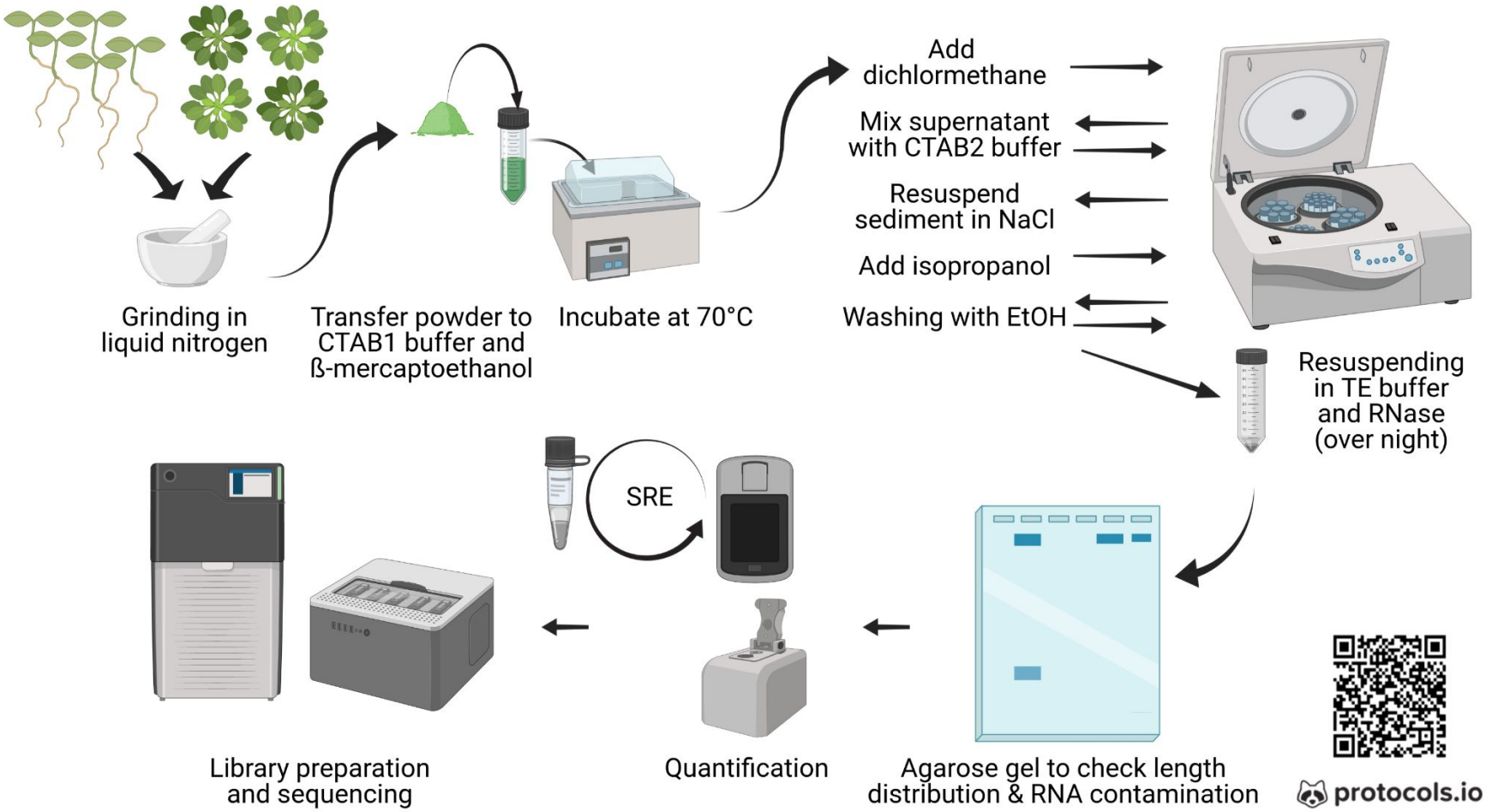
- DNA repair (companion module)
- Magnetic beads for purification
- Library preparation kit
- Sequencing kit
- Wash kit with DNase

Workflow

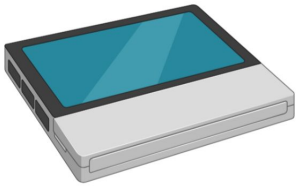
		task	consumed time	hands-on time	equipment	estimated costs of consumables	estimated costs of lab equipment
A		plant incubation in darkness	2-3d	1h			
B		non-destructive sampling	-	1h			
C		DNA extraction	1d	8h	waterbath, centrifuge	$50	$1000 $8000
D		quality control	1h	1h	NanoDrop, Qubit	$20	
E		short 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		

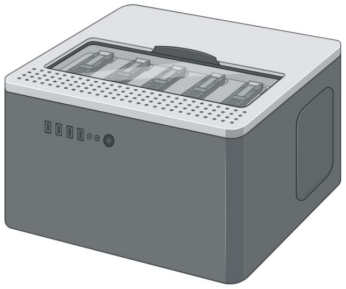
Pucker et al., 2022: 10.1017/qpb.2021.18

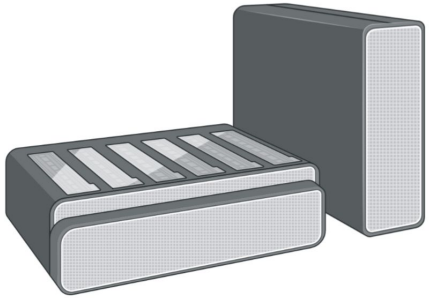
DNA extraction workflow



ONT sequencing devices



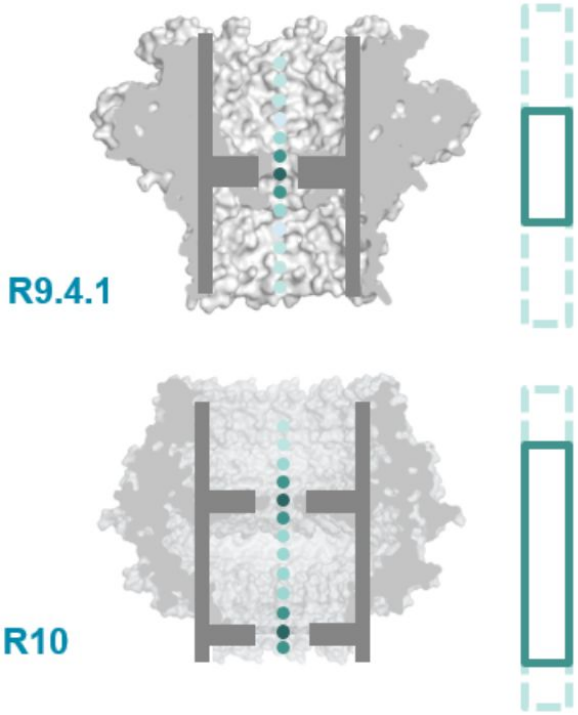




ONT vs. PacBio

	ONT	PacBio (HiFi)
Maximal read length	DNA molecule size	25kb
Raw read accurracy	99% with Q20+	99.5%
DNA input	1 µg	3 µg
Instrument costs	$1000 (MinION)	High
Costs per genome	$3000 per Gbp	

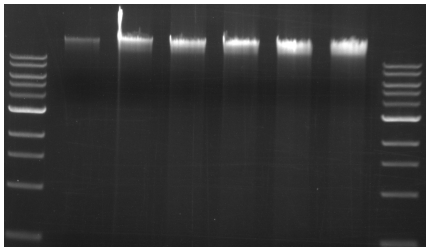
Nanopore comparison



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

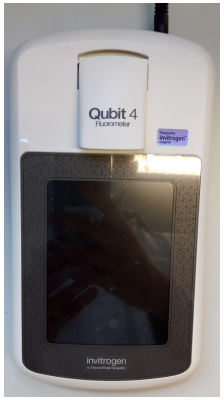
Quality control

- Agarose gel electrophoresis



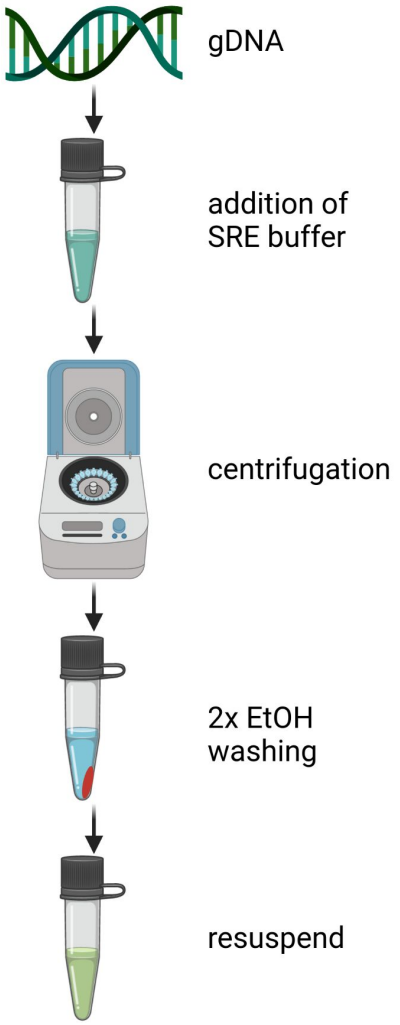
- Photometric measurement via NanoDrop

- Quantification with Qubit



Short Read Eliminator (SRE)

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



DNA repair

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

Library preparations

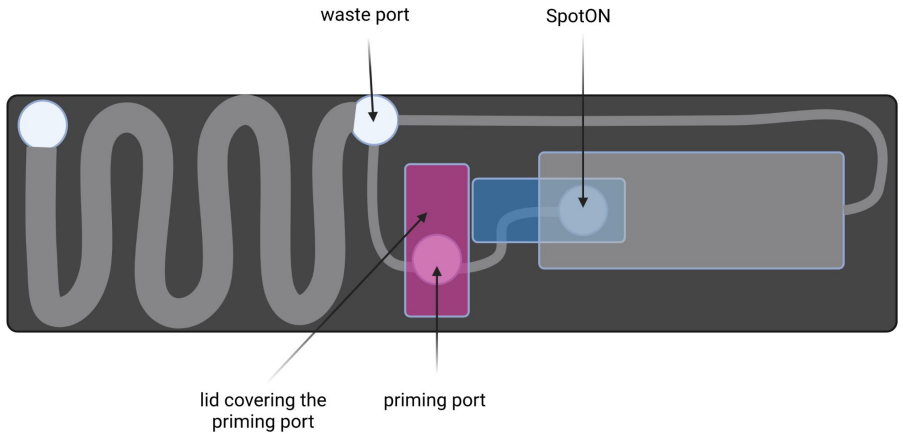
- Repaired DNA is subjected to library preparation
- Addition of adapters to DNA fragments
- Concentration of DNA can be quantified via Qubit measurement (optional)
  - Control step to ensure that library construction is working
- Purification of DNA with magnetic beads

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
- 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

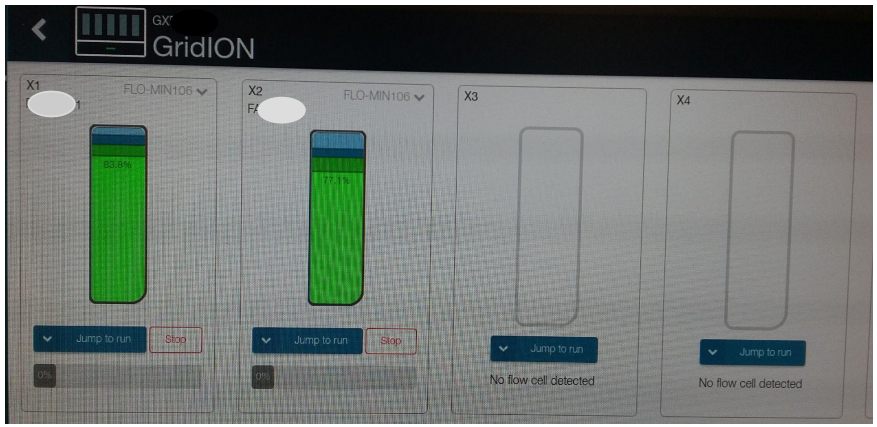


Starting sequencing

- Define the flow cell type (R9.4.1)
- Select parameters (default)
- Set the output location, run name, …
- Start the sequencing about 30 minutes after loading the flow cell
  - This allows the DNA to get into contact with the nanopores

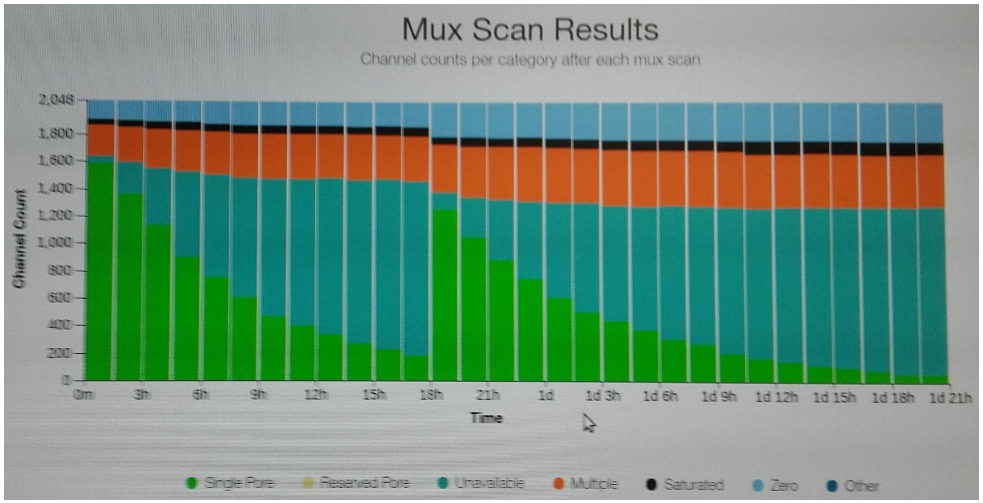
Monitor sequencing

- 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



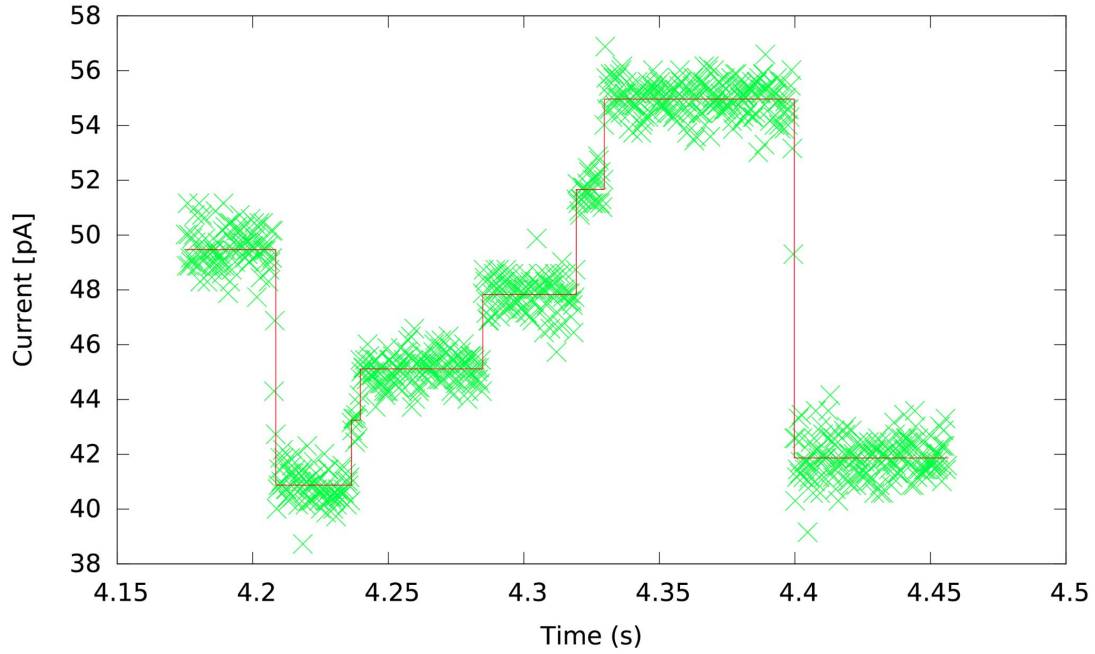
Stop & wash

- Sequencing is stopped once the number of active pores gets low
- Washing with DNase to free blocked nanopores
- Flow cells are regenerated and can be re-used
- Process can be repeated multiple times (3-5x)



Basecalling

- Electric signal is converted into sequence information (bascalling)
- Algorithmic improvement lead to higher read accuracy
- Raw sequencing data need to be stored



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

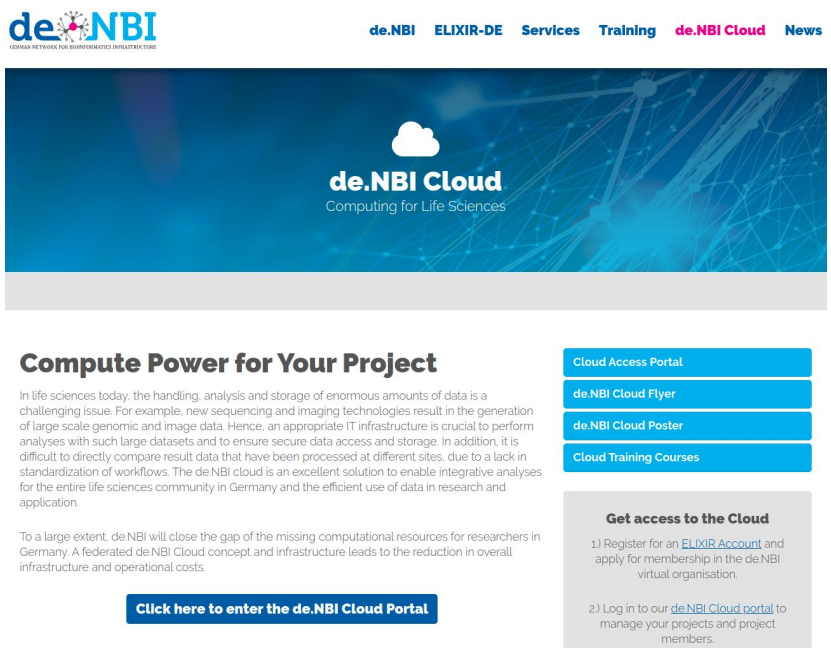
Bioinformatics / Data analysis

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/

de.NBI cloud

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



Connecting to virtual machine

- Secure shell (ssh)
- Private & public keys

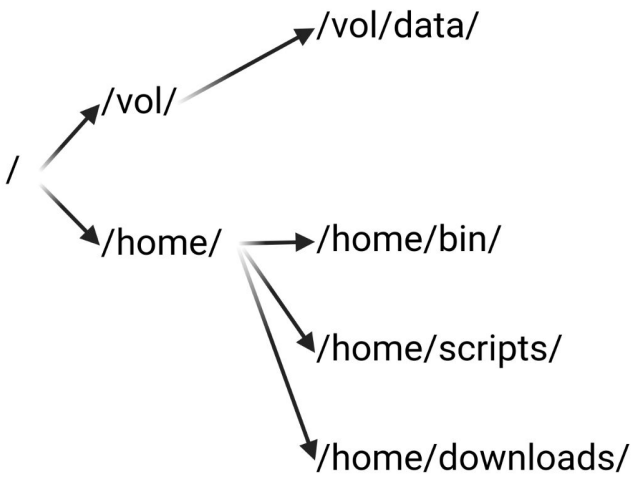
X2GO

- Adding a graphical user interface to the virtual machine (VM)
- Sessions keeps running when user disconnects
- Excellent solution for users with individual VMs
- Problem: only one user can access VM at a time

https://wiki.x2go.org/doku.php

Introduction to 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 ‘$’



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

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
```

Running other tools

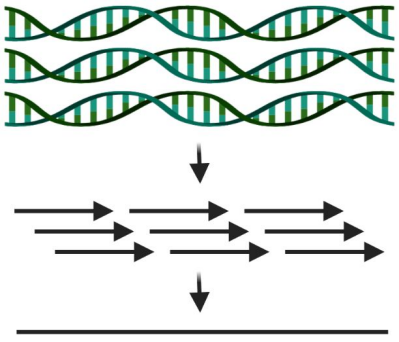
- Show help message:
  - $ <NAME_OF_TOOL>
  - $ <NAME_OF_TOOL> -h
  - $ <NAME_OF_TOOL> -- help

- Providing arguments is different for each tool:

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

Genome sequence assembly

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



DNA

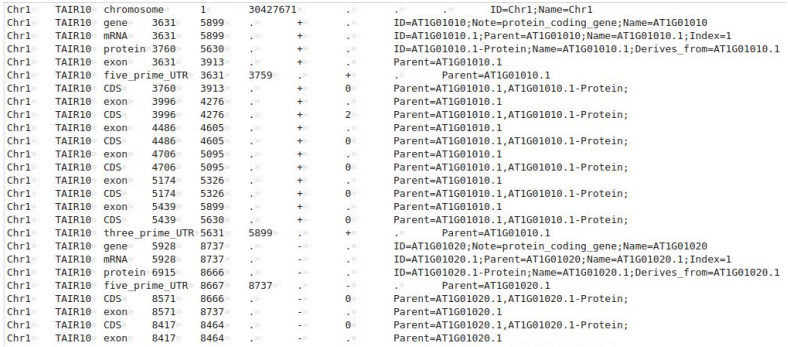
Reads

Assembly

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

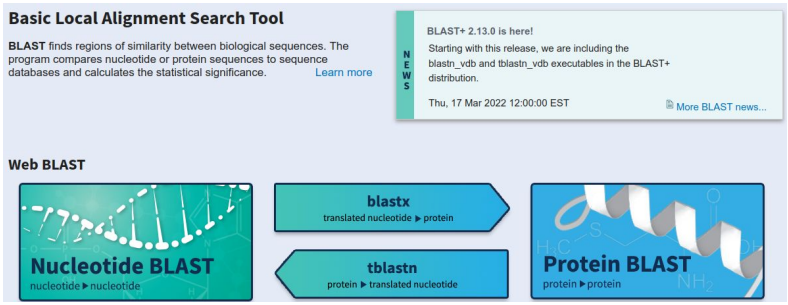
Annotation

- 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=AT1G01010;Note=protein_coding_gene;Name=AT1G01010
Chr1    TAIR10  mRNA    3631    5899    .       +       .       ID=AT1G01010.1;Parent=AT1G01010;Name=AT1G01010.1;Index=1
Chr1    TAIR10  protein 3760    5630    .       +       .       ID=AT1G01010.1-Protein;Name=AT1G01010.1;Derives_from=AT1G01010.1
Chr1    TAIR10  exon    3631    3913    .       +       .       Parent=AT1G01010.1
Chr1    TAIR10  five_prime_UTR  3631    3759    .       +       .       Parent=AT1G01010.1
Chr1    TAIR10  CDS     3760    3913    .       +       0       Parent=AT1G01010.1,AT1G01010.1-Protein;
Chr1    TAIR10  exon    3996    4276    .       +       .       Parent=AT1G01010.1
Chr1    TAIR10  CDS     3996    4276    .       +       2       Parent=AT1G01010.1,AT1G01010.1-Protein;
Chr1    TAIR10  exon    4486    4605    .       +       .       Parent=AT1G01010.1
Chr1    TAIR10  CDS     4486    4605    .       +       0       Parent=AT1G01010.1,AT1G01010.1-Protein;
Chr1    TAIR10  exon    4706    5095    .       +       .       Parent=AT1G01010.1
Chr1    TAIR10  CDS     4706    5095    .       +       0       Parent=AT1G01010.1,AT1G01010.1-Protein;
Chr1    TAIR10  exon    5174    5326    .       +       .       Parent=AT1G01010.1
Chr1    TAIR10  CDS     5174    5326    .       +       0       Parent=AT1G01010.1,AT1G01010.1-Protein;
Chr1    TAIR10  exon    5439    5899    .       +       .       Parent=AT1G01010.1
Chr1    TAIR10  CDS     5439    5630    .       +       0       Parent=AT1G01010.1,AT1G01010.1-Protein;
Chr1    TAIR10  three_prime_UTR 5631    5899    .       +       .       Parent=AT1G01010.1
Chr1    TAIR10  gene    5928    8737    .       -       .       ID=AT1G01020;Note=protein_coding_gene;Name=AT1G01020
Chr1    TAIR10  mRNA    5928    8737    .       -       .       ID=AT1G01020.1;Parent=AT1G01020;Name=AT1G01020.1;Index=1
Chr1    TAIR10  protein 6915    8666    .       -       .       ID=AT1G01020.1-Protein;Name=AT1G01020.1;Derives_from=AT1G01020.1
Chr1    TAIR10  five_prime_UTR  8667    8737    .       -       .       Parent=AT1G01020.1
Chr1    TAIR10  CDS     8571    8666    .       -       0       Parent=AT1G01020.1,AT1G01020.1-Protein;
Chr1    TAIR10  exon    8571    8737    .       -       .       Parent=AT1G01020.1
Chr1    TAIR10  CDS     8417    8464    .       -       0       Parent=AT1G01020.1,AT1G01020.1-Protein;
Chr1    TAIR10  exon    8417    8464    .       -       .       Parent=AT1G01020.1
```

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



https://blast.ncbi.nlm.nih.gov/Blast.cgi

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 Integrated Genomics Viewer (IGV)

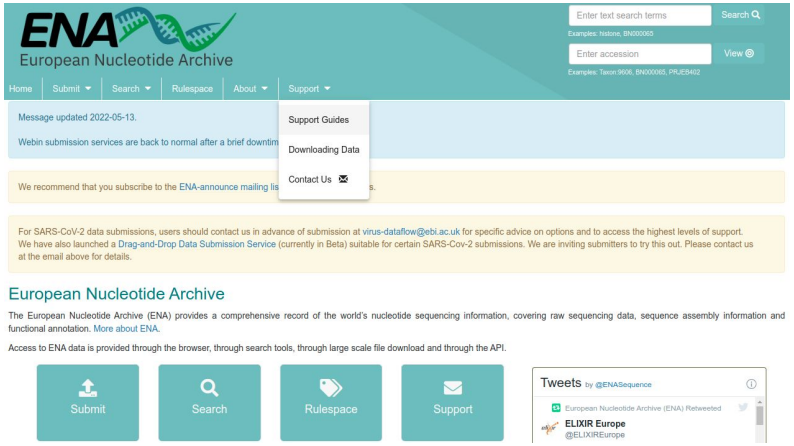
Variant calling

- 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



Nanopore sequencing trends

Barcoding

- 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

- Examples: UNCALLED (Kovaka et al., 2020); COSMIC (Payne et al., 2020)

- 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 Ca9 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

Time for questions!

Questions

1. Which methods allow a genome size estimation?
2. What are the important steps of a high molecular weight DNA extraction protocol?
3. Which preparations are required for nanopore sequencing?
4. Which steps are involved in a nanopore sequencing workflow?
5. What can be monitored during the sequencing run?
6. Which methods can be used to transfer the data?
7. Which databases can be used to store the sequencing data?