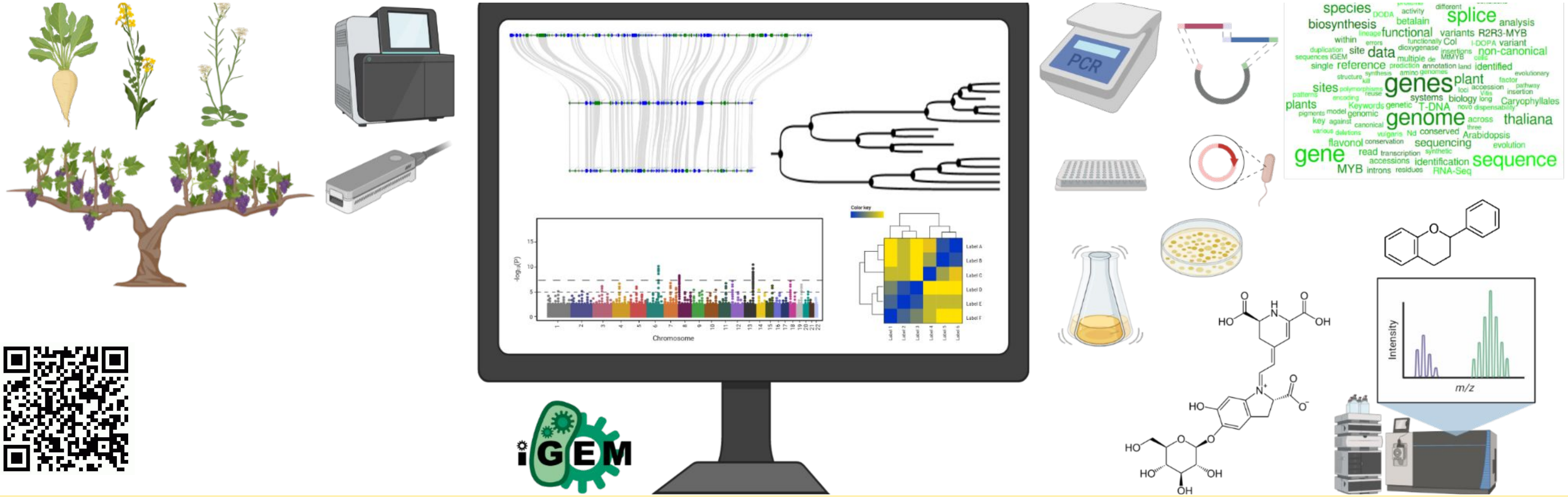
Technische
Universität
Braunschweig



# Genome Sequence Assembly

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

# Availability of slides

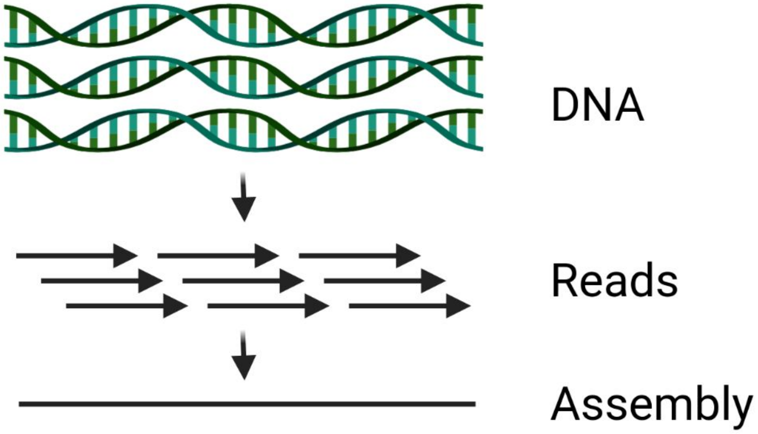
- All materials are freely available (CC BY) - after the lectures:
  - StudIP: **GE31/MM12**
  - GitHub: https://github.com/bpucker/teaching

- Questions: Feel free to ask at any time

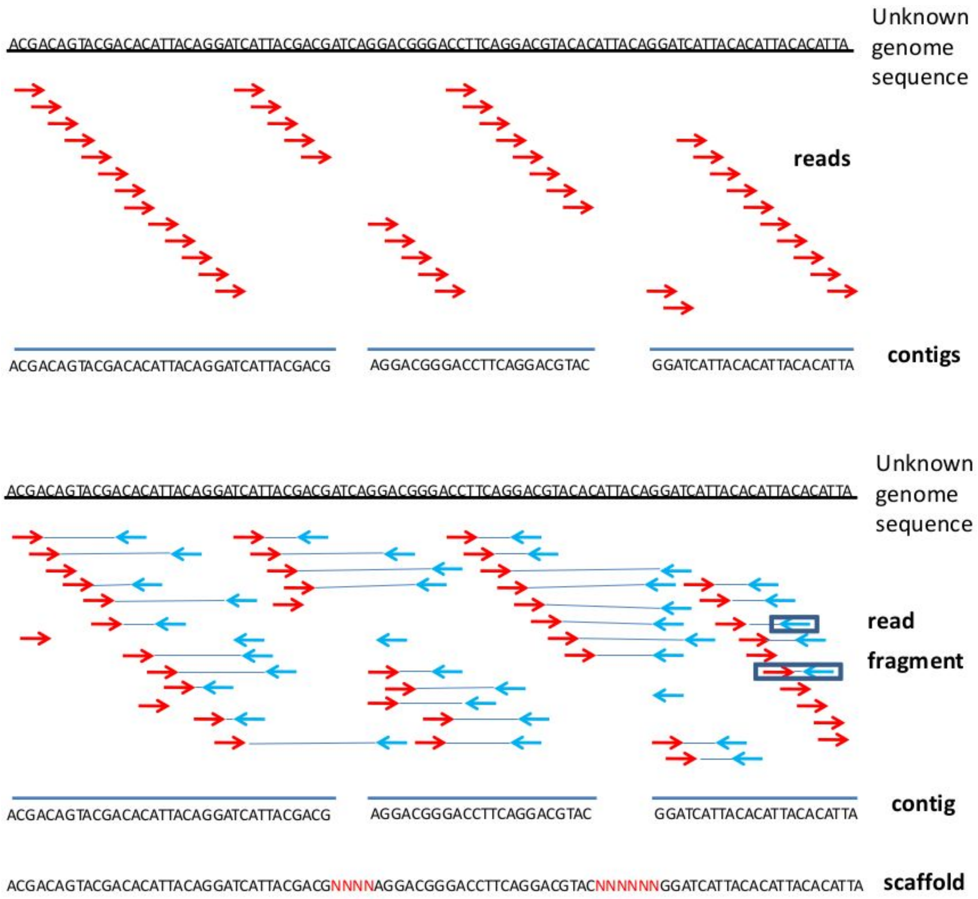
- Feedback, comments, or questions: b.pucker[a]tu-braunschweig.de

# 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



ACGACAGTACGACACATTACAGGATCATTACGACGATCAGGACGGGACCTTCAGGACGTACACATTACAGGATCATTACACATTACACATTA

Unknown genome sequence

reads

ACGACAGTACGACACATTACAGGATCATTACGACG AGGACGGGACCTTCAGGACGTAC GGATCATTACACATTACACATTA

contigs

ACGACAGTACGACACATTACAGGATCATTACGACGATCAGGACGGGACCTTCAGGACGTACACATTACAGGATCATTACACATTACACATTA

Unknown genome sequence

read

fragment

ACGACAGTACGACACATTACAGGATCATTACGACG AGGACGGGACCTTCAGGACGTAC GGATCATTACACATTACACATTA

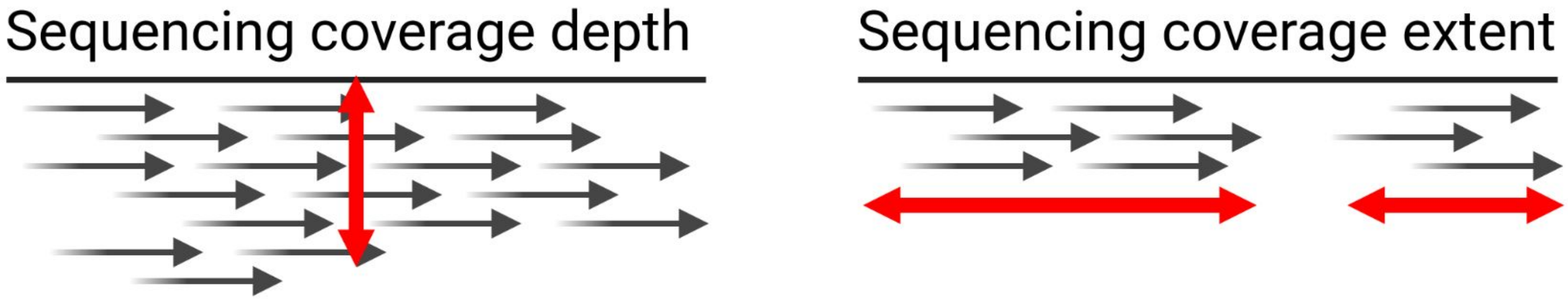
contig

ACGACAGTACGACACATTACAGGATCATTACGACGNNNNAGGACGGGACCTTCAGGACGTACNNNNNNGGATCATTACACATTACACATTA

scaffold

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

# Sequencing coverage depth vs. coverage extent



# Sequencing coverage depth

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

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

- Coverage (depth) reflects total amount of sequencing data

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

# Sequencing coverage extent

- Coverage extent = ratio of genome covered by at least one base

- Informative to calculate required sequencing depth for a project

- Coverage extent follows a Poisson distribution

- Calculation of coverage extent (c):
  - Non-coverage extent: $P(X=0) = e^{-c}$
  - Probability of coverage extent: $P(X>0) = 1 - e^{-c}$

- Complete coverage of genome requires $G \times e^{-c} < 1$

- Larger genomes required higher sequencing depth

- Real coverage extent is often suffering from sequencing bias

# Assembly paradigms

- Greedy
- Overlap layout consensus (OLC)
- De Bruijn Graph (DBG)

# Greedy

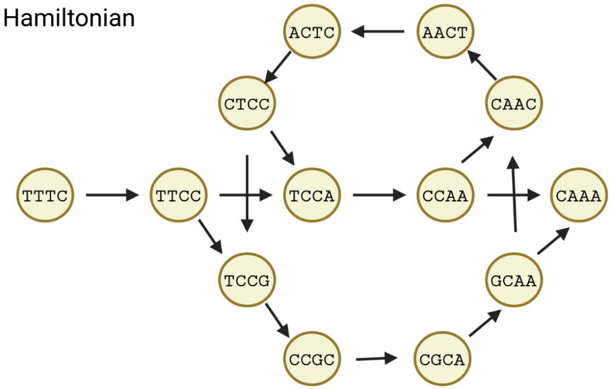
- Assemble reads that overlap best
- Choices are inherently local => paired-end / mate pair information not used
- Heuristics needed to avoid misassembling repeats
- Paradigm was used for early assemblers: phrap, TIGR

# Overlap layout consensus (OLC)

- 1) Identification of all reads with sufficient overlaps (all vs. all comparison)
- 2) Construction of graph based on layout information
  - Nodes represent reads
  - Edges represent overlaps
- 3) Inference of consensus sequence from graph
- Complexity/Problems:
  - Number of nodes is equal to number of reads
  - Number of edges increases logarithmically
  - Hamiltonian path problem is NP-hard

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

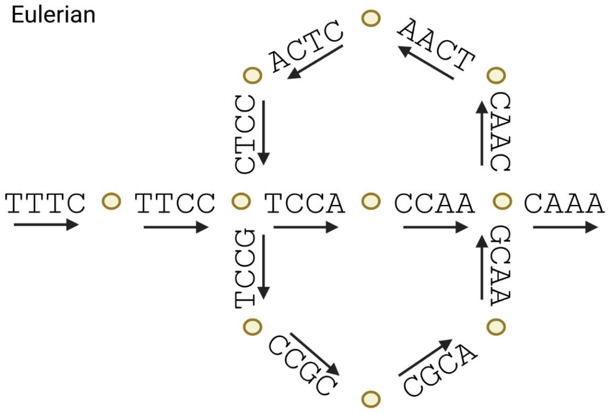
Hamiltonian



# De Brujin graph (DBG)

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

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

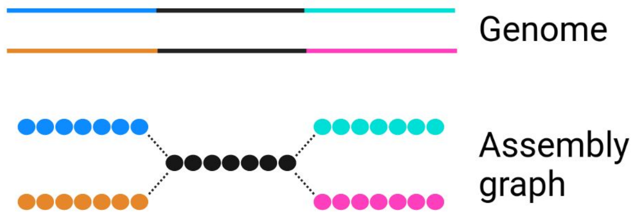


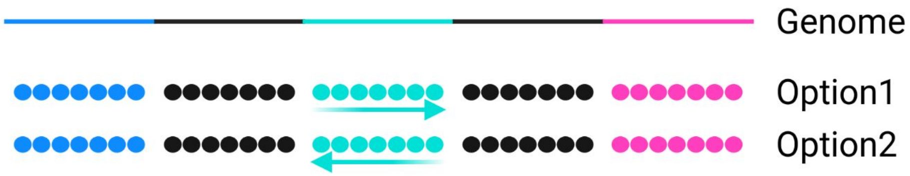
# K-mer size

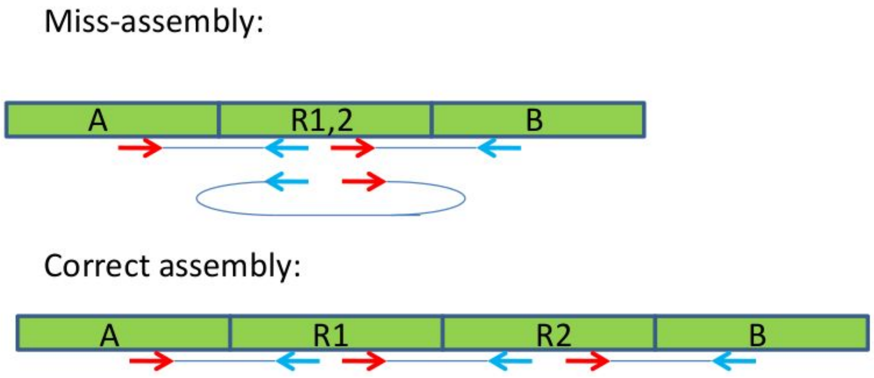
- Larger k-mers increase the assembly continuity by spanning repeats
- Larger k-mers are more sensitive to sequencing errors
- Removal of low abundance k-mers (caused by sequencing errors)
- K-mer size should be 0.5 to 0.75 of the read length
- Typical k-mer sizes:
  - 87 or 97 for 2x150bp reads

# Assembly challenges

- Collapsed repeats
- Inversions
- Overstretched repeats

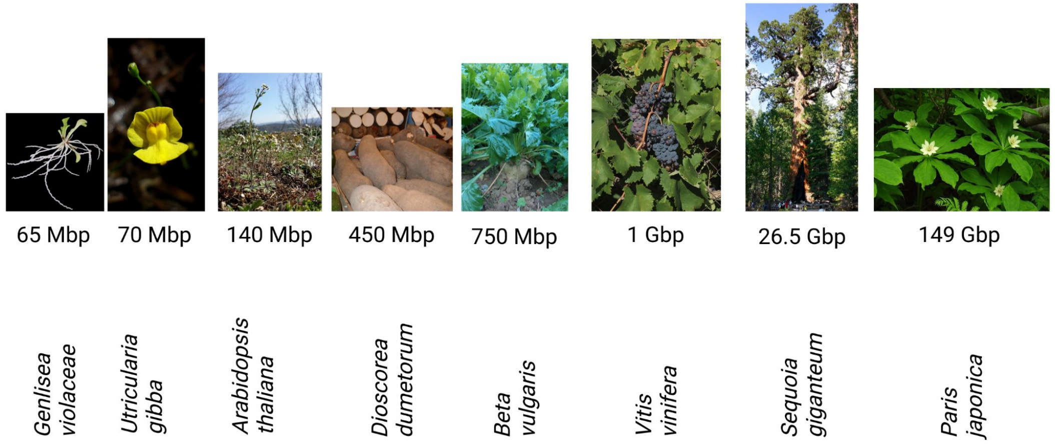






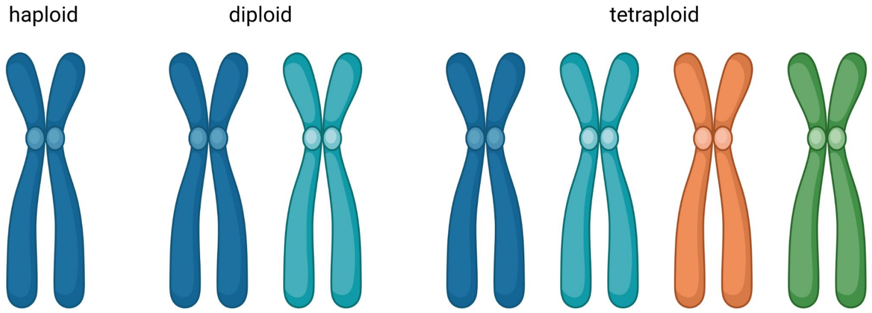
# Assembly challenges (2)

- Genome size: variation from 65 Mbp to 149 Gbp



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

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



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

# Assembly polishing

ONT
assembly

Illumina
reads

```
...ACGTTACGGTACGACATGACGTAAAAAAAAACGATAGTACGTGTTAGCGTACGTACGTAACGTT...

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

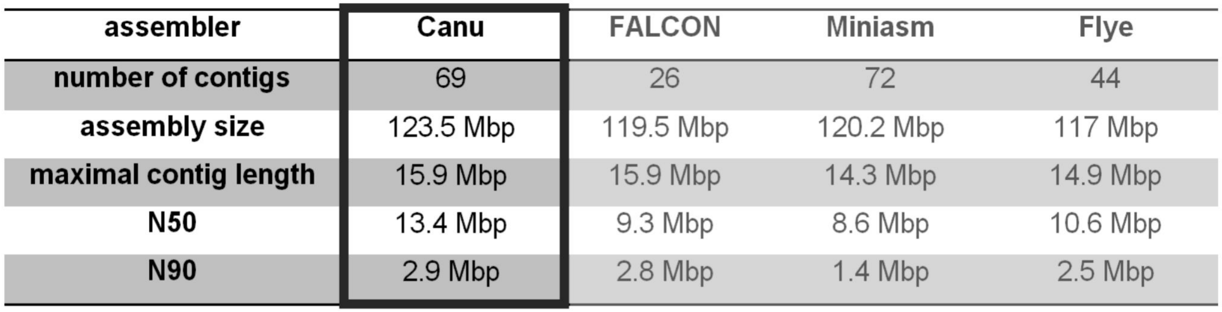
Technische Universität Braunschweig

# Assembly evaluation

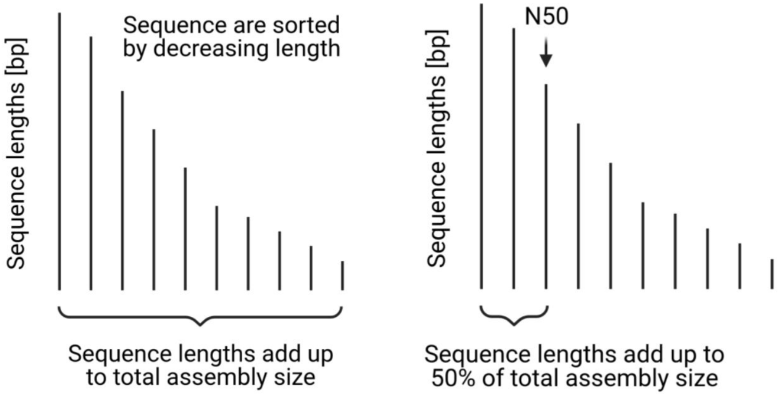
- Continuity: Does the assembly represent a genome in a small number of contigs?
- Completeness: Are all parts of the genome represented?
- Correctness: Is the assembly a correct representation of the genome?

# Evaluation: continuity

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



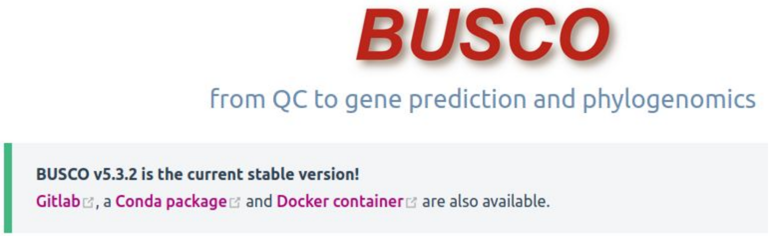
| assembler | Canu | FALCON | Miniasm | Flye |
|---|---|---|---|---|
| number of contigs | 69 | 26 | 72 | 44 |
| assembly size | 123.5 Mbp | 119.5 Mbp | 120.2 Mbp | 117 Mbp |
| maximal contig length | 15.9 Mbp | 15.9 Mbp | 14.3 Mbp | 14.9 Mbp |
| N50 | 13.4 Mbp | 9.3 Mbp | 8.6 Mbp | 10.6 Mbp |
| N90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


Sequence lengths [bp]
Sequence are sorted by decreasing length
Sequence lengths add up to total assembly size
Sequence lengths [bp]
N50
Sequence lengths add up to 50% of total assembly size

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

# Evaluation - completeness

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



BUSCO

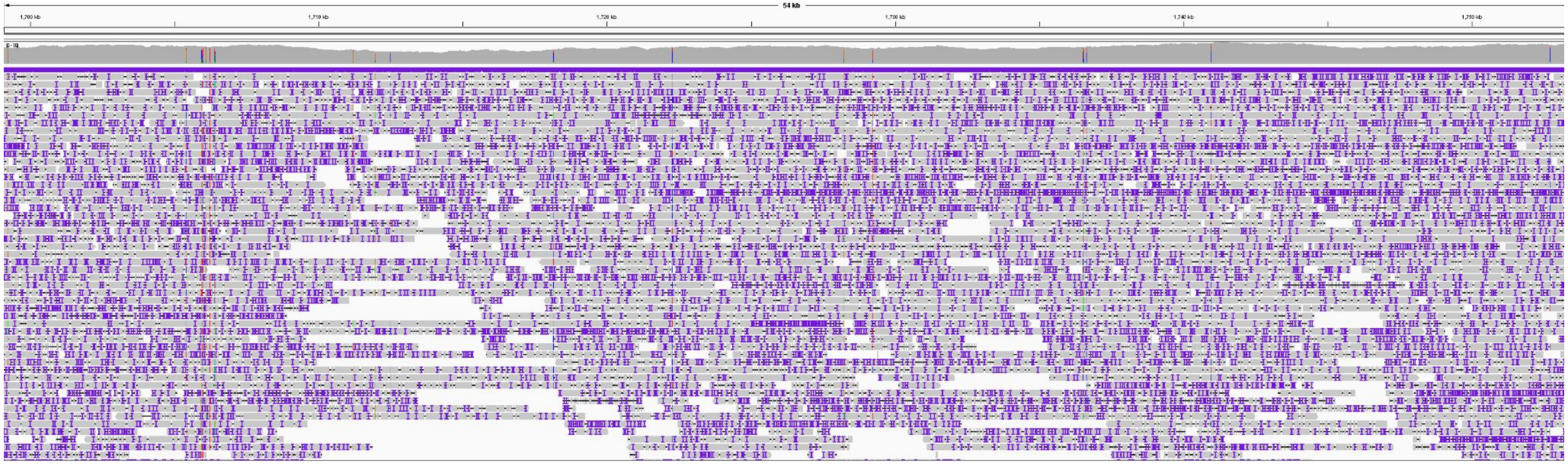
from QC to gene prediction and phylogenomics

**BUSCO v5.3.2 is the current stable version!**

Gitlab, a Conda package and Docker container are also available.

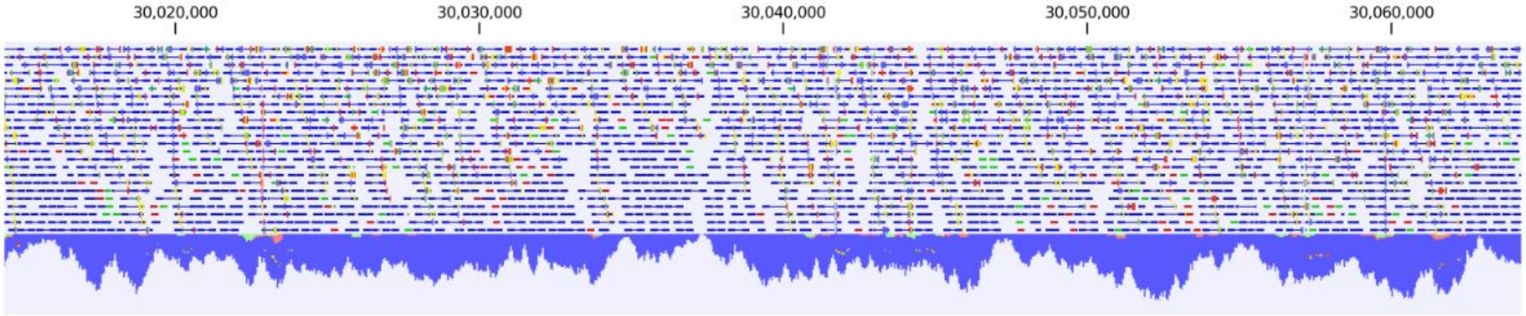
# Evaluation - assembly correctness

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



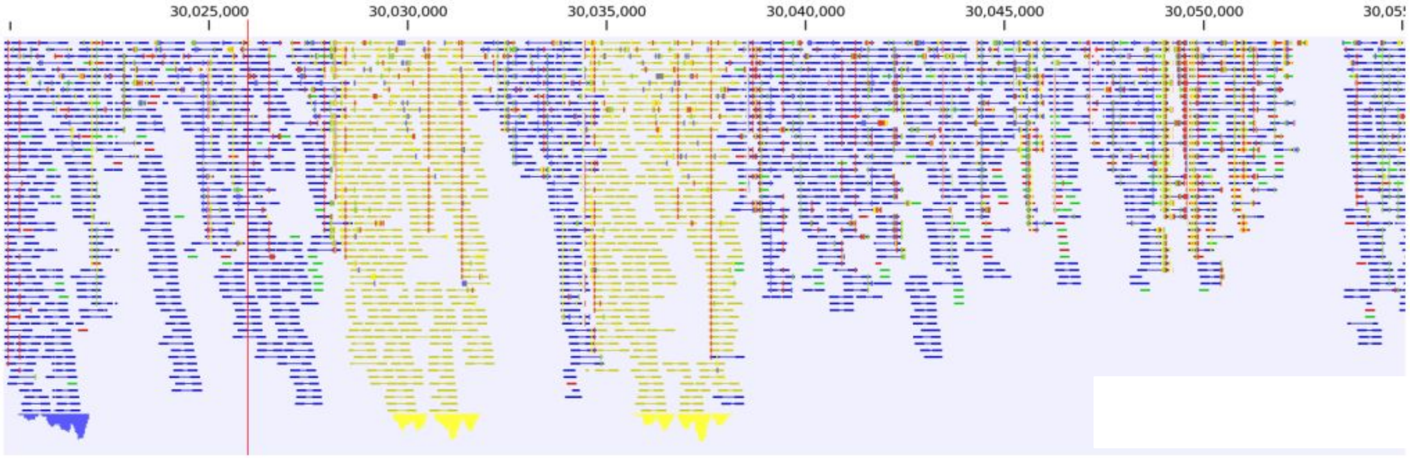
IGV: Thorvaldsdottir et al., 2013: 10.1093/bib/bbs017
REAPR: Hunt et al., 2013: 10.1186/gb-2013-14-5-r47
SQUAT: Yang et al., 2019: 10.1186/s12864-019-5445-3

# Read mappings (1)

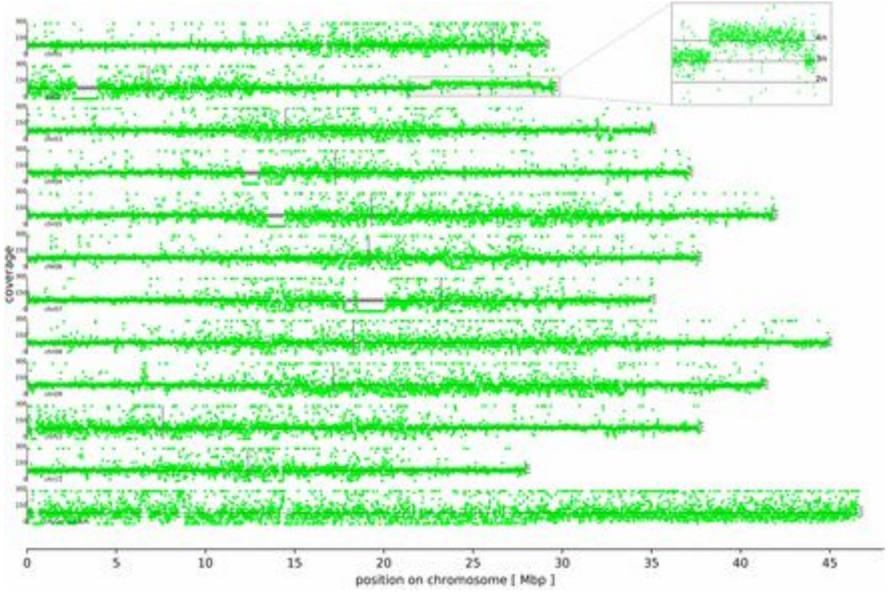


Read mapping of paired-end sequenced fragments (blue) to assembly

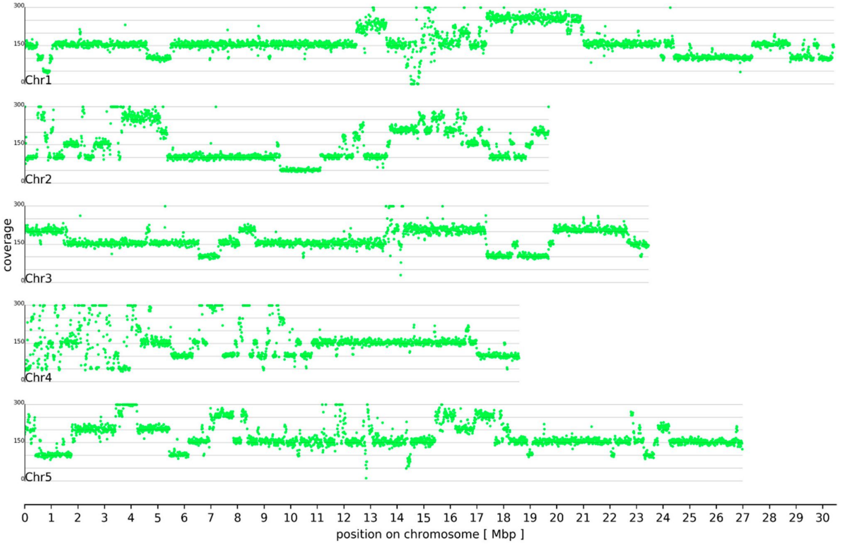
Coverage is too high to show all individual fragments at some positions



# Read mappings (2)



*Musa acuminata* (banana) read mapping



*Arabidopsis thaliana* At7 read mapping

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

# Advantages of long reads

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

# Long read assemblers

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

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

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

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

# Importance of error rate

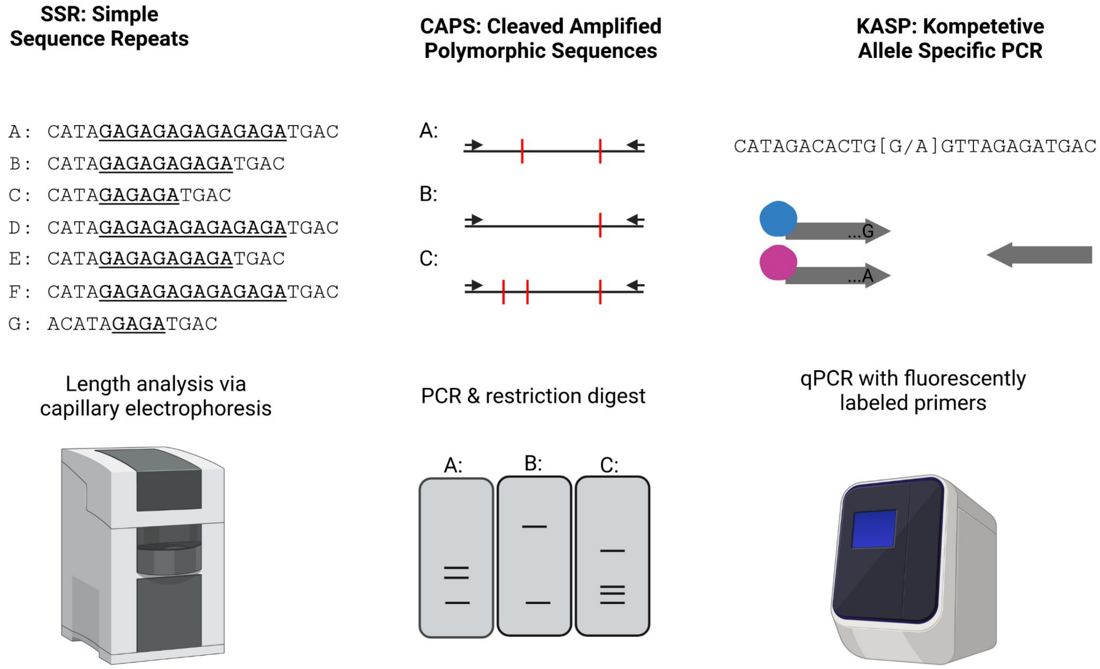
- Correction step is computationally extremely intense
    - Computation of all-vs-all alignments

- Direct assembly is possible with >99% raw read accuracy

- Higher accuracy allows to filter read overlaps more strictly

# Integration of genetic linkage information (1)

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



SSR: Simple Sequence Repeats

CAPS: Cleaved Amplified Polymorphic Sequences

KASP: Kompetetive Allele Specific PCR

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

```
CATAGACACTG[G/A]GTTAGAGATGAC
```

Length analysis via capillary electrophoresis

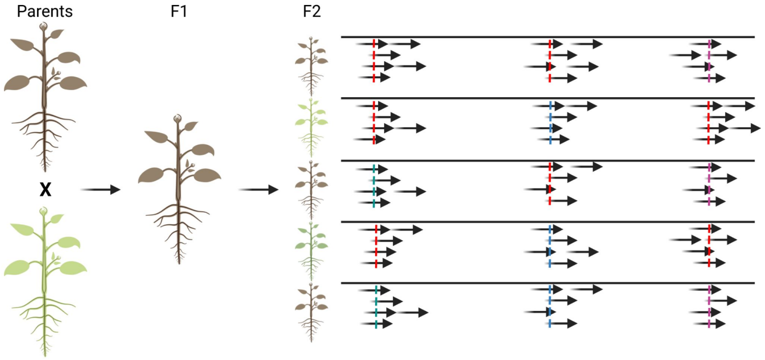
PCR & restriction digest

qPCR with fluorescently labeled primers

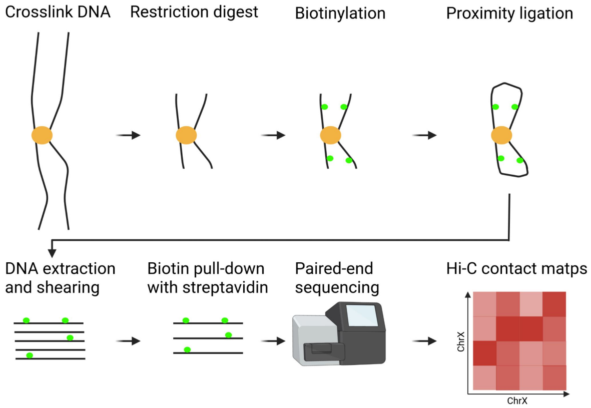
SSR: Holtgräwe et al., 2020: 10.3389/fpls.2020.00156
CAPS: Konieczny & Ausubel, 1993: 10.1046/j.1365-313x.1993.04020403.x
KASP: He et al., 2014: 10.1007/978-1-4939-0446-4_7

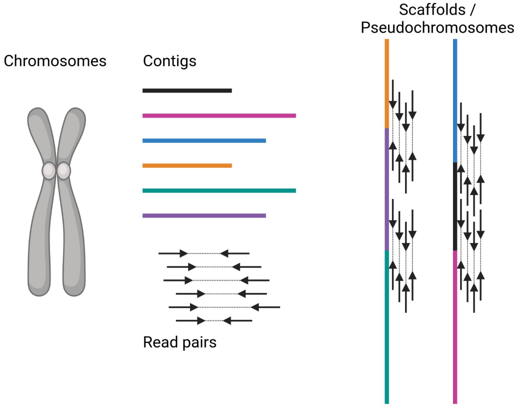
# Integration of genetic linkage information (2)

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



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



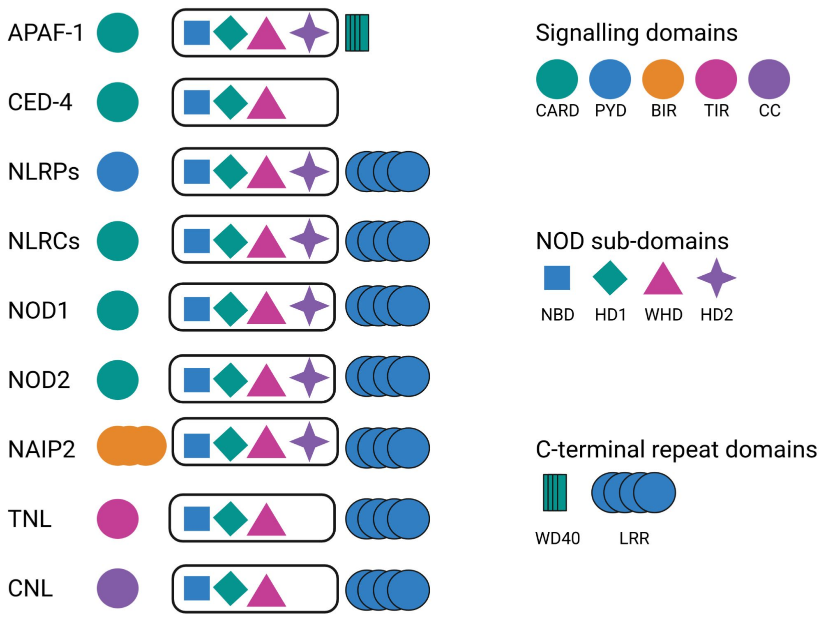


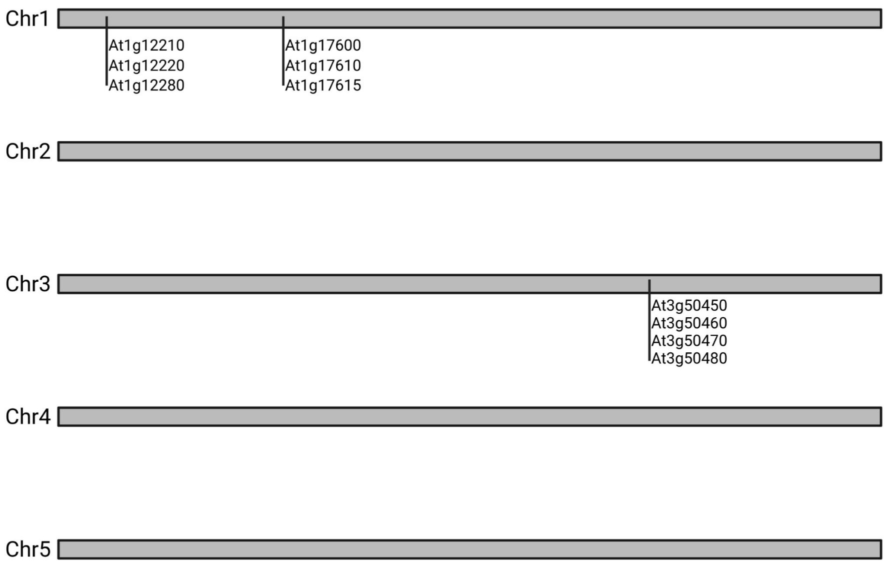
# Checking challenging regions

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

# NLRome

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





(only selected genes displayed)

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

# Time for questions!

# Questions

1. What is the difference between a contig and a scaffold?
2. What is sequencing coverage depth and coverage extent?
3. What is the N50 length?
4. Which methods can be used to assess the assembly quality?
5. What information can be used for the construction of pseudochromosomes?
6. Which aspects of the assembly can be evaluated?
7. How is the completeness of an assembly evaluated?