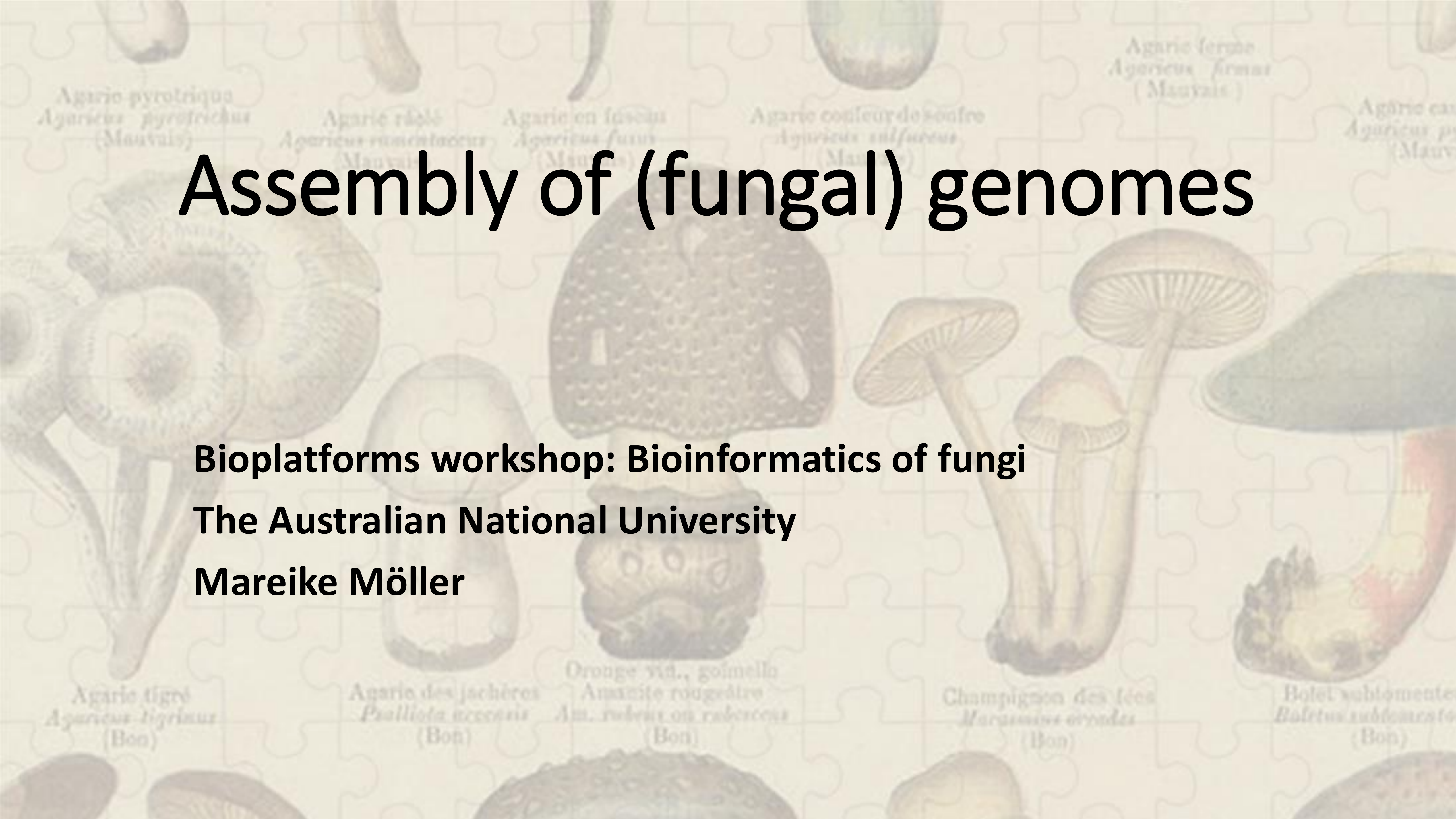
Assembly of (fungal) genomes



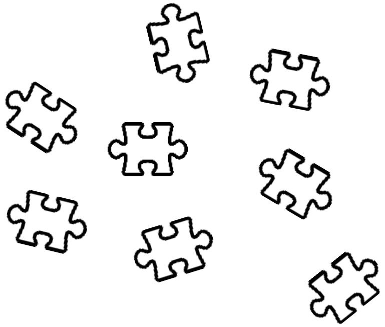
Bioplatforms workshop: Bioinformatics of fungi

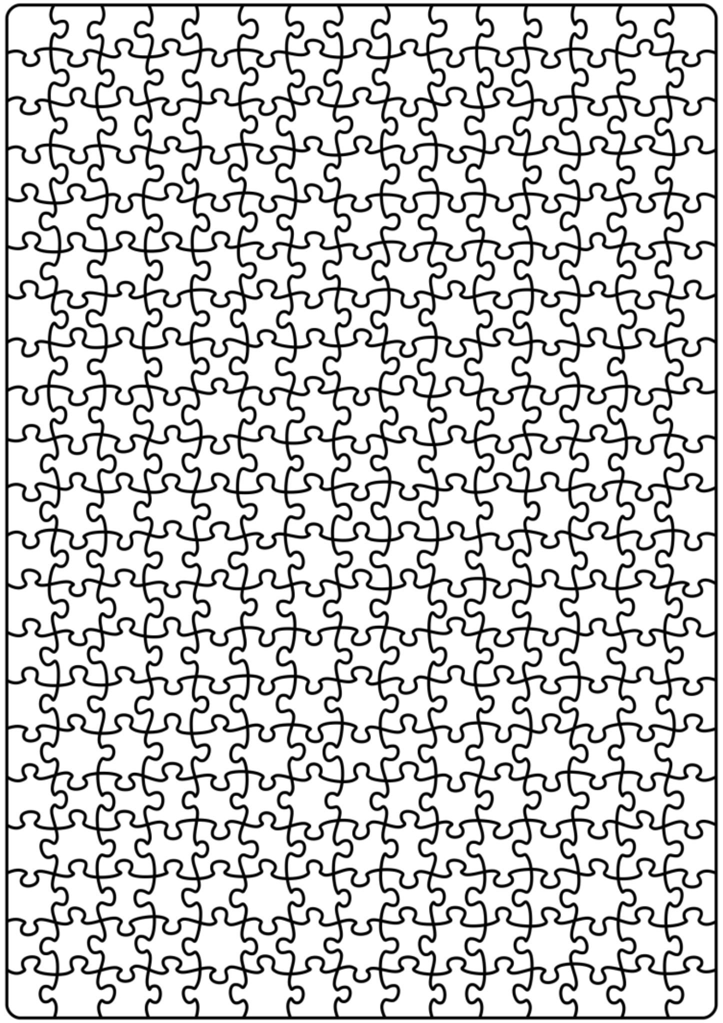
The Australian National University

Mareike Möller

What is a genome assembly?

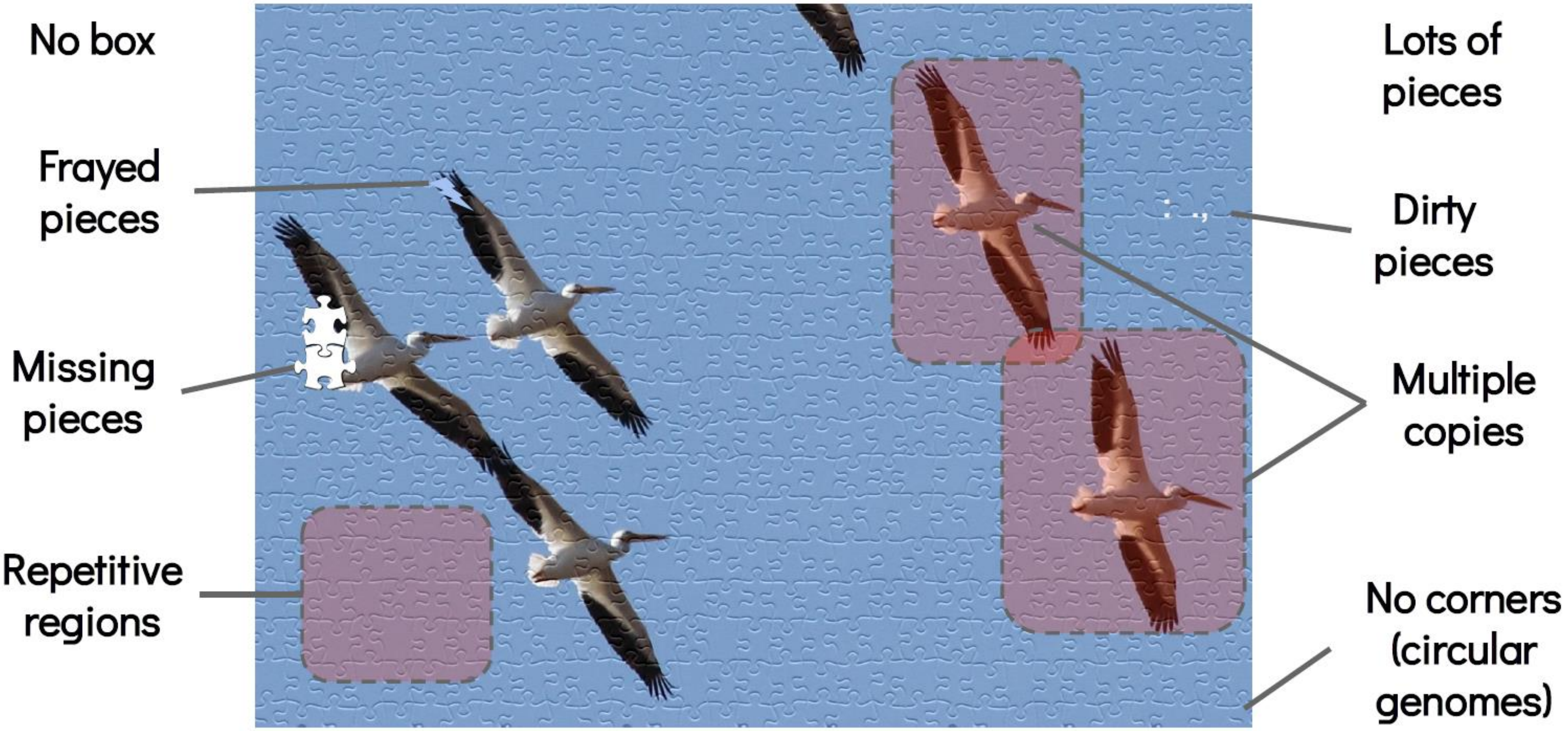
- DNA sequencing is generating pieces of the puzzle
- Assembly putting the puzzle together





Ash Jones

What makes a genome (jigsaw) puzzle hard?



Torsten Seemann (Bioinformatics Winter School)

Ash Jones

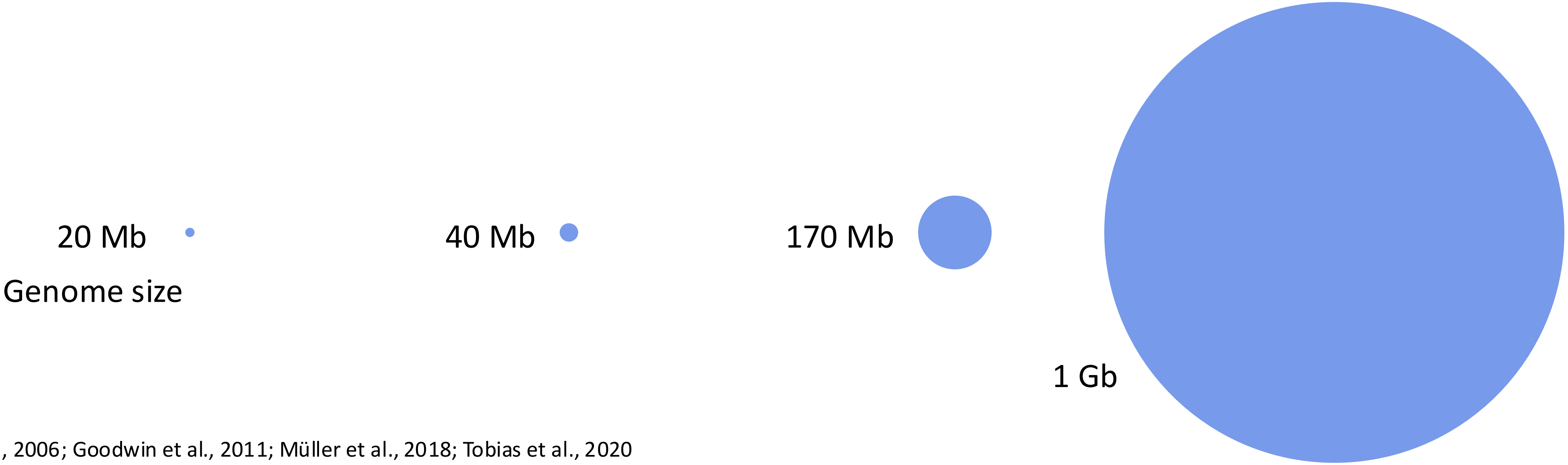
Fungal genomes are highly variable

Ustilago maydis

Zymoseptoria tritici

Blumeria graminis f. sp. *tritici*

Austropuccinia psidii



Kämper et al., 2006; Goodwin et al., 2011; Müller et al., 2018; Tobias et al., 2020

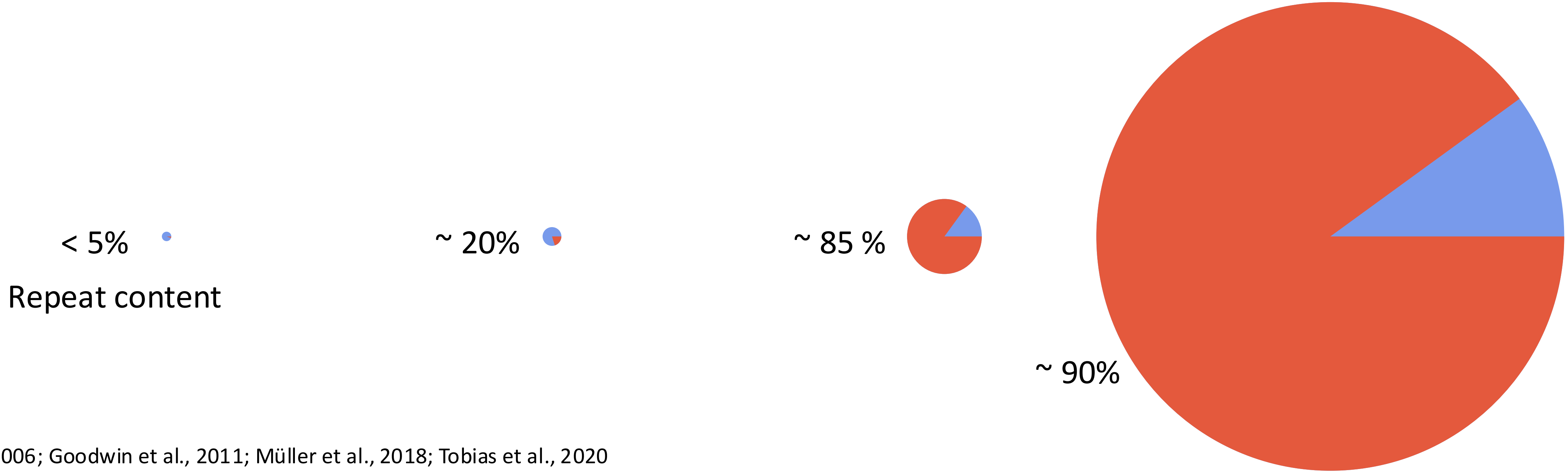
Genome size correlates with repeat content

Ustilago maydis

Zymoseptoria tritici

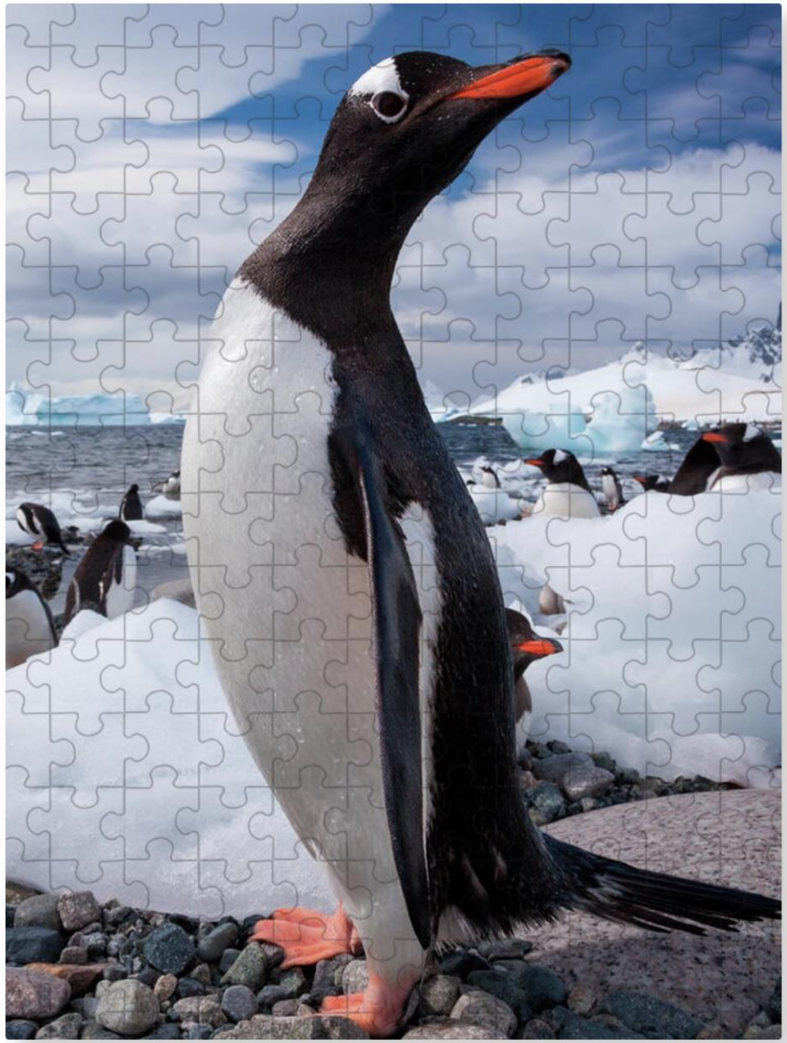
Blumeria graminis f. sp. *tritici*

Austropuccinia psidii



Kämper et al., 2006; Goodwin et al., 2011; Müller et al., 2018; Tobias et al., 2020

Challenges for assembling genomes

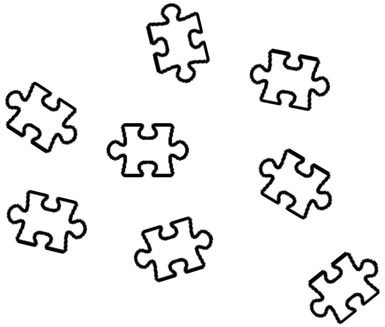


Small genome

Many unique sequences

Few repeats

Haploid

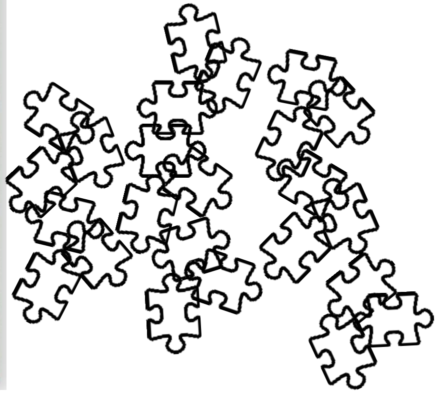


Large genome

Few unique sequences

Repeat rich

Diploid or polyploid



Why do we want a genome assembly?

- **Get assembly that provides sufficient information to address research question**

- Studying genome structure and repeats? – complete genome assembly

- Interested in gene content? – genome assembly that contains protein coding regions

What is important for a (good) genome assembly?

- Plan your sequencing project based on research question and properties of sequenced genome

- What is the expected genome size?
  - How many reads to get good coverage?
- What is the expected repeat content?
  - Short reads will result in highly fragmented assemblies in repeat rich genomes
- What is the expected ploidy?
  - Coverage and read length to be considered if phased assembly is the goal, possibly additional data to phase the genome

Data quality is important for a good genome assembly!

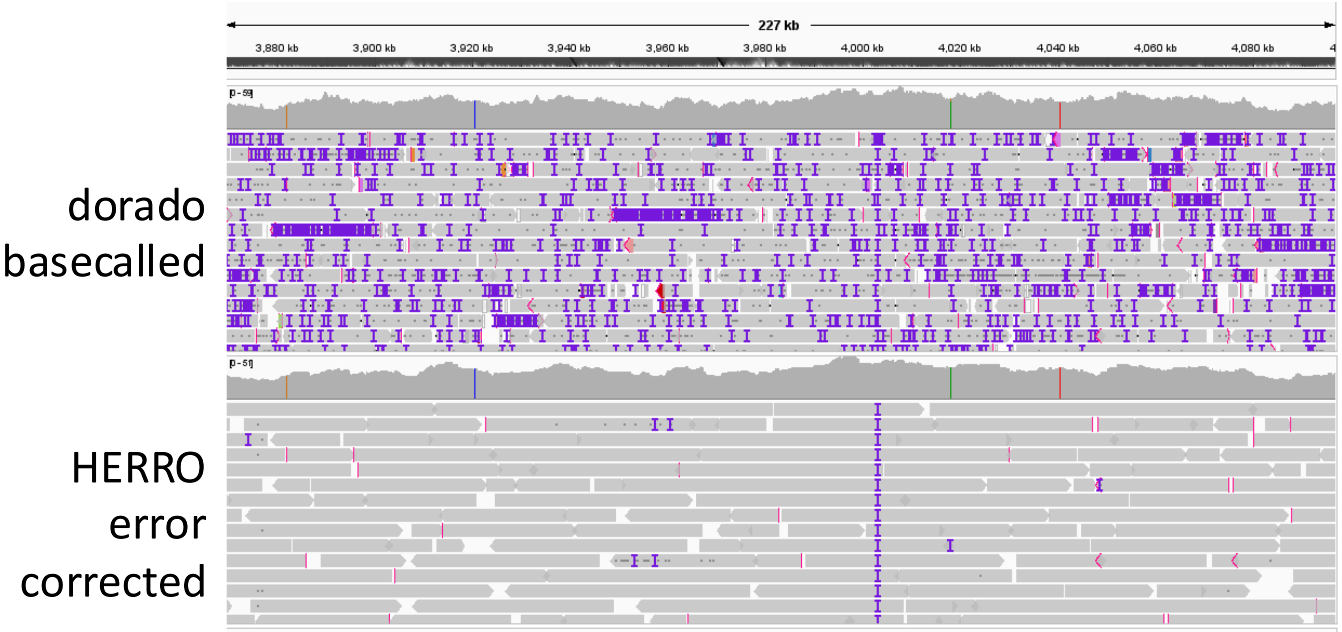
Data quality!

- High quality input DNA
- Avoid contaminants!
- Sequencing errors? Can we trust Nanopore data?

Nanopore reads can be error corrected

- Sequencing errors? Can we trust Nanopore data?

Mapping of dorado basecalled and HERRO error corrected reads to reference genome



Short vs long reads

- With short-reads: Low genome contiguity, no structure, no haplotype resolution.
- Long reads: better, however, noisy long-reads or low-molecular weight DNA challenging.

- **Moving to long-read sequencing is essential for assembling complete genomes**
- High molecular weight DNA is challenging.
- Input material may be limiting factor
- Complex sugars.
- Polyphenols.
- Low-throughput protocols

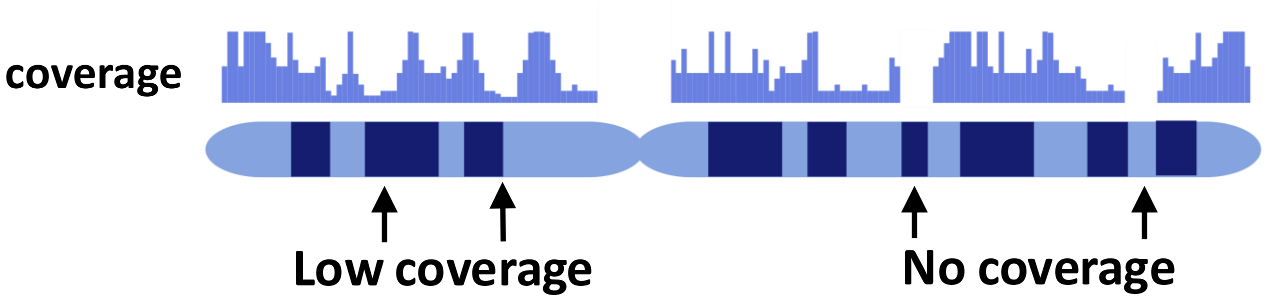
Coverage is important for complete assemblies

- Coverage

How much coverage do I need to assemble my genome?

- 30x coverage in general advisable

Short-read sequencing can miss up to 30% of the genome
(will vary by species).
Unequal coverage (sequence bias).
Complex repeats in genome can exceed 20 kb.
PCR duplicates can be problematic.

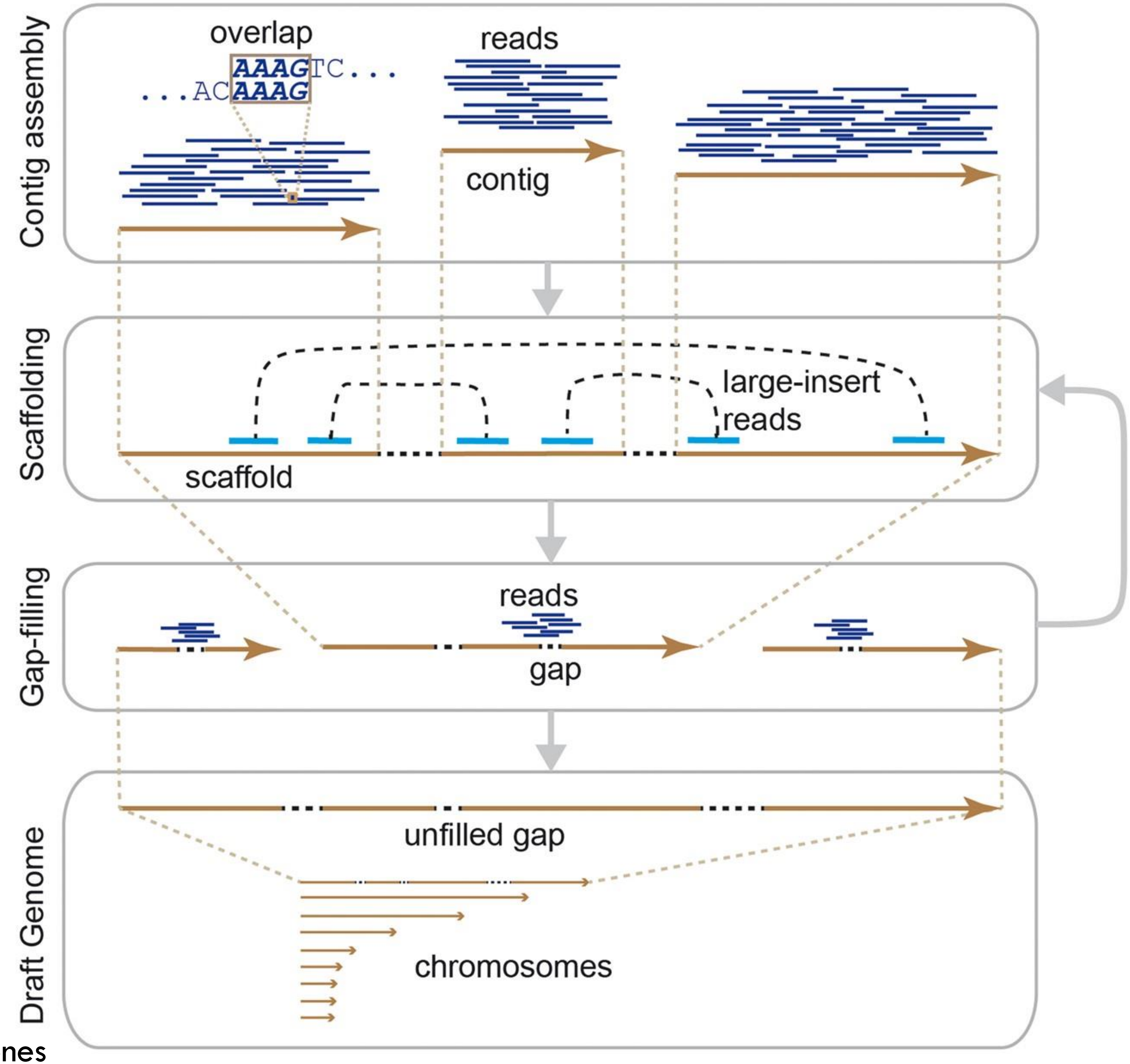


Ash Jones

What are the different genome assemblers?

- Various assemblers available
  - Short read (**SPAdes**, velvet)
  - Long-read (**hifiasm**, verkko)
  - Hybrid (flye)

- Choose best assembler based on:
  - Data type
  - Read length
  - Genome size
  - ploidy



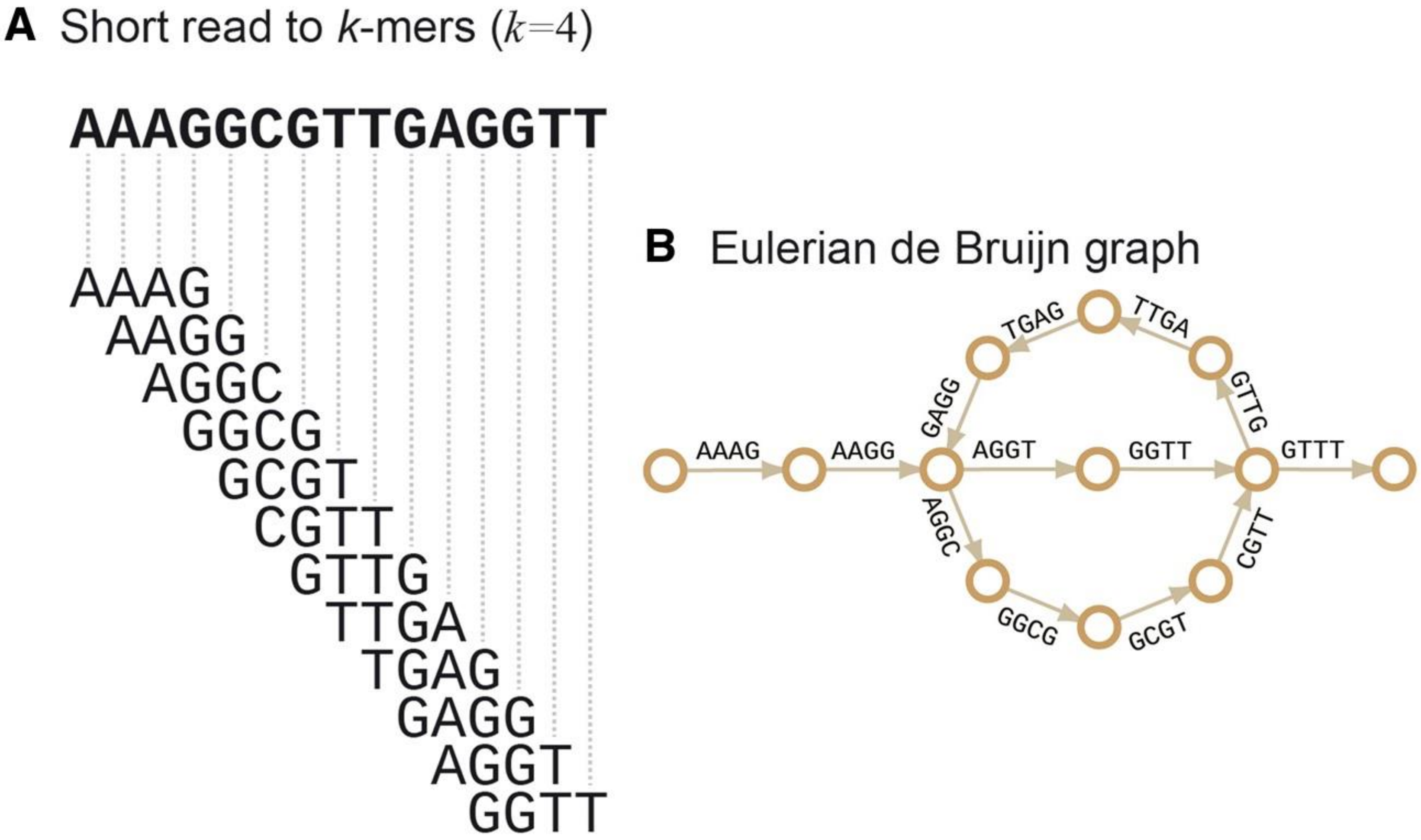
Assembly

- General workflow of the *de novo* assembly of a whole genome.
- Align and overlap sequencing reads to get bigger contiguous fragments (contigs)
- Can scaffolding by large-insert reads /contact maps if available (e.g. Hi-C).
- Gap-filling steps can be iteratively performed until no most gaps are filled.
- A draft genome consisting of scaffolds is built. Hopefully chromosomes!

ohn and Nam (2018). Briefings in Bioinformatics

Ash Jones

De Bruijn graph: Break into small manageable pieces (k-mers)



A de Bruijn graph breaks the reads into small overlapping sequences, called k-mers

Each k-mer becomes a node in the graph

An edge connects two k-mers if they overlap

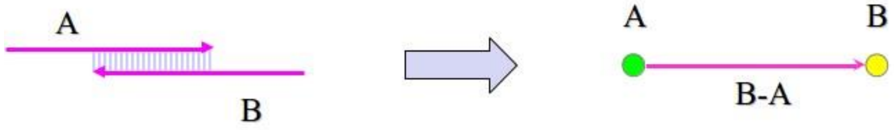
Simplifies overlapping short reads, can handle massive amounts of **short-read data**

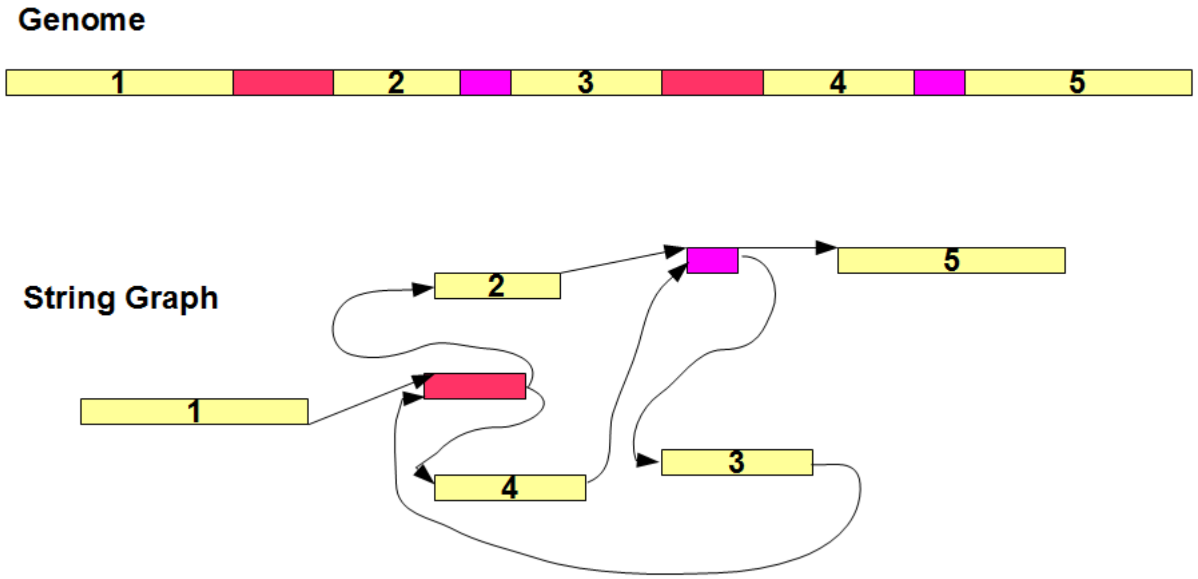
Complex repeats are challenging to assemble

Sohn and Nam (2018). Briefings in Bioinformatics

Overlapping reads using string graphs

- String graph represents whole reads as nodes and overlaps as edges
- Keeps reads intact, maintains long-range connections -> **good for long-read data**
- Can span repeats and complex regions
- Computationally intense, not ideal for short-read data

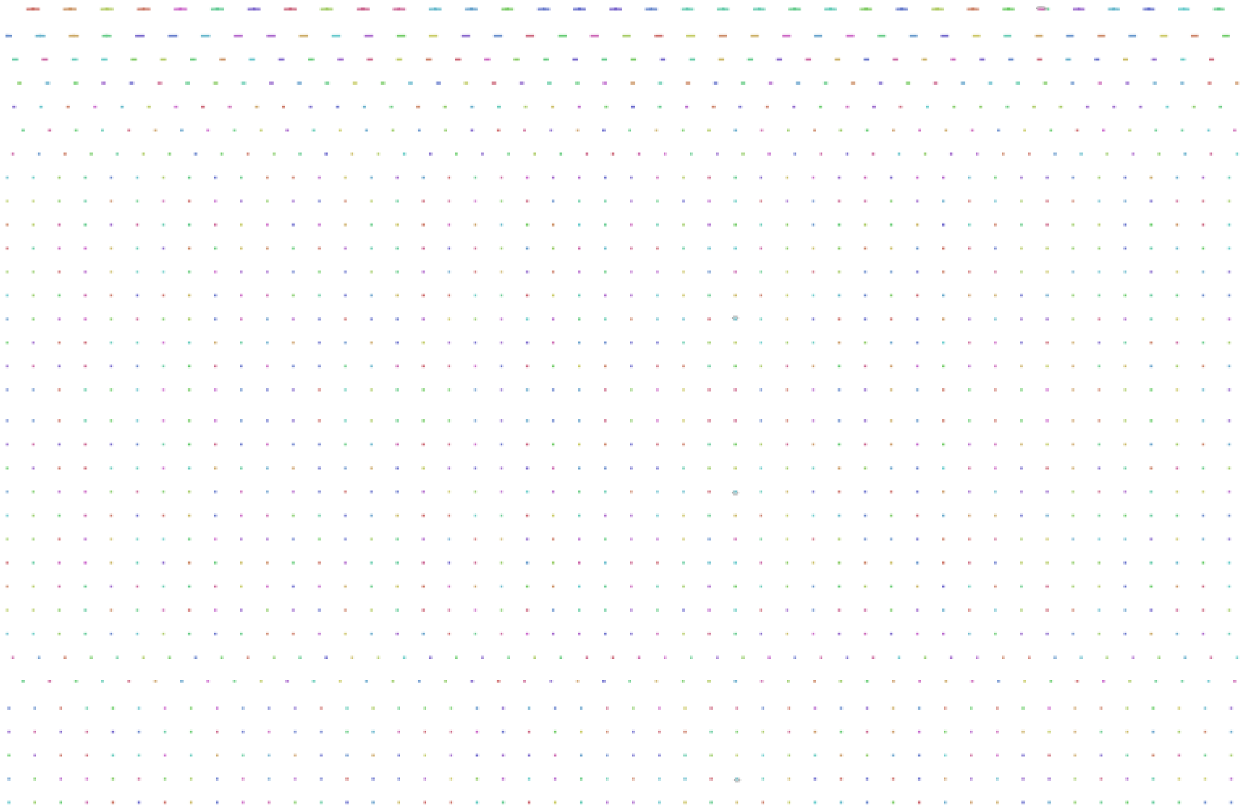




Short and long read data produce different assemblies

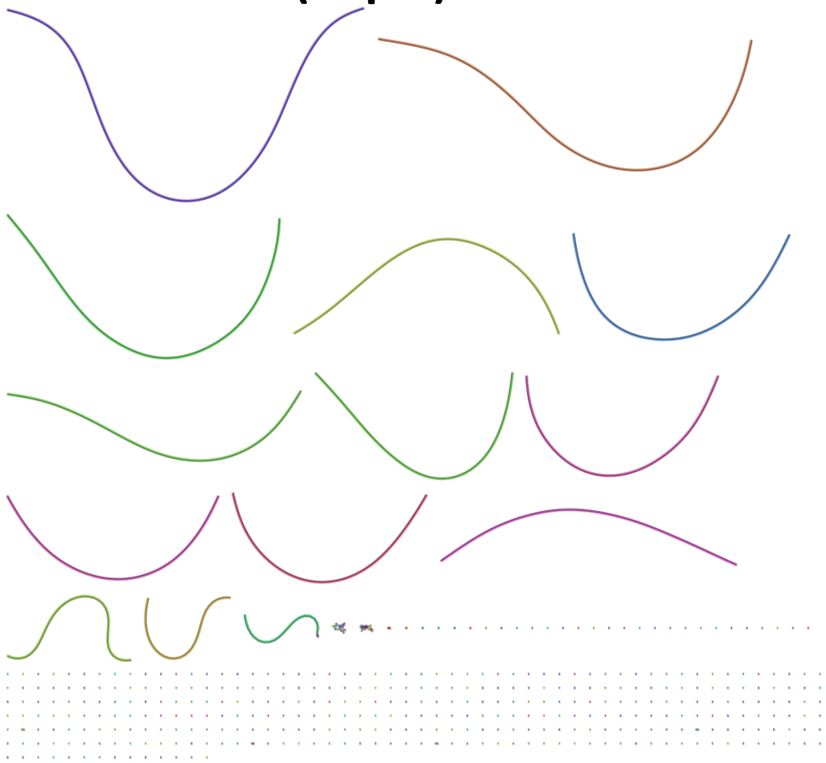
Assemblies of Eucalyptus species

Short-read assembly



Low genome contiguity, no structure, no haplotype resolution

Long-read assembly (hap 1)



Telomere to telomere haplotype resolved assembly

Ash Jones

Examples of genome assemblers

Short read: SPAdes

Long read: hifiasm

Short read assemblies with SPAdes

- **Short read**
- SPAdes - St. Petersburg genome assembler

- assembling and analyzing sequencing data from Illumina paired end sequences
- Different version optimized for transcriptome, metagenome, single cell, plasmids...

Prjibelski et al., 2020

Short read assemblies with SPAdes

SPAdes Input data

- Illumina paired-end libraries
- Illumina + PacBio (not very relevant anymore)

Parameters

- Can adjust k-mer size but automatic setting seems to be the best way to start with

Prjibelski et al., 2020

Short read assemblies with SPAdes

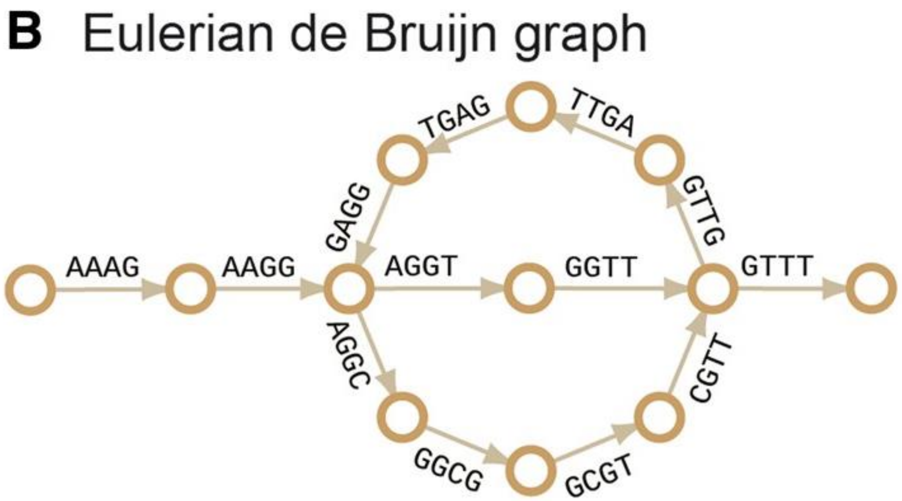
Error correction cleans up sequencing errors to produce a more accurate graph.

Building de Bruijn Graphs with Multiple k-mers instead of using just one k-mer length, SPAdes builds **several graphs** with different k-mer lengths (like 21, 33, 55, etc.), which helps SPAdes capture different levels of detail, making the assembly more accurate.

SPAdes simplifies the graph by removing errors or ambiguities. For example, it gets rid of **bubbles** (small alternative paths created by sequencing errors or slight differences) and tries to **resolve repetitive regions**, which can cause confusing loops or branches in the graph.

Contig extraction follows Eulerian paths through the cleaned graph to form assembled sequences.

Scaffolding links contigs into larger sequences using paired-end reads.



Prjibelski et al., 2020

Short read assemblies with SPAdes

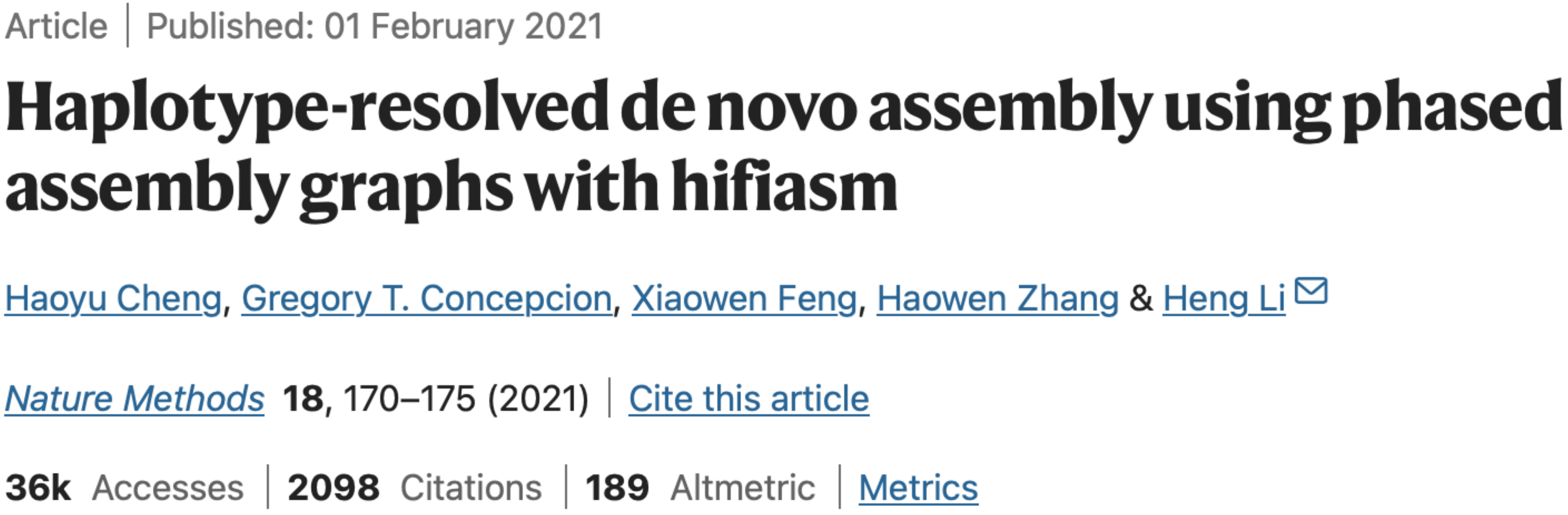
- SPAdes Output

- scaffolds.fasta - resulting scaffolds (recommended for use as resulting sequences)
- contigs.fasta - resulting contigs
- assembly_graph.gfa

Prjibelski et al., 2020

Long read assemblies with hifiasm

- Long read assembler hifiasm



Article | Published: 01 February 2021

Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm

Haoyu Cheng, Gregory T. Concepcion, Xiaowen Feng, Haowen Zhang & Heng Li ✉

Nature Methods **18**, 170–175 (2021) | Cite this article

36k Accesses | **2098** Citations | **189** Altmetric | Metrics

Long read assemblies with hifiasm

- Long read hifiasm input data

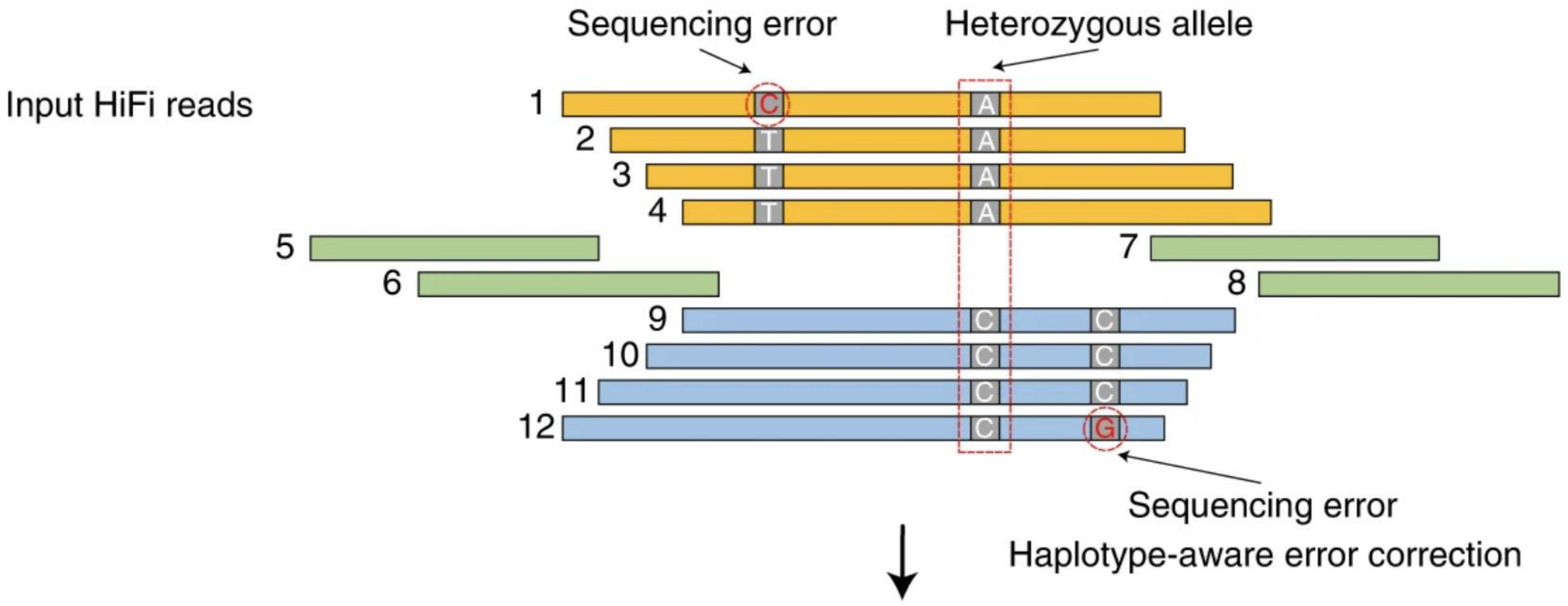
Input:

- PacBio Hifi or error corrected ONT reads (essential)
- Ultra-long ONT reads (optional)
- Hi-C reads to phase haplotypes (optional)

Cheng et al. 2021

Long read assemblies with hifiasm

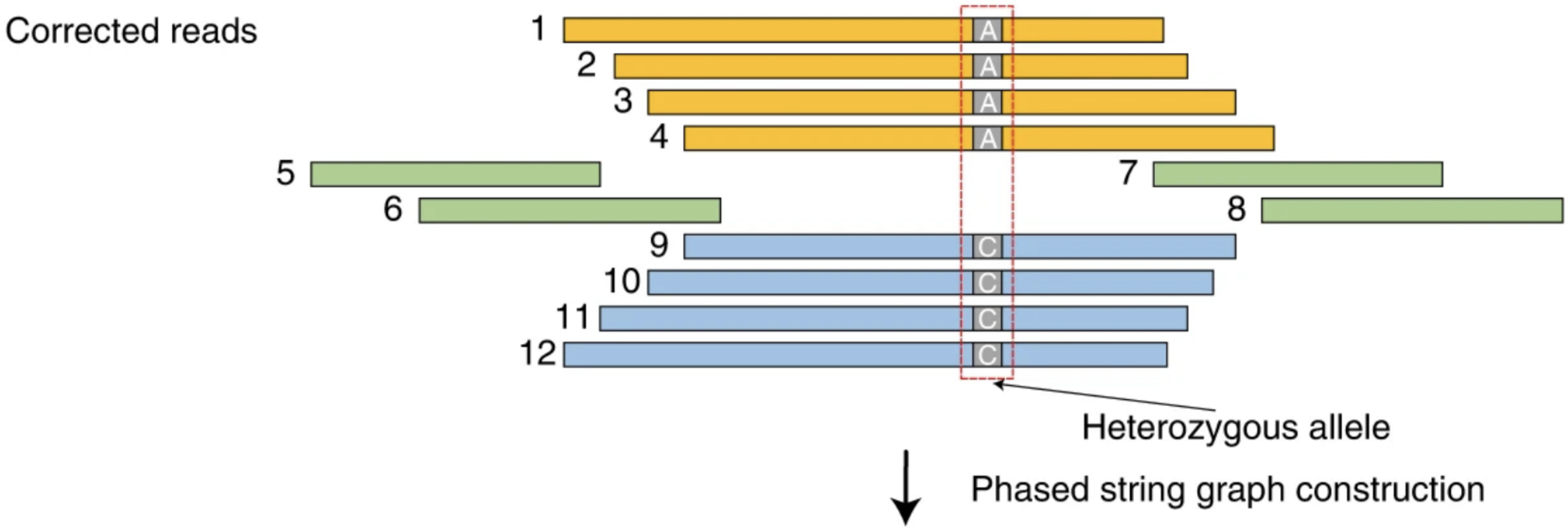
Error correction of reads



Cheng et al. 2021

Long read assemblies with hifiasm

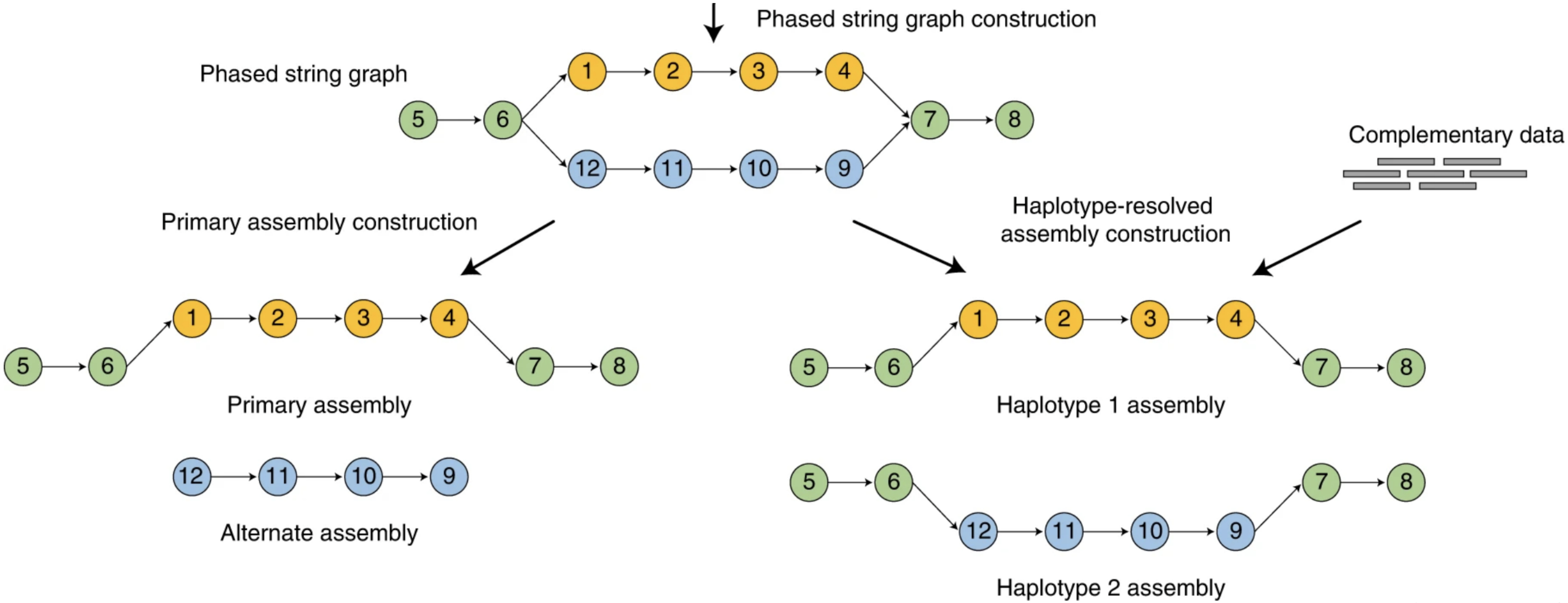
Error corrected reads as input for string graph construction



Cheng et al. 2021

Long read assemblies with hifiasm

String graphs to construct phased assembly



Cheng et al. 2021

Output of long read assemblies with hifiasm

- Outputs a variety of assembly graphs based on input data

- **Primary Assembly (e.g., *.p_ctg.gfa and *.p_ctg.fa)**
- **File format**: GFA and FASTA
- **Description**: Represent the main sequences of the assembled genome, usually corresponding to the haploid representation. In diploid organisms, these primary contigs are a consensus of the two haplotypes.
- **Usage**: This is often used as the main assembly for downstream analyses when a single representative genome is required.

Cheng et al. 2021

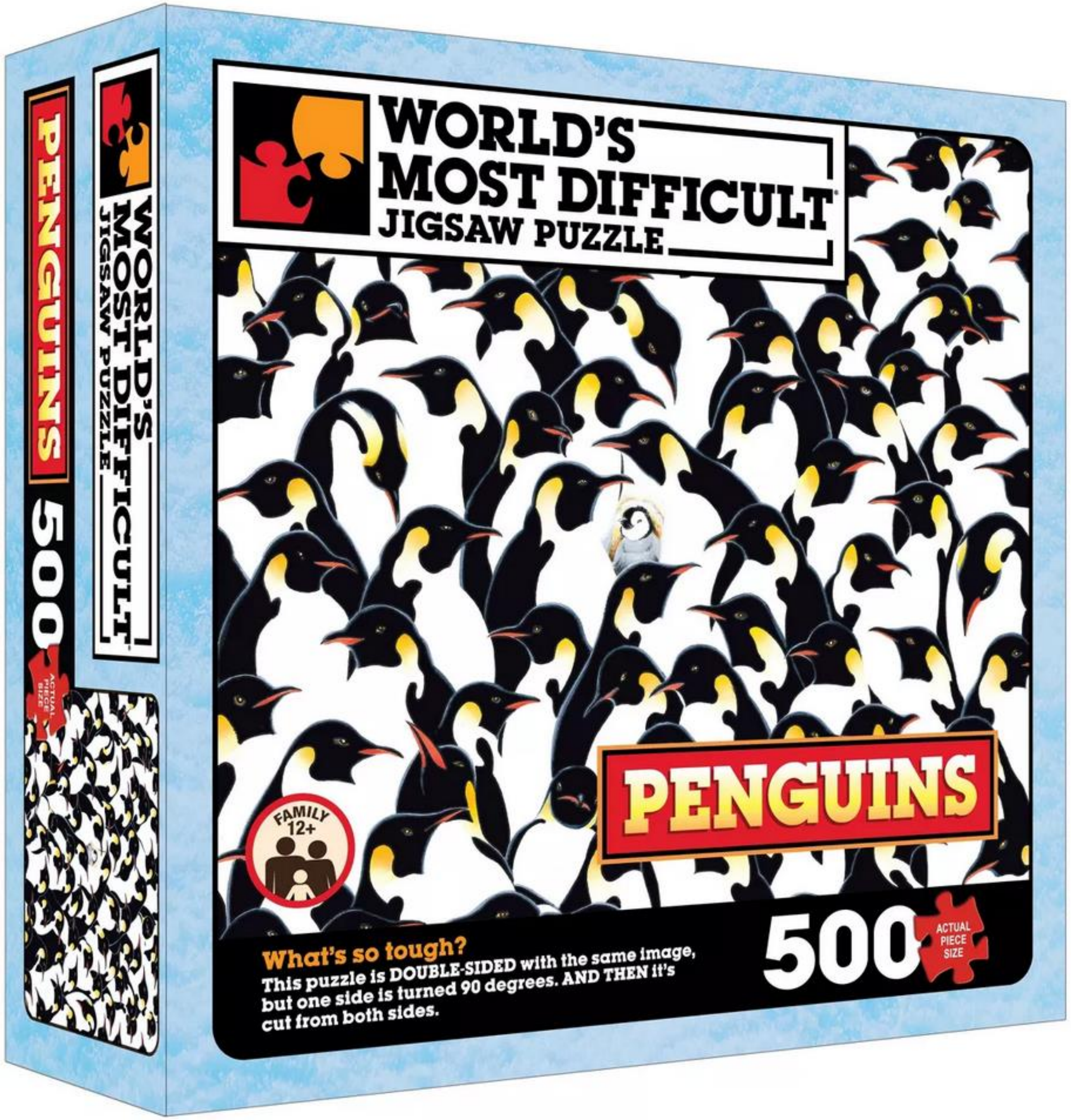
Choosing the right assembler for your project

- Based on available data and project needs

- May need to test out different assemblers

- Combine different data types to improve assembly, can consider to generate additional data to improve

Let’s assemble a genome now!!


WORLD'S
MOST DIFFICULT
JIGSAW PUZZLE
PENGUINS
FAMILY 12+
What's so tough?
This puzzle is DOUBLE-SIDED with the same image, but one side is turned 90 degrees. AND THEN it's cut from both sides.
500
ACTUAL PIECE SIZE

Genome assembly resources

Long-read assemblies:

Recent review:

Li, H., Durbin, R. Genome assembly in the telomere-to-telomere era. *Nat Rev Genet* **25**, 658–670 (2024). https://doi.org/10.1038/s41576-024-00718-w

Hifiasm:

Cheng, H., Concepcion, G.T., Feng, X. *et al.* Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021). https://doi.org/10.1038/s41592-020-01056-5

Verkko

Rautiainen, M., Nurk, S., Walenz, B.P. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* **41**, 1474–1482 (2023). https://doi.org/10.1038/s41587-023-01662-6

Flye

Kolmogorov, M., Yuan, J., Lin, Y. *et al.* Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**, 540–546 (2019). https://doi.org/10.1038/s41587-019-0072-8

Short-read assemblies:

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using SPAdes de novo assembler. *Current Protocols in Bioinformatics*, 70, e102. doi: 10.1002/cpbi.102