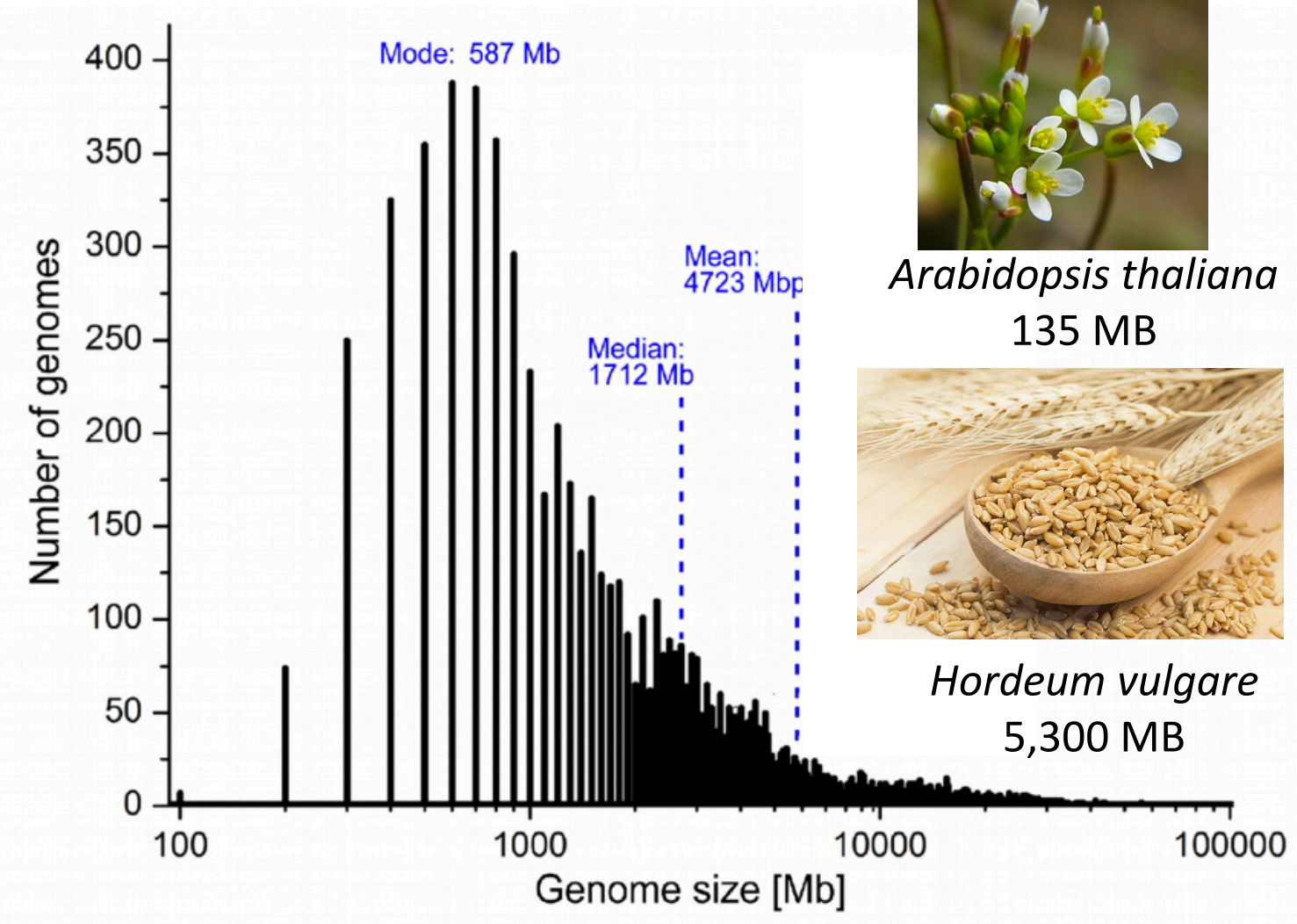
Genome assembly

Data transfer

- Before we start on the specific, please check that you have all the data you will need or follow the iget commands to transfer all relevant files for this module

Genome size of angiosperms



Adapted from Wicker et al. 2017 *Mobile DNA*

Arabidopsis thaliana
135 MB

Oryza sativa
430 MB

Zingiber officinale
1,582 MB

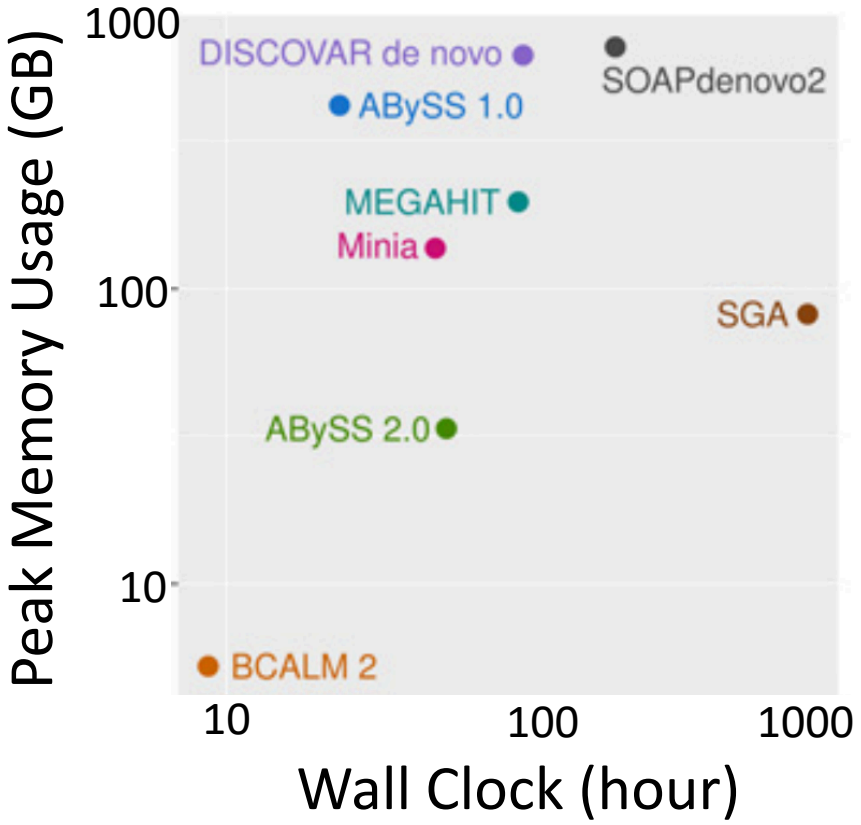
Hordeum vulgare
5,300 MB

Allium cepa
16,000 MB

Tulipa sylvestris
59,241 MB

Problems with assembling large genomes

- Many assemblers are designed for genomes equal or smaller to the human genome (3 GB)
- Larger genomes are more repetitive, with roughly the same number of genes
- Computation resources intense for deep coverage need
  - Memory and wall-time



Adapted from Jackman et al. 2017

de novo assembly

- Illumina only
  - High quality reads with fewer errors
- Hybrid option
  - Nanopore or PacBio + Illumina
  - Either raw or error corrected long-reads
- Long-read only
  - Raw typically works better
  - Need to polish with Illumina data to fix errors
- Recommendation from Li and Harkess 2018 is 100x coverage short-reads and 50x long-reads

Illumina vs Nanopore

Illumina

Reads

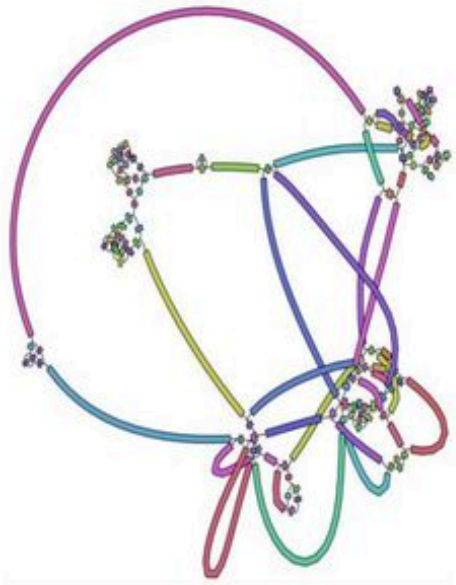
- 100–250 bp reads, 300–1000 bp fragments (shorter than repeats)
- Very accurate

Assemblies

- Fragmented
- Small N50: 10s–100s of kbp
- Very accurate

Uses

- SNPs
- Phylogenetics
- Specific alleles



MinION

Reads

- Wide length distribution, 20+ kbp common (longer than repeats)
- 90–95% accuracy

Assemblies

- Complete
- 98+% accuracy

Uses

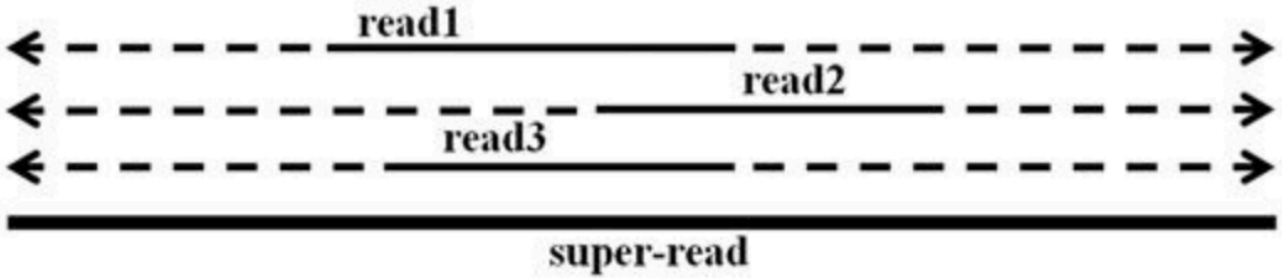
- Large-scale structure
- Horizontal gene transfer

From Louise Judd ANU Nanopore Workshop

MaSuRCA – short-read and hybrid

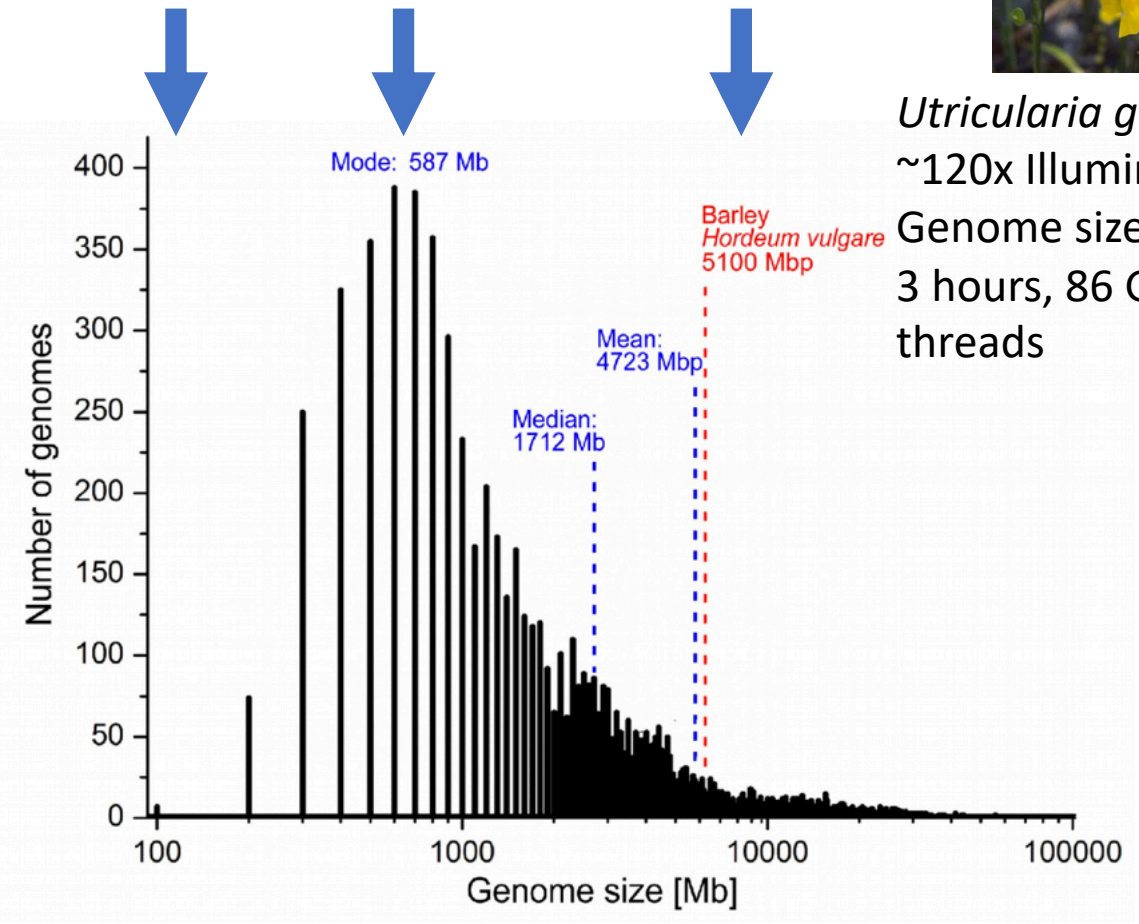
- A fast and accurate option, produces longer N50s and more BUSCO genes than other assemblers given the same data
- Combines de Bruijn graph and Overlap-Layout-Consensus

Creates ”Super-reads” from Illumina data

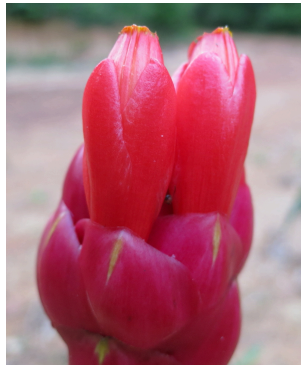


Assembler	Quast contig NGA50	Quast contig misas-semblies	NGA50 scaffold (Kb)	Scaffold misas-semblies/ MB
Allpaths-LG	28	**175**	261	**0.03**
SOAPdenovo2	8	369	1828	0.17
MaSuRCA	**56**	283	**3445**	0.19
Assemblies including some Long Read (LR) data				
MaSuRCA + 1× LR	70	256	4472	**0.04**
MaSuRCA + 2× LR	82	248	3704	0.21
MaSuRCA + 4× LR	**102**	**246**	**4511**	0.21

Examples for requirements



Utricularia gibba
~120x Illumina, 100x PacBio
Genome size:78 MB
3 hours, 86 GB storage, 16 threads

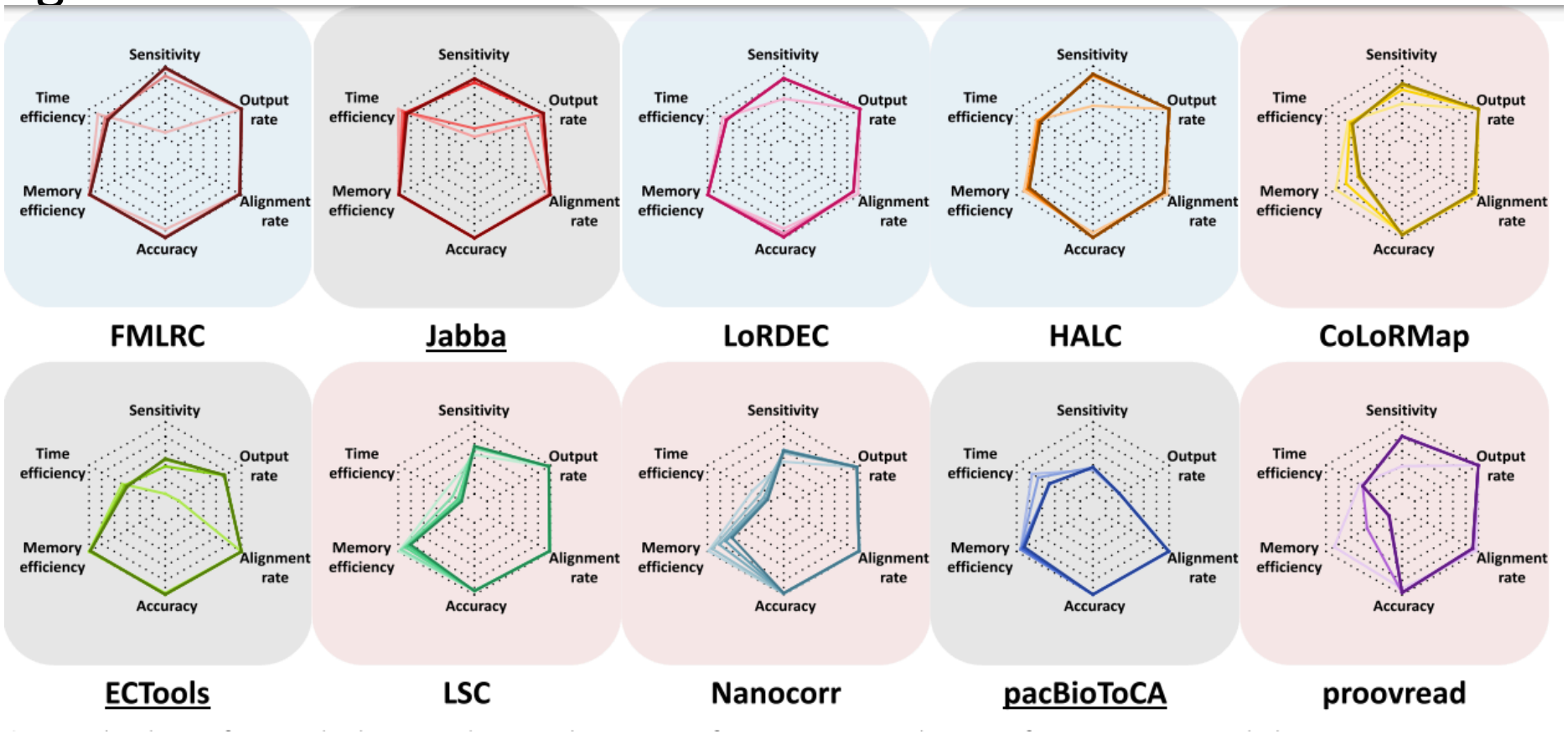


Costus spiralis
80x Illumina, 20x Nanopore
Genome size: 1 GB
3 days, 200 GB storage, 28 threads

Calochortus venustus
40x Illumina, 4x Nanopore
Genome size: 5.5 GB
22 days, 3.7 TB storage, 28 threads

Error correction

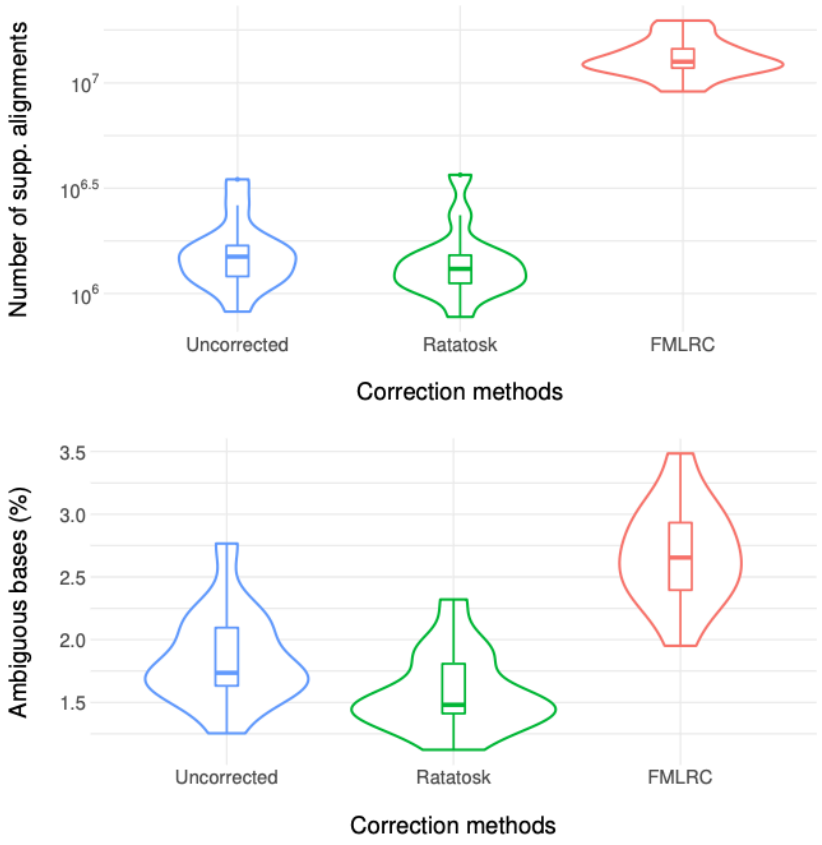
- Hybrid (using Illumina data) works better than just using depth of long-read



Fu et al., 2019
Zhang et al., 2019

New methods continuously coming out

- Ratatosk – a new program that just came out in preprint



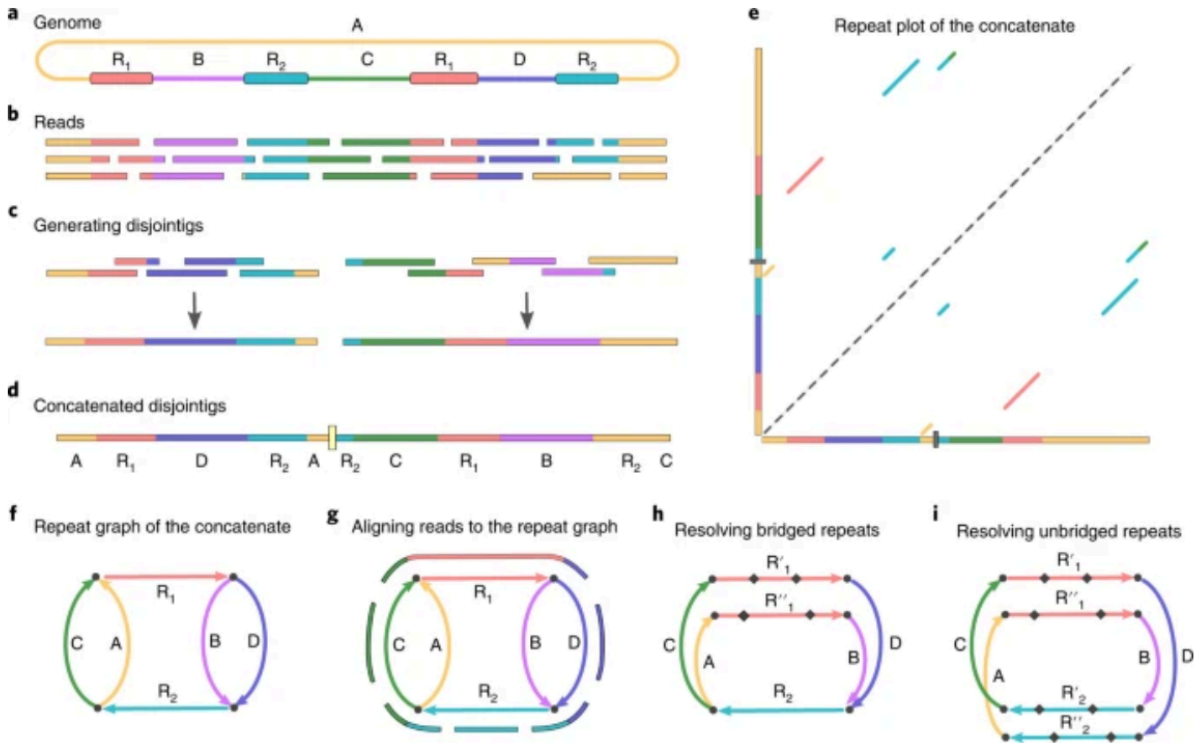
	Time (CPU h.)		Peak RAM (GB)	
	Min	Max	Min	Max
Ratatosk	11,698	22,252	213.22	305.29
FMLRC	4,228	10,828	122.69	253.49

Holley et al., BioRxiv

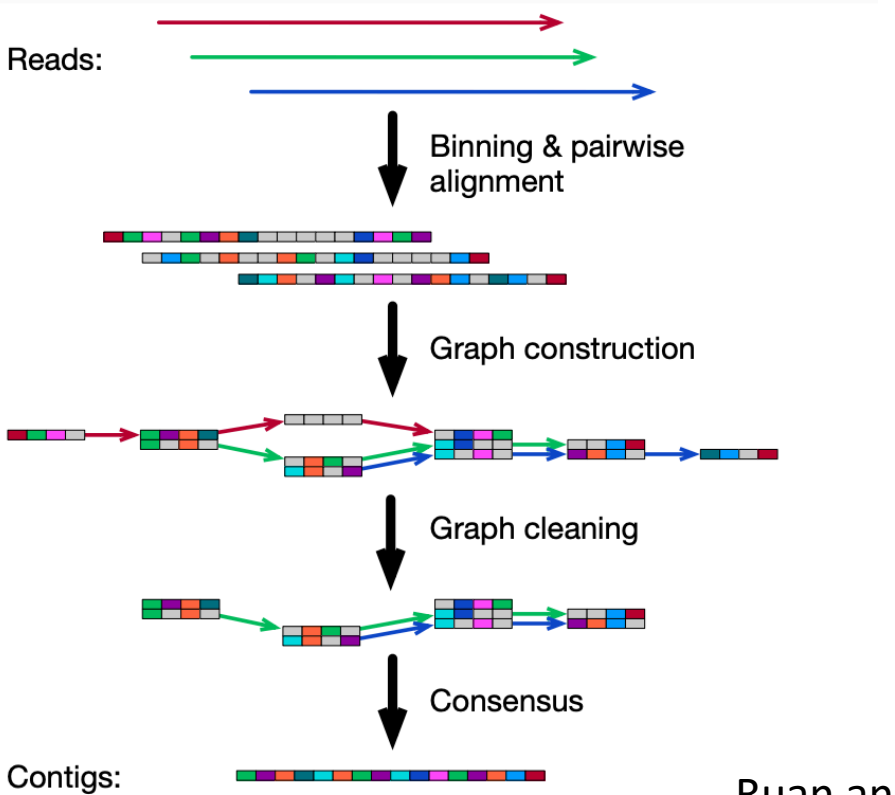
Long-read assemblers

- Many options out there; size of the genome can be make or break

Flye – repeat graphs



wtdbg2 – fuzzy-Bruijn assembly graph



Kolmogorov et al. 2019

Ruan and Li, BioRxiv

Polishing genome assemblies

- Many options for this as well; we'll use polca (part of the MaSuRCA package) to error correct the assemblies from the long-read assemblers

Genome assembly

↓ bwa mem

Map Illumina reads

↓ samtools

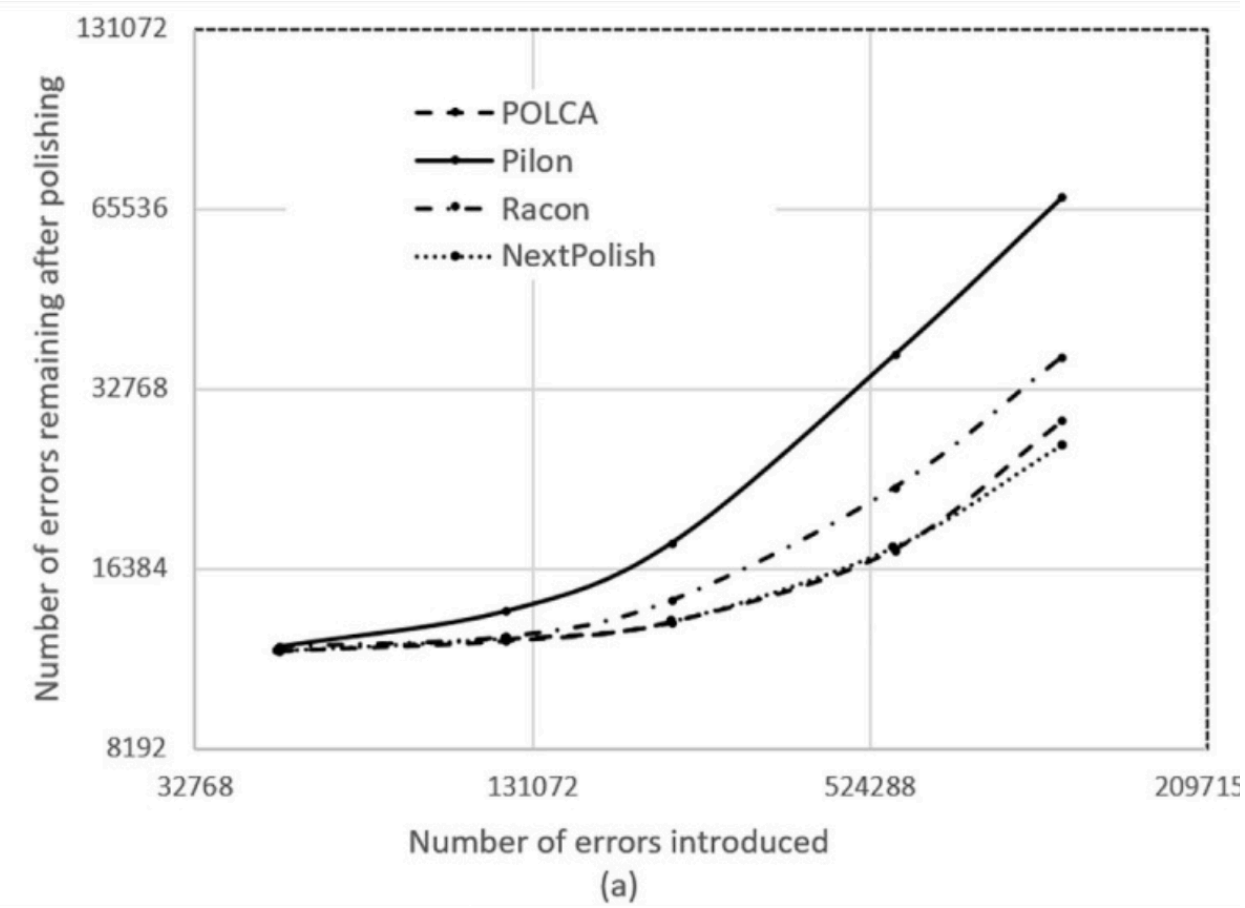
Sort by position

↓ Freebayes

Call variants

↓ polca

Replace errors with
Highest scoring alternative



BUSCO

- Benchmarking Universal Single-Copy Orthologs
- A way to judge completeness of an assembly based on the number of single copy genes expected to find through BLAST
- Some lineage specific libraries, but most will use broader categoires such as embryophyta or chlorophyta
  - Brassicales, Solanales, Poales, Fabales, or Eudicots are options