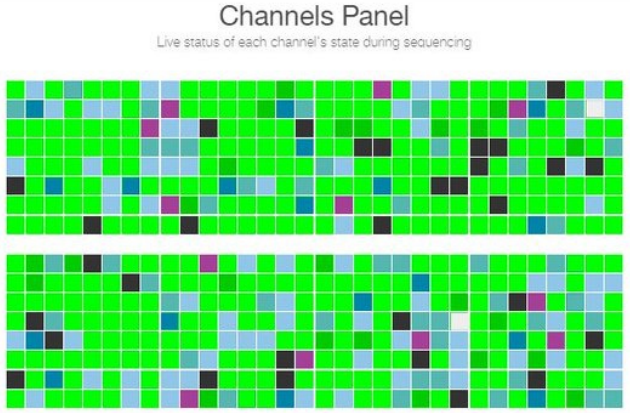
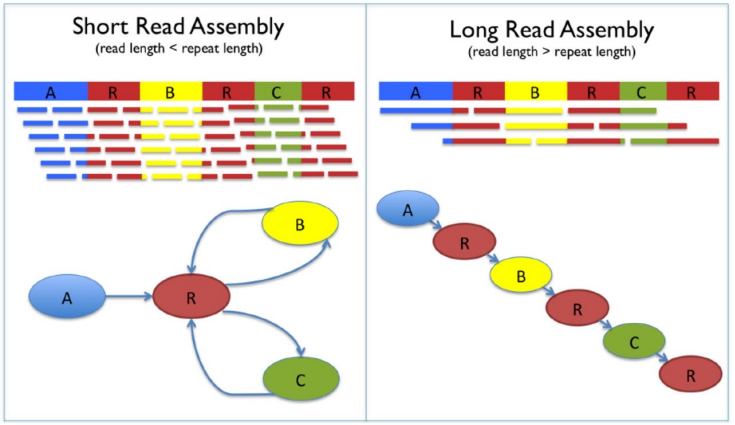


PACBIO®

de novo genome assemblies

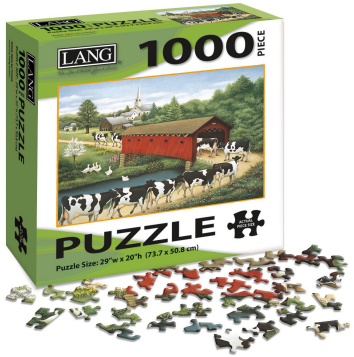
Jacob B. Landis

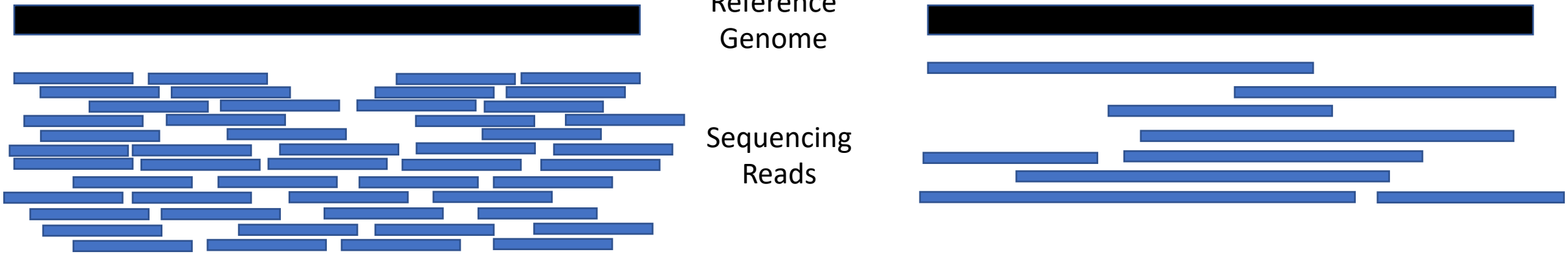
July 18th, 2021



Which puzzle is easier to put together?

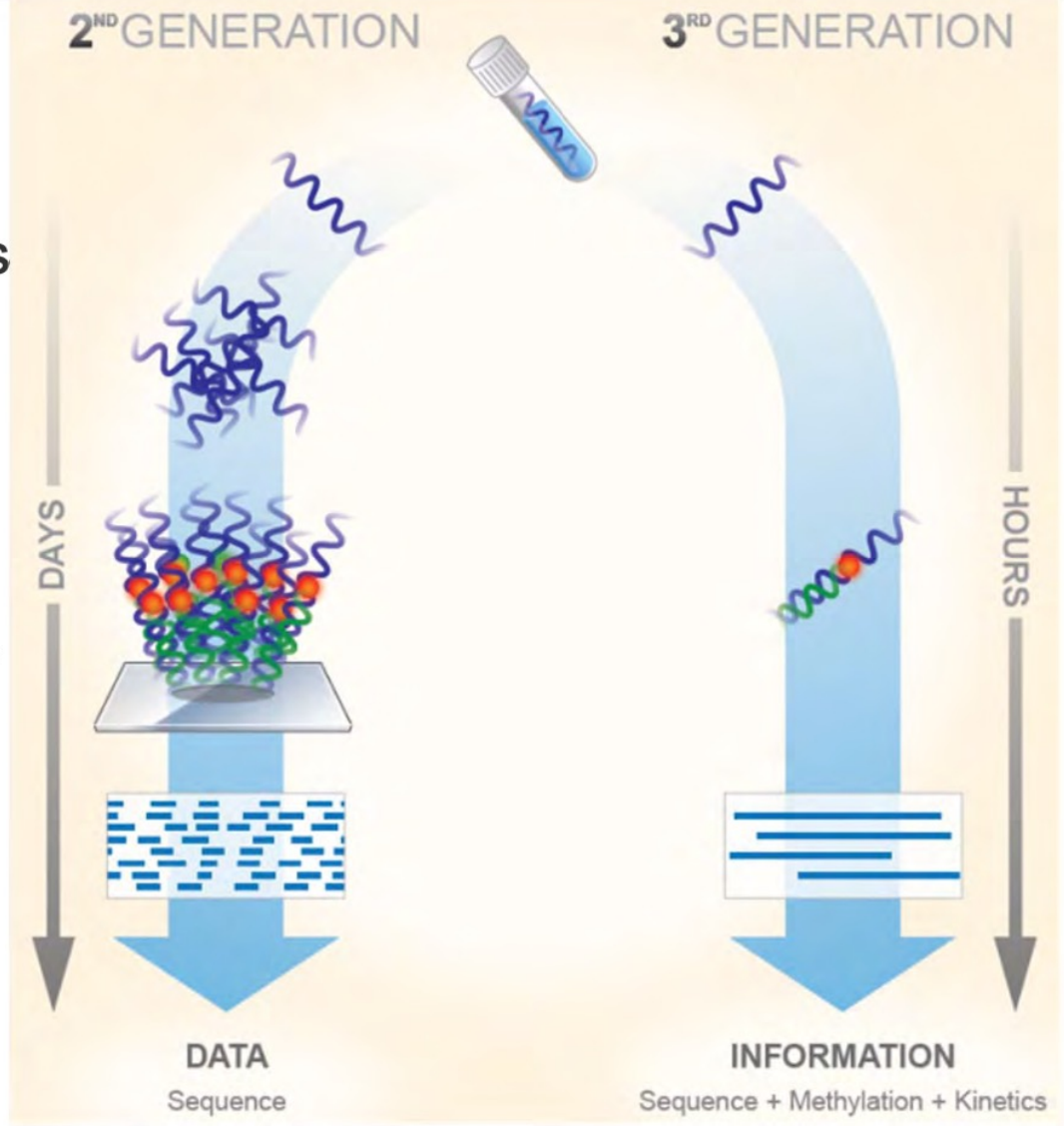
In the world of genomes





Short vs Long-reads

- Short reads
- Amplification errors and bias
- Several enzymatic steps
- Multi-molecule raw accuracy
- Errors tend to be systematic
- More coverage required



- Long reads
- No required amplification
- Simple sample prep
- Single molecule raw accuracy
- Errors tend to be random (vs. systematic)
- Less coverage required

https://get.genotoul.fr/wp-content/uploads/2017/06/150519_Technology_and_Applications_PACBIO_GerritKuhn_low.pdf

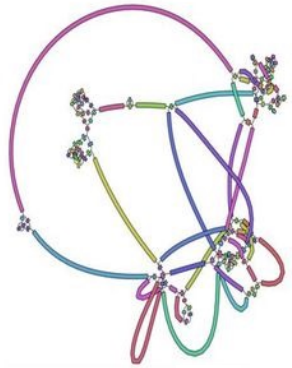
de novo assembly

- 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 after with Illumina data to fix errors
- Recommendation is 50x coverage short-reads and 50x long-reads
- So how much data do I need?

Short-read

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

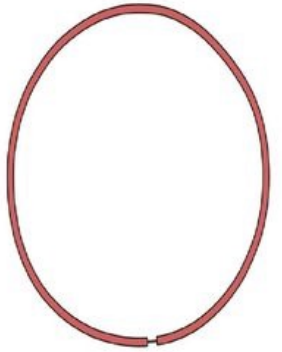
Uses
- SNPs
- Phylogenetics
- Specific alleles



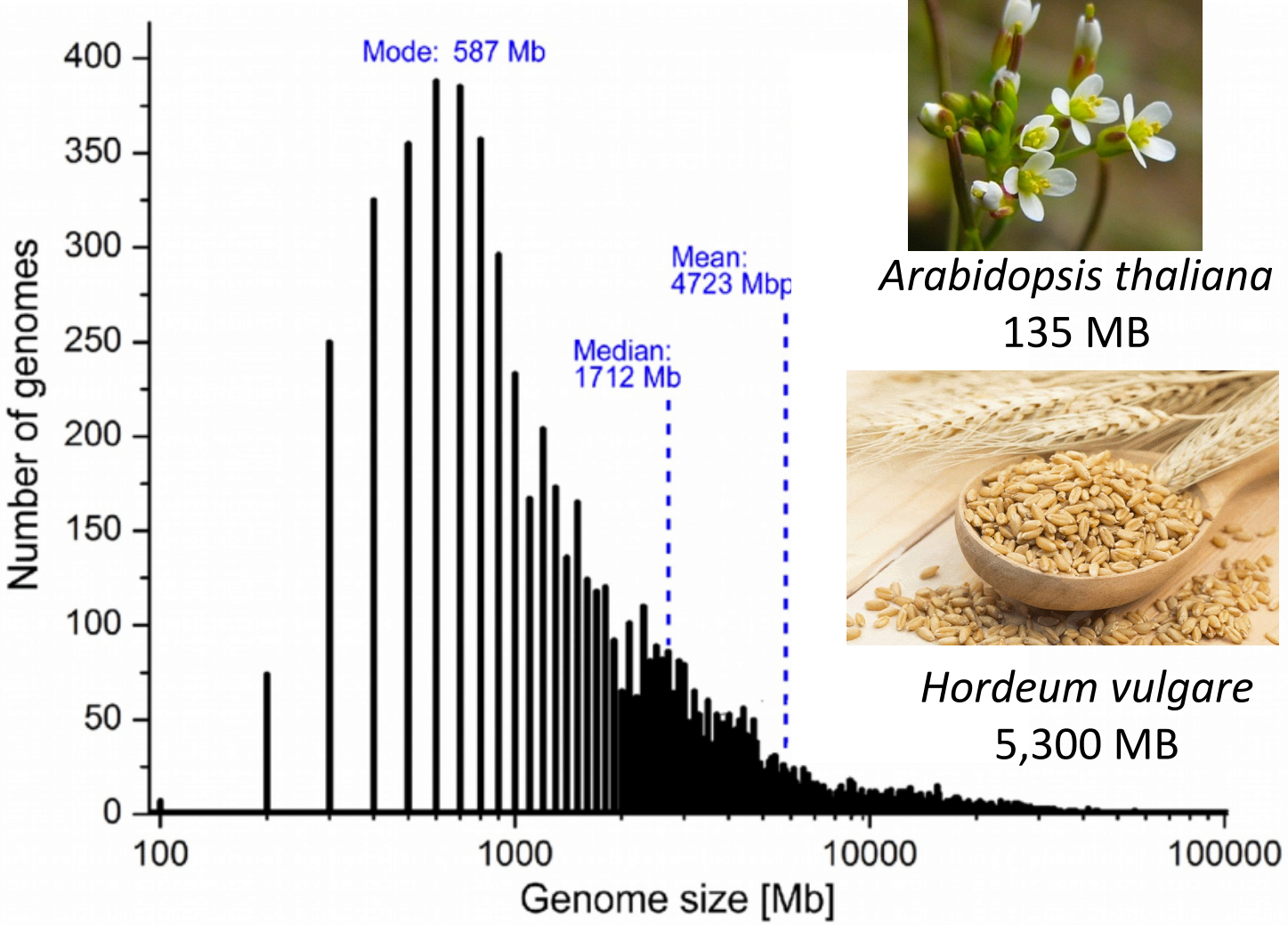
Long-read

Assemblies
- Complete
- 98+% accuracy

Uses
- Large-scale structure
- Horizontal gene transfer



Genome size of angiosperms



Arabidopsis thaliana
135 MB

Oryza sativa
430 MB

Zingiber officinale
1,582 MB

Hordeum vulgare
5,300 MB

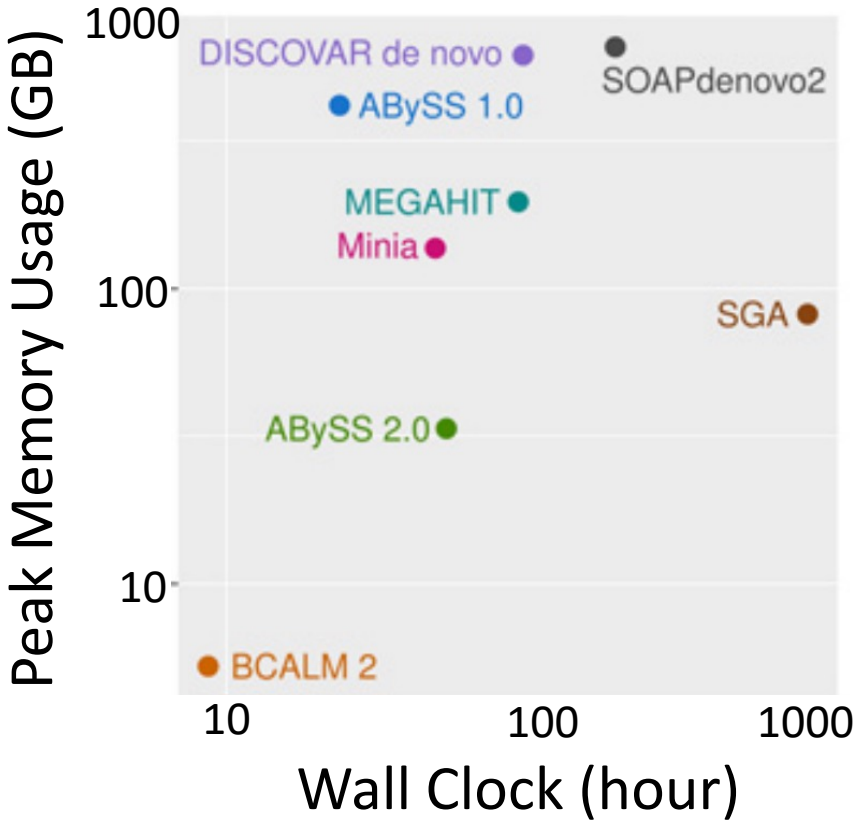
Allium cepa
16,000 MB

Tulipa sylvestris
59,241 MB

Adapted from Wicker et al. 2017 *Mobile DNA*

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

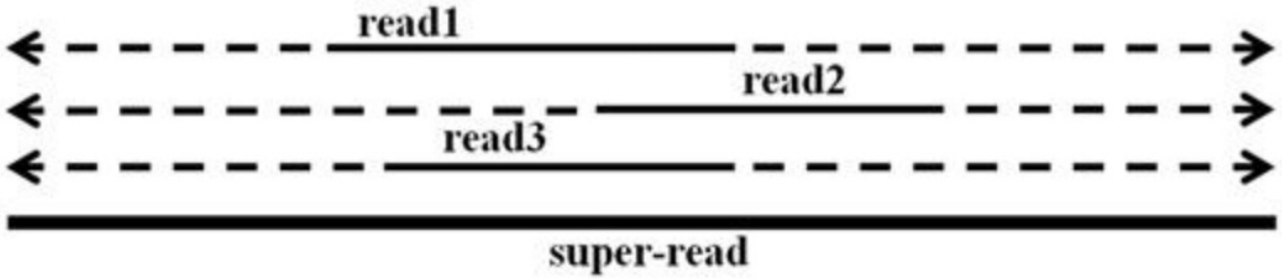


Adapted from Jackman et al. 2017

MaSuRCA – short-read and hybrid

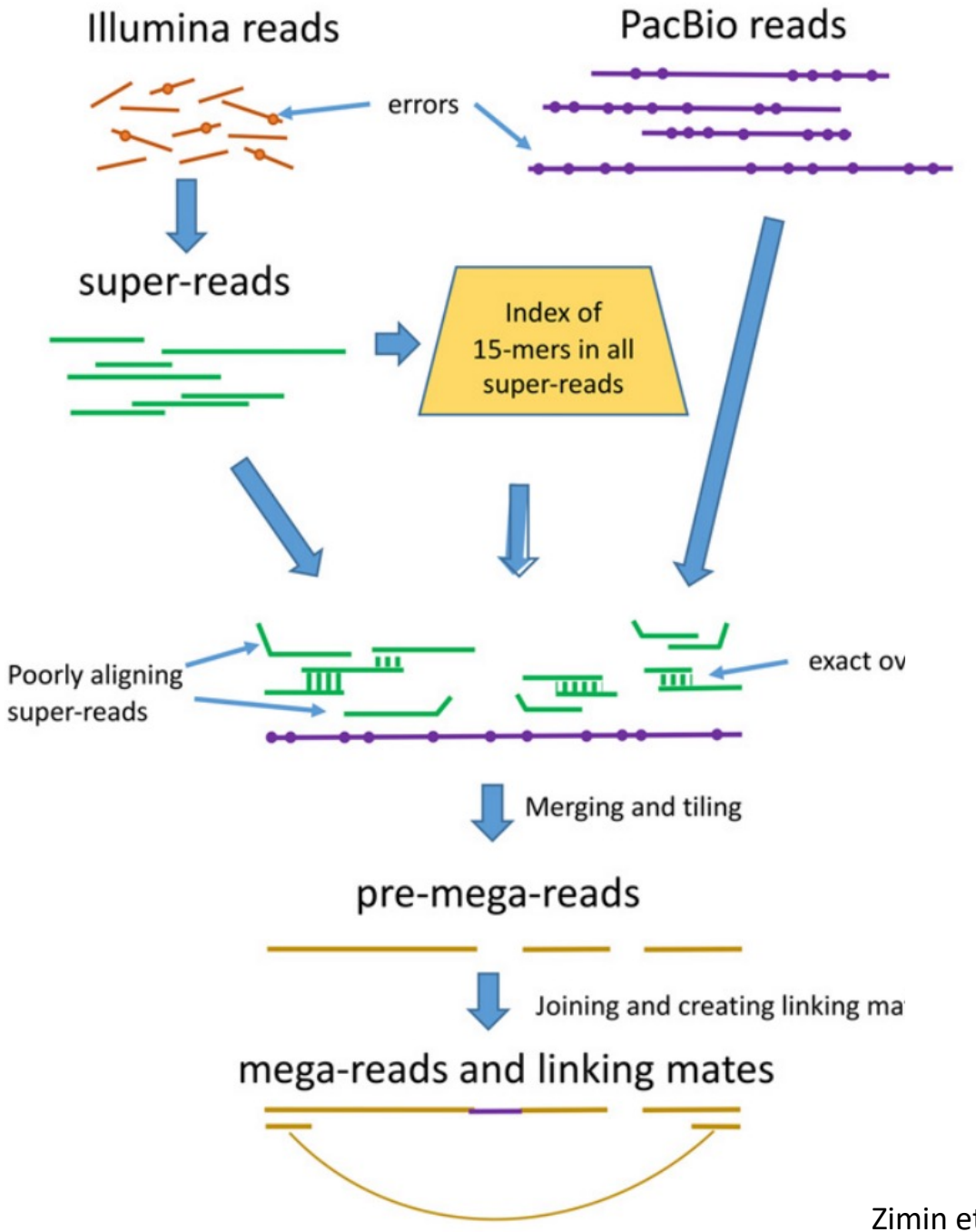
- A fast and accurate option, produces longer N50s and more BUSCO genes than other assemblers given the same data
- Combines de Bruijn graph and Overlap-Layout-Consensus

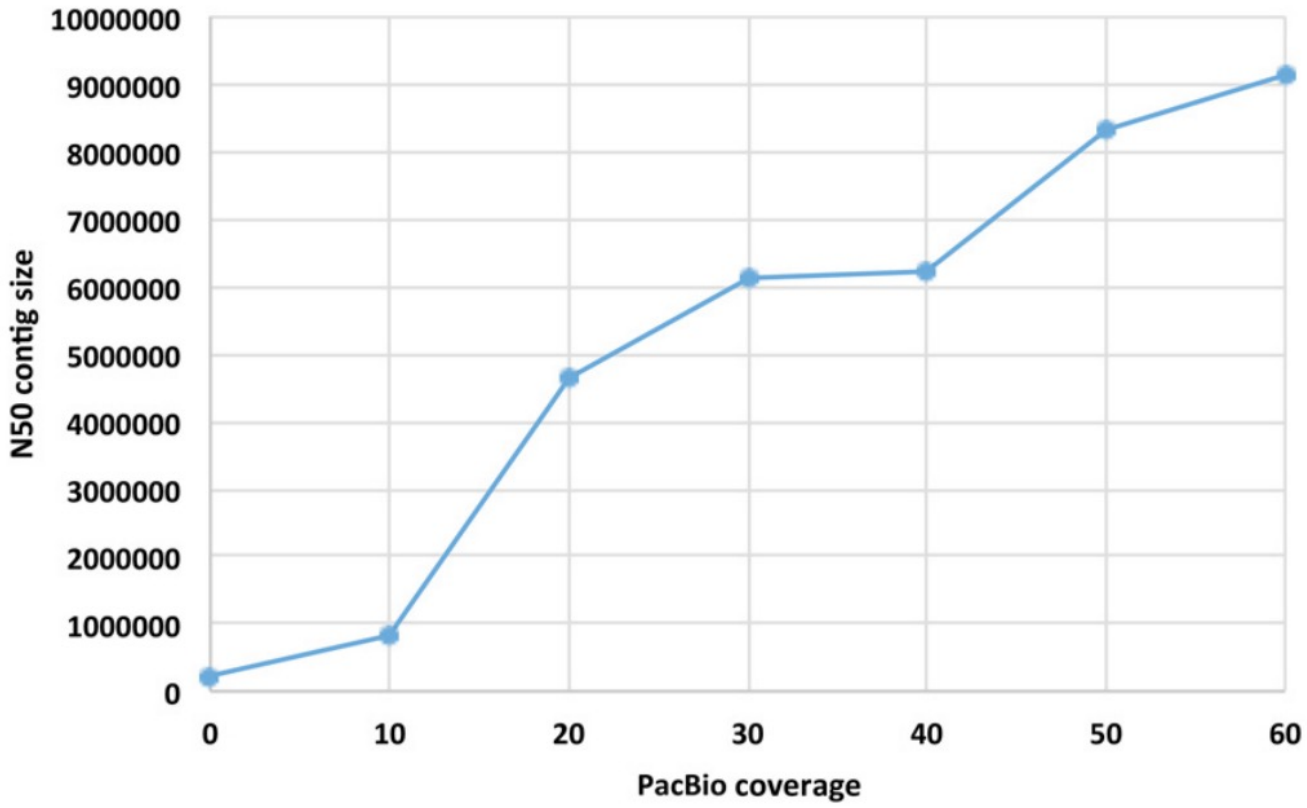
Creates ”Super-reads” from Illumina data



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

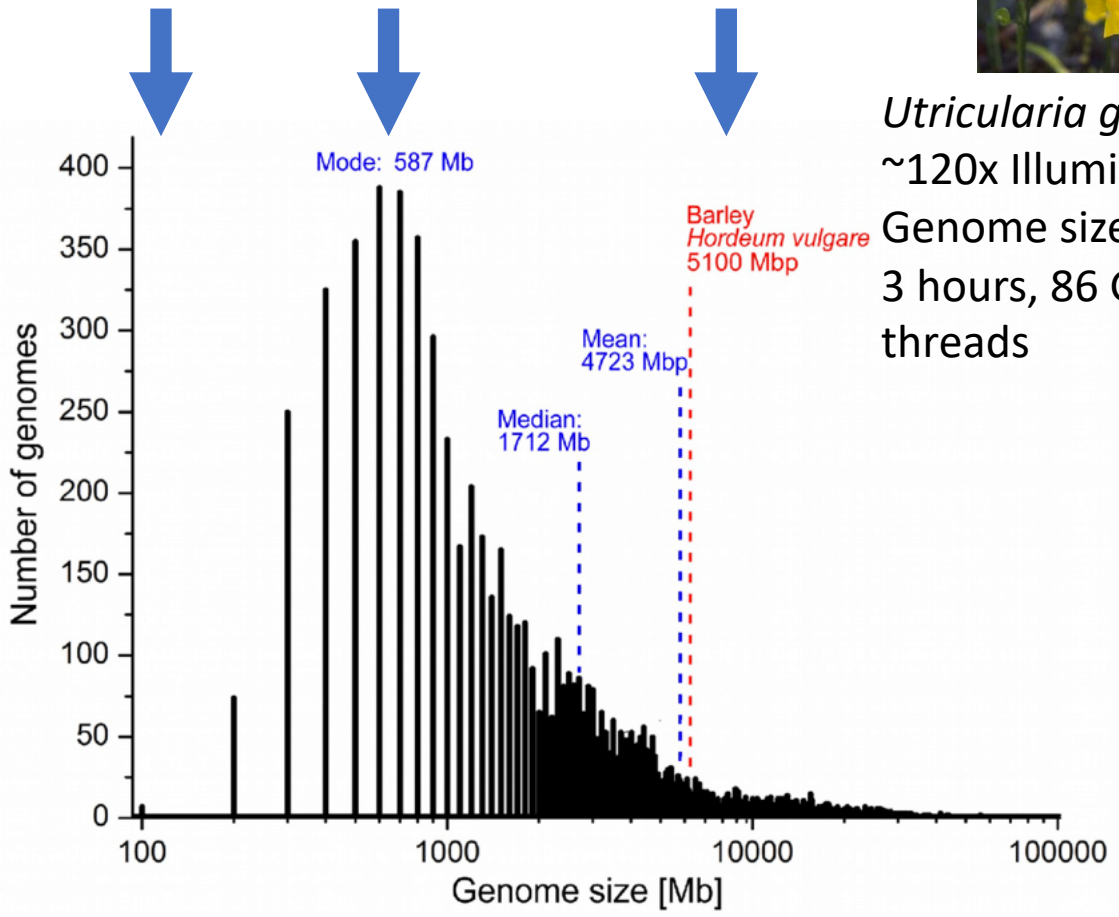
MaSuRCA –hybrid



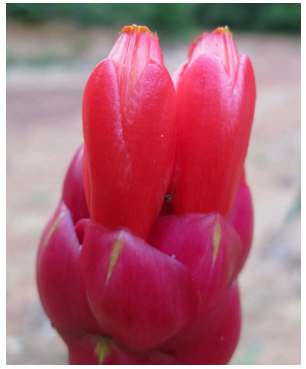


Zimin et al. 2017; *Genome Research*

Examples for requirements



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



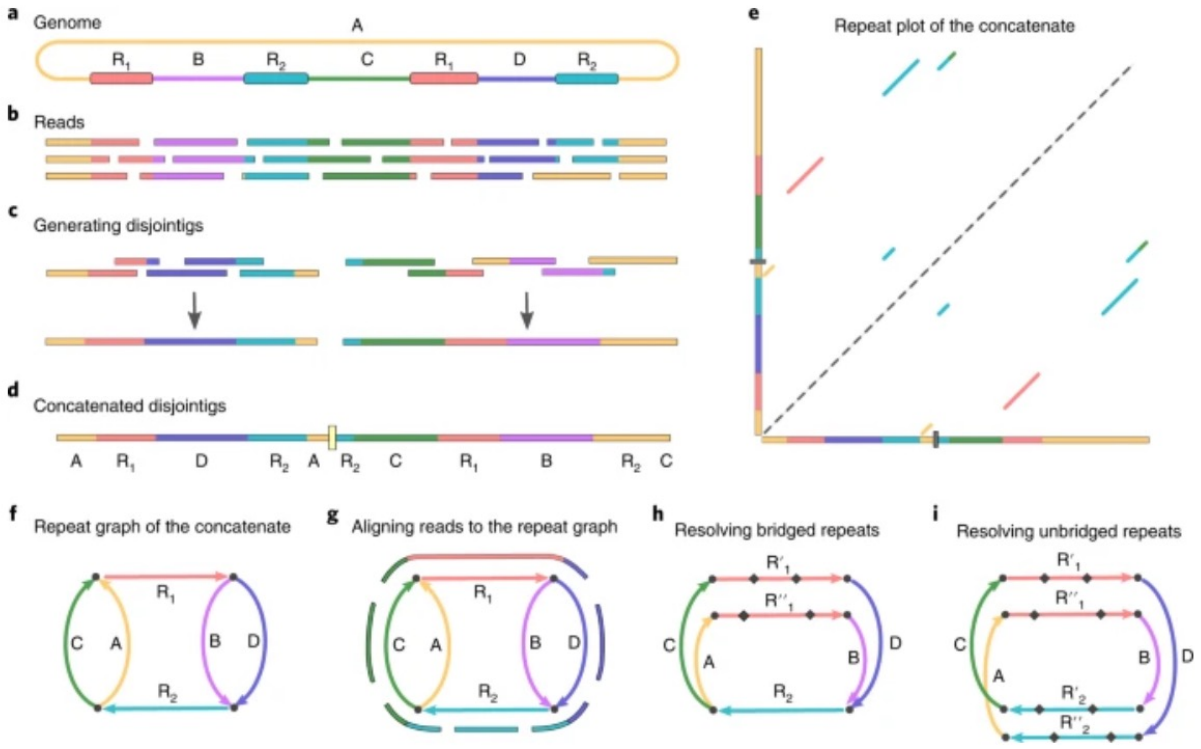
Costus spiralis
80x Illumina, 20x Nanopore
Genome size: 1 GB
3 days, 200 GB storage, 28 threads

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

Long-read assemblers

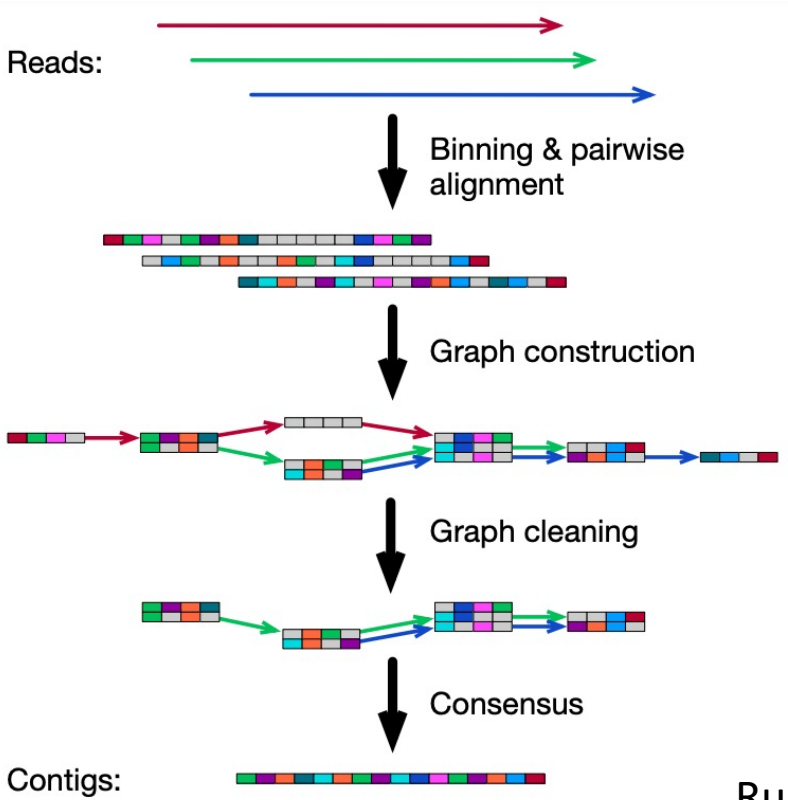
- Many options out there; size of the genome can be “make or break”

Flye – repeat graphs



Kolmogorov et al. 2019

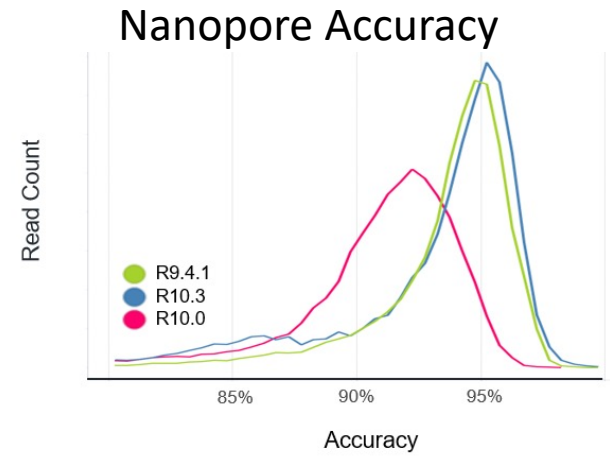
wtdbg2 – fuzzy-Bruijn assembly graph

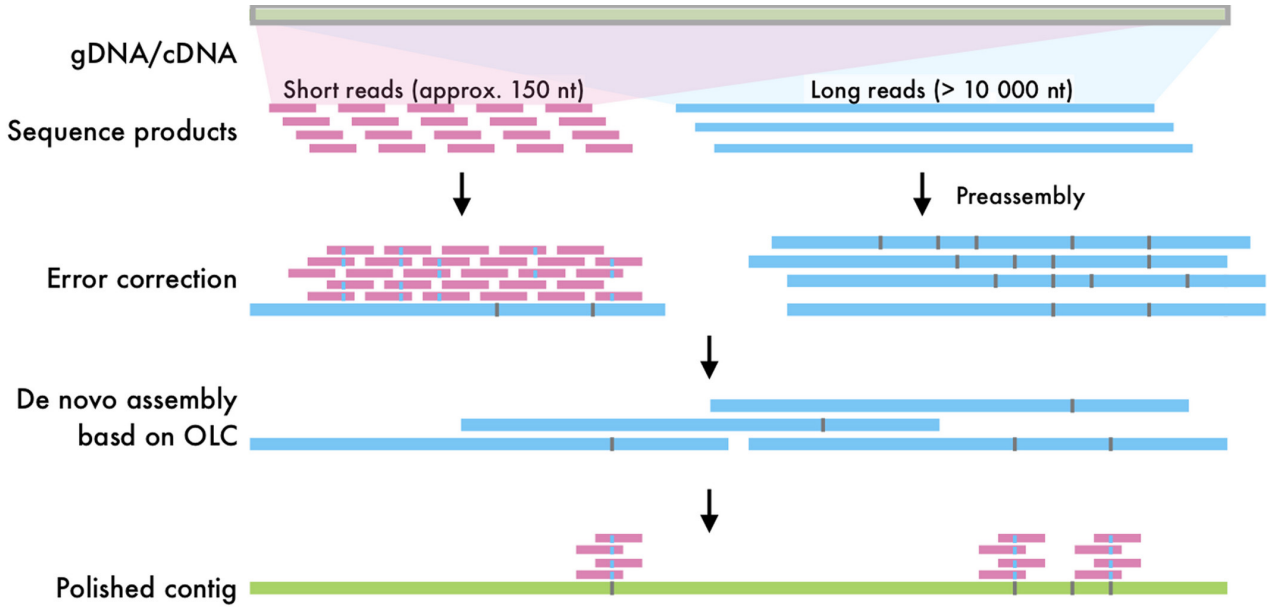


Ruan and Li, BioRxiv

Is the error rate a problem?

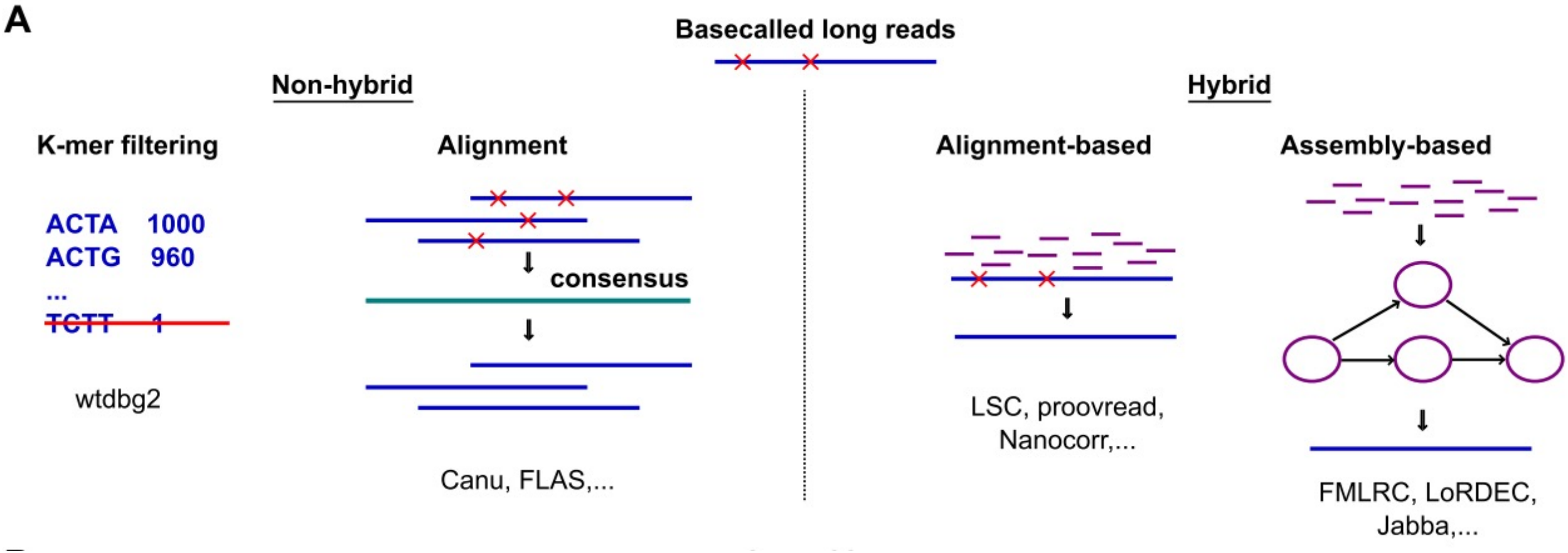
- Definitely can be, but both Nanopore and PacBio have made recent improvements
- How can we fix it?
  - Error correct the long reads prior to assembly
  - Polish the assembly afterwards
- Illumina data very useful but not mandatory





Kono and Arakawa 2019; *Development, Growth, and Differentiation*

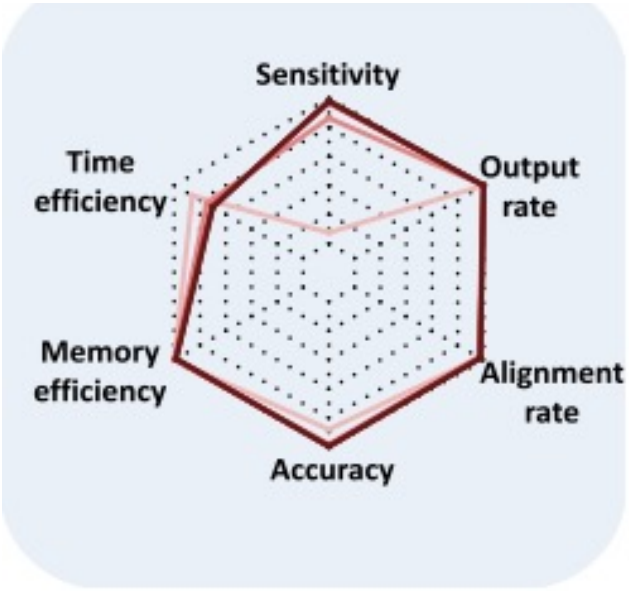
Error Correcting long-reads



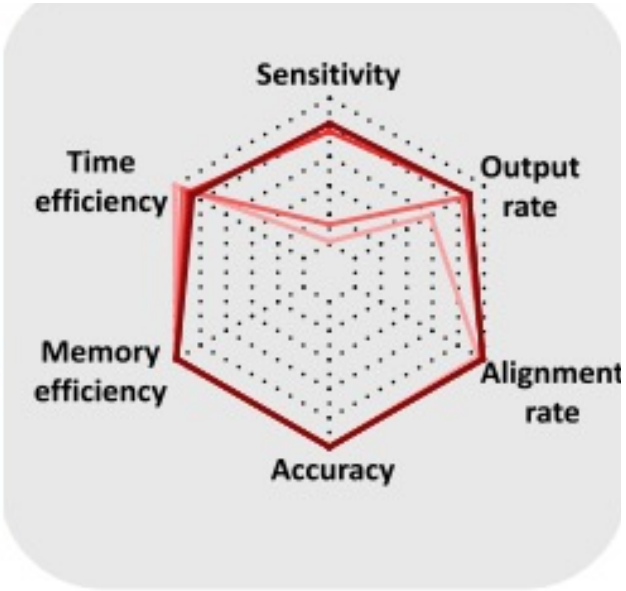
Amarasinghe et al. 2020; *GBE*

Error correction

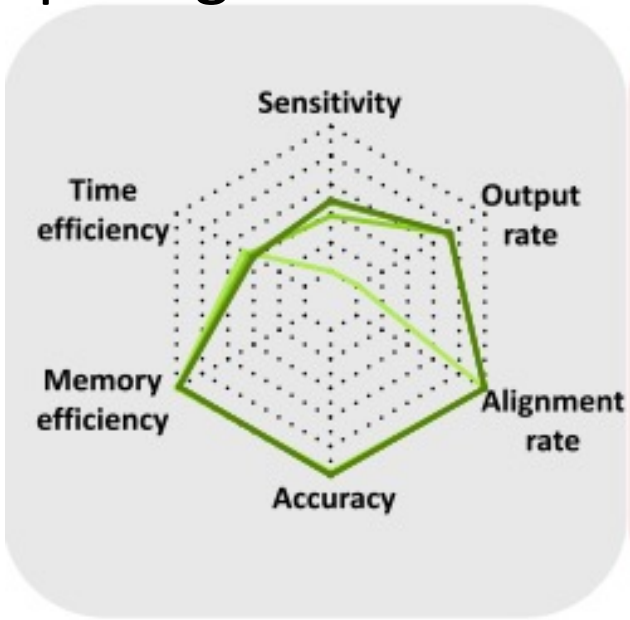
- Hybrid (using Illumina data) works better than just using depth of long-read
- Many factors to consider when comparing methods



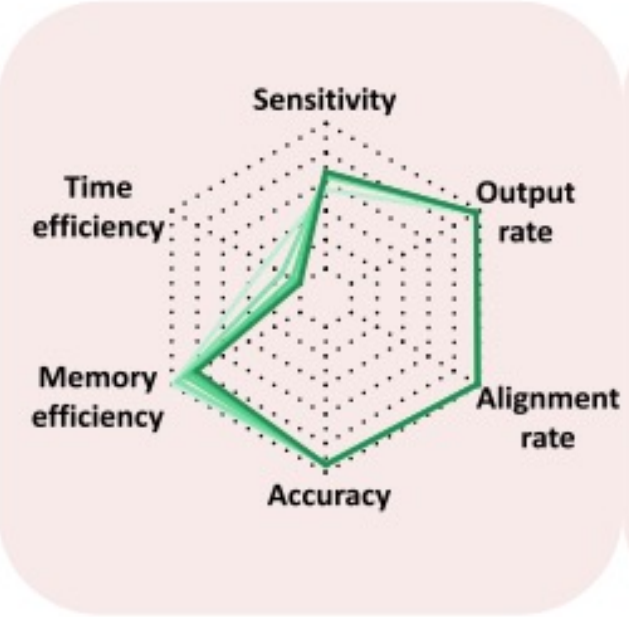
FMLRC



Jabba



ECTools



LSC

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

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

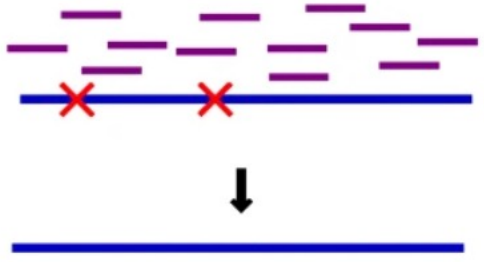
Sort by position

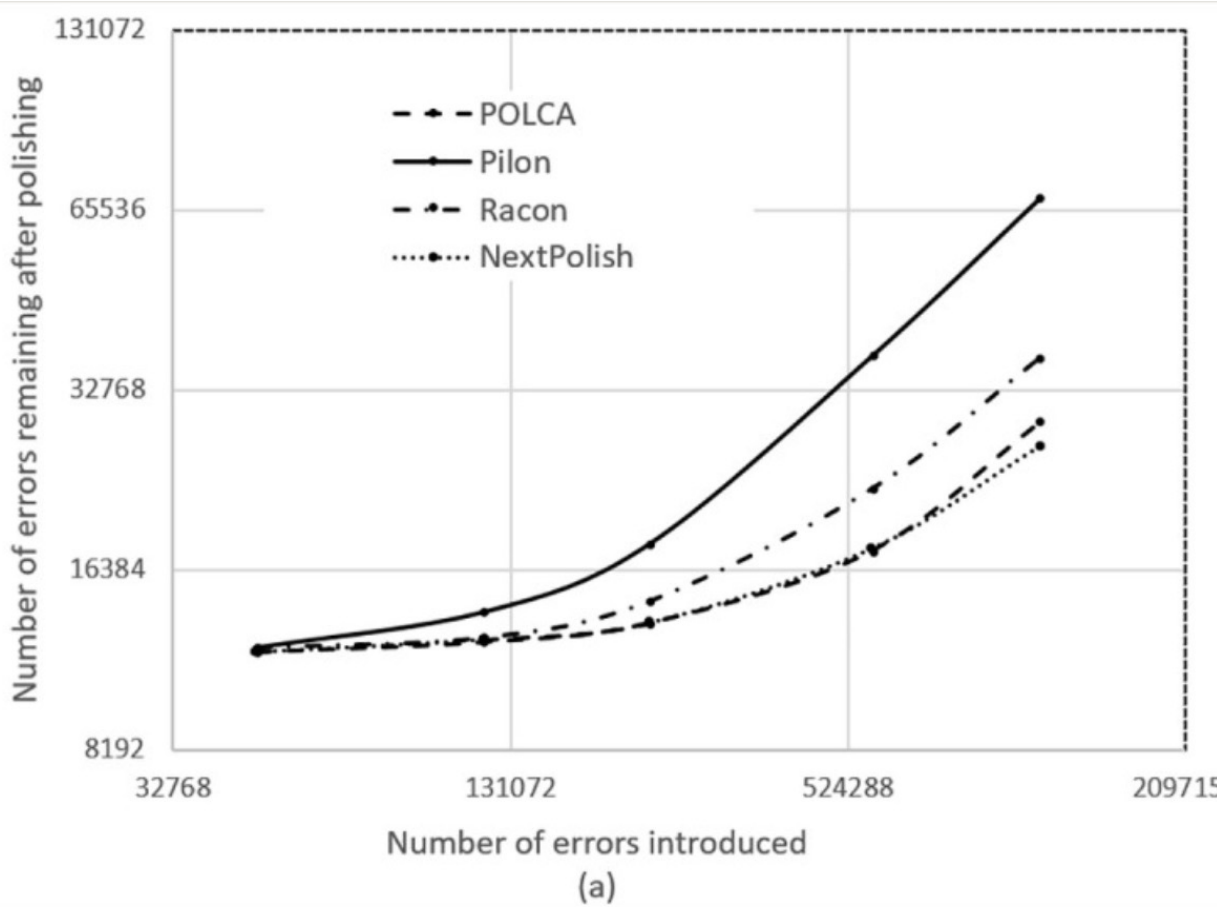
↓ Freebayes

Call variants

↓ polca

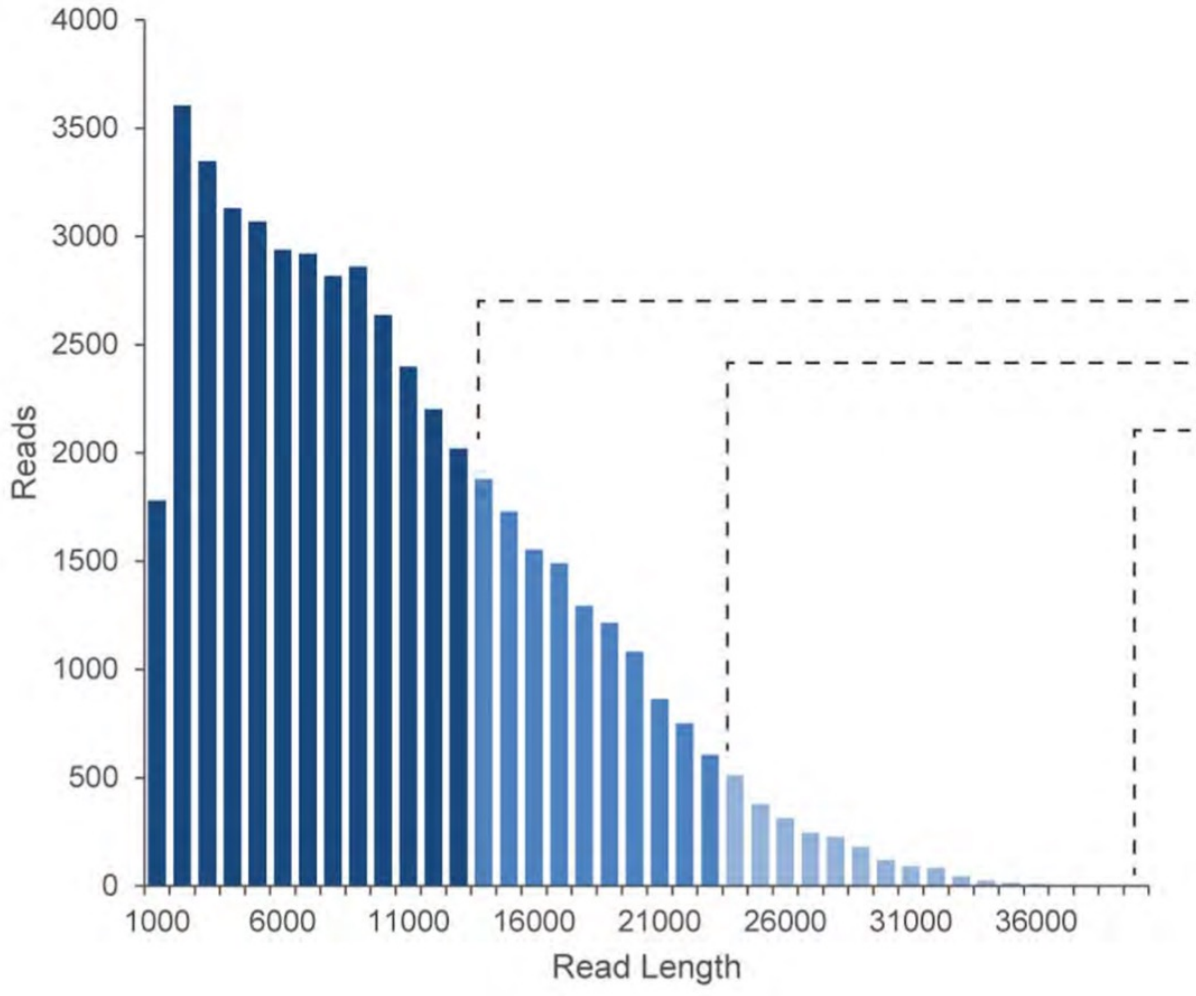
Replace errors with
Highest scoring alternative





Zimin and Salzberg 2020; *PLoS Comp Biology*

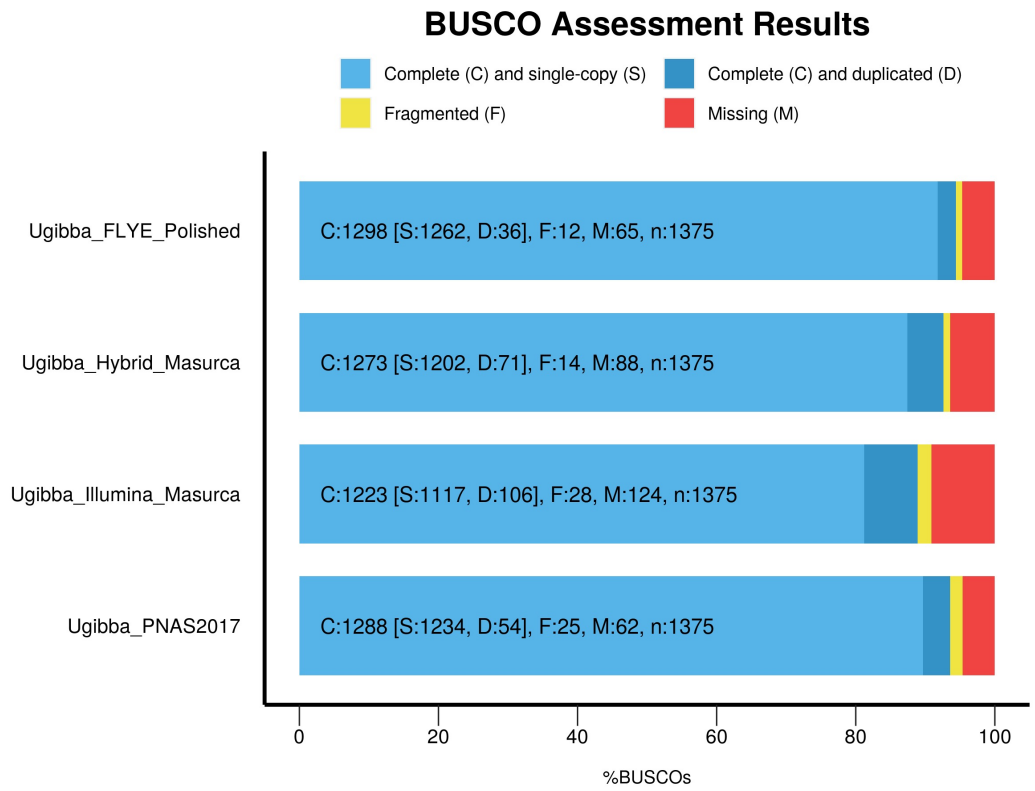
Quality Control Metrics



Half of data in reads/contigs (N50): > 14 kb
Top 5% of reads: >24 kb
Maximum read length: >40 kb
Number of reads: 28,000

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



Long-read example data set

- For the walkthrough we will be using *Utricularia gibba*
  - Humped bladderwort; carnivorous aquatic species
  - Small genome size (76 MB) with only 3% of the genome non-coding
- Small data set that can be run locally or on a VM from CyVerse and all analyses should finish quickly
- Incorporates publicly available data using a high-quality genome assembly and RNA-Seq data for multiple organ types
  - Bladder, leaf, rhizoid, and stem

Utricularia gibba
Humped bladderwort

U. gibba traps

Tutorial

Assembly	Number of Contigs	N50 (bp)	Variants polished
MaSuRCA short-read only			
wtdbg2 long-read only			
Flye long-read only			
MaSuRCA hybrid (short- and long-read)			

Questions

@JLandisBotany jbl256@cornell.edu