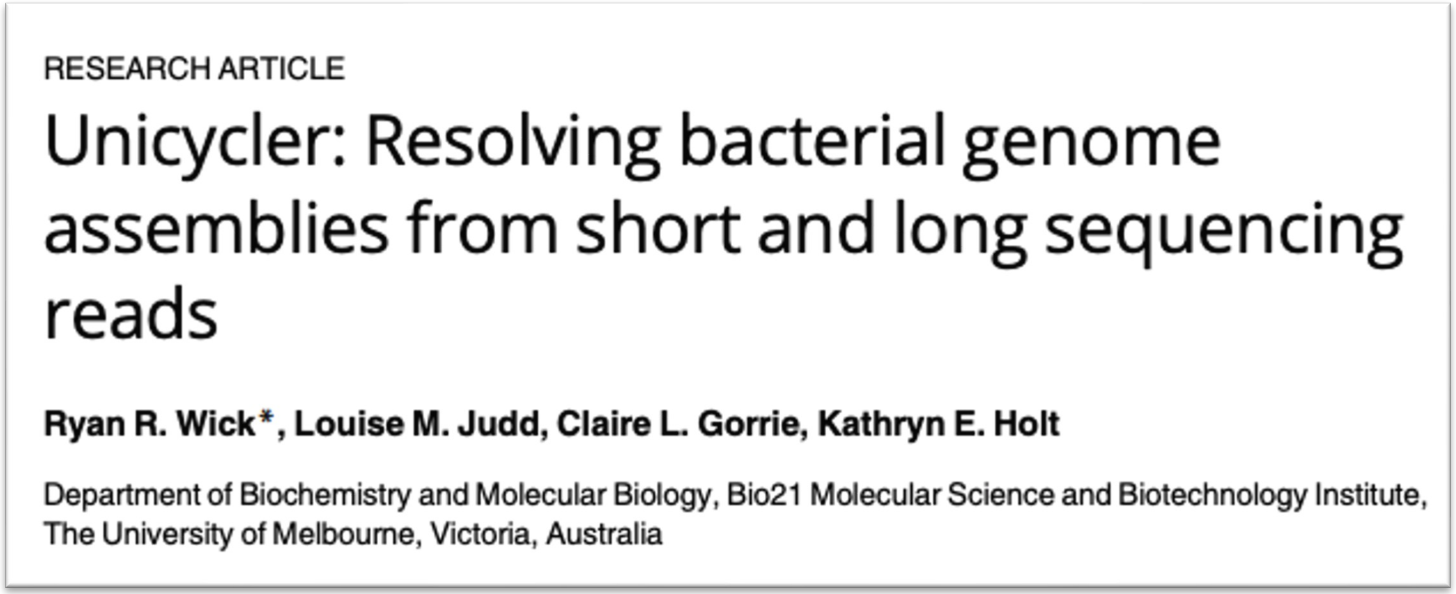
Unicycler: bacterial genome assemblies from short and long read sequences

Journal Club

Zoey Germuskova



RESEARCH ARTICLE

Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads

Ryan R. Wick*, Louise M. Judd, Claire L. Gorrie, Kathryn E. Holt

Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Victoria, Australia

Unicycler

PLOS Computational Biology | https://doi.org/10.1371/journal.pcbi.1005595 June 8, 2017

https://github.com/rrwick/Unicycler#2022-update

What is Unicycler

= Assembly pipeline for bacterial genomes

Use cases:

- Short reads only
- Long reads only
- Short and long reads from the same isolate (best case)

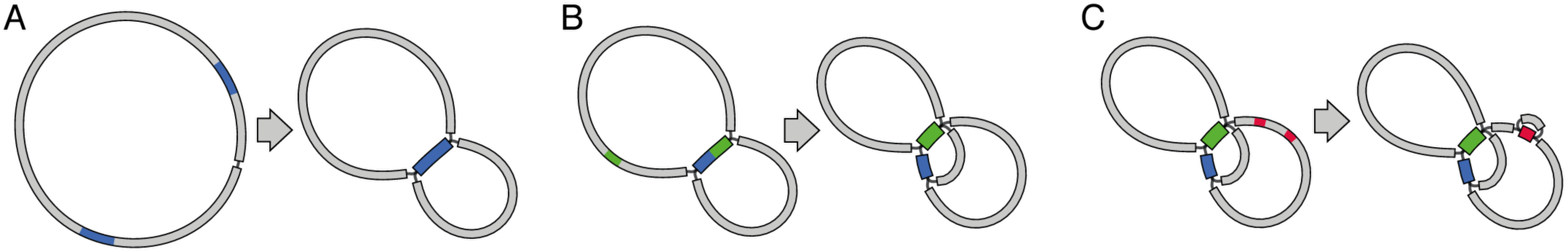
Why use Unicycler:

- It circularizes replicons
- Produces assembly graph in addition to contigs FASTA file
- Handles plasmid-rich or repetitive genomes
- Can use long reads of any depth and quality
- Filters out low-depth contigs (useful in case of low level contamination)

Background

Limitations of short reads

- Accurate but short reads are smaller than many repetitive elements in bacterial genomes
- Assembly of a full genome often not possible
- Incomplete genomes hinder large-scale comparative genomic studies
- **Short** reads don’t have enough info to resolve repeats but **long** reads do => **hybrid assembly**



As repeats are added, the graph becomes increasingly tangled

Source: https://github.com/rrwick/Unicycler#2022-update

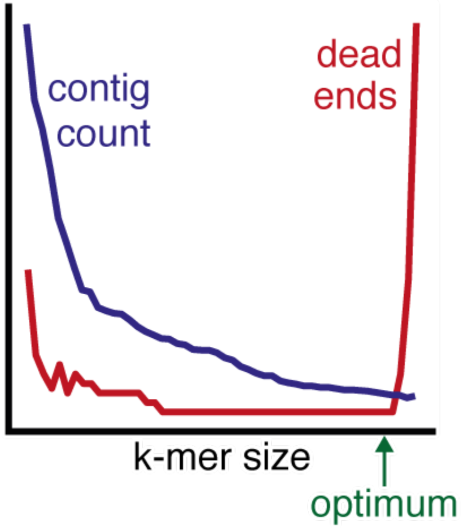
Unicycler pipeline

SPAdes

- Short read assemby
- Made by performing a **De Brujin graph** assembly with a **wide range of k-mer** size
- Each assembly builds on the previous one, allowing SPAdes to get the advantages from:
  - Small k-mer assemblies = more connected graph
  - Large k-mer assemblies = better resolved repeats

- When assembling Illumina reads,

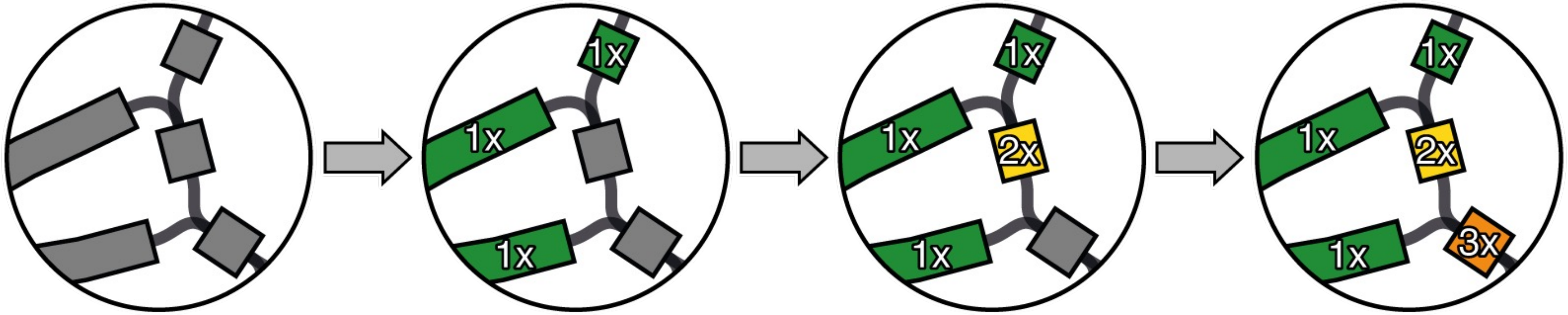
  Unicycler functions as SPAdes optimizer



Unicycler pipeline

Multiplicity

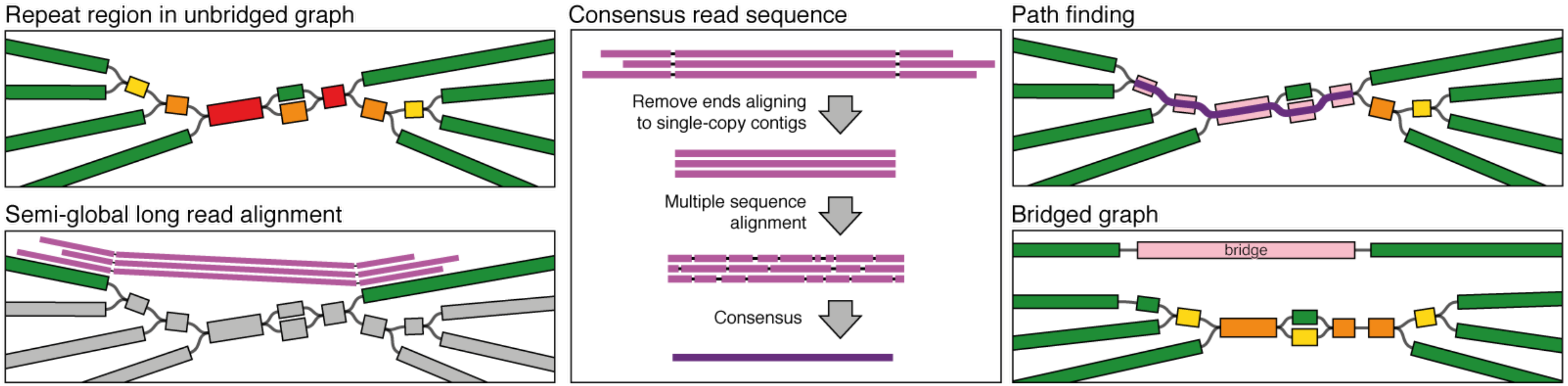
- Goal: distinguish between single copy contigs and repeat contigs
- Read depth can be indicative of multiplicity if genome single chromosome
- Greedy alorithm uses **read depth** and **connectivity**
- Multiplicity of 1 assigned to contigs close to median depth and with one connection at either end



Unicycler pipeline

Bridging

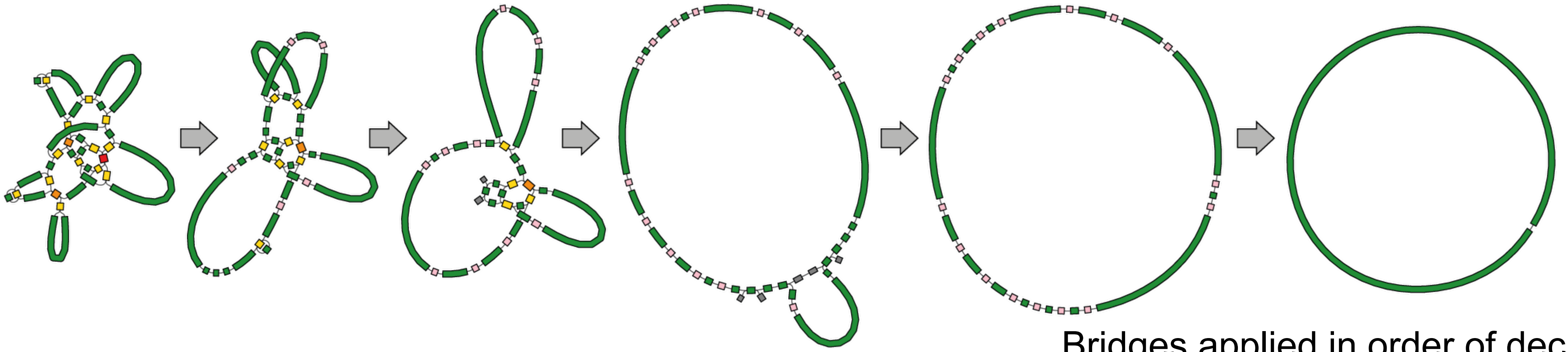
- Scaffolding by building bridges between single copy contigs using the path information in the SPAdes assembly
- Bridges made using long reads can resolve larger repeats


Repeat region in unbridged graph
Semi-global long read alignment
Consensus read sequence
Remove ends aligning to single-copy contigs
Multiple sequence alignment
Consensus
Path finding
Bridged graph
bridge

Unicycler pipeline

Bridge application

- Multiple possible bridges, some may be conflicting
- **Quality score** assigned to each bridge
  - Number of reads supporting the bridge
  - Length of contigs to be bridged
  - Alignment quality between read consensus and graph path
  - Read depth consistency between contigs

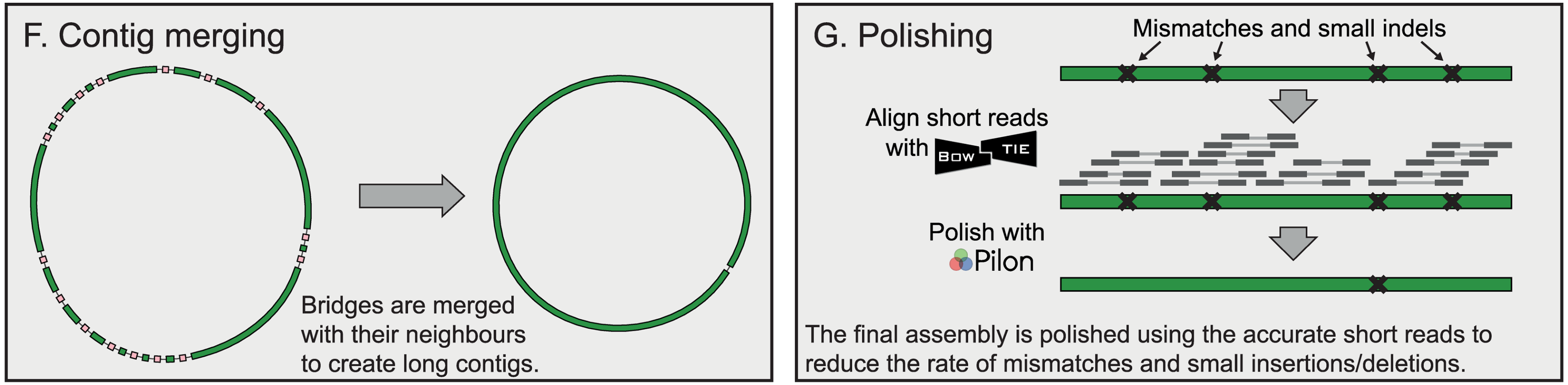


Bridges applied in order of decreasing quality

Unicycler pipeline

Merging & polishing

- Reducing rate of mismatches and small errors using the accurate short reads
- Pilon



Bridges are merged with their neighbours to create long contigs.

The final assembly is polished using the accurate short reads to reduce the rate of mismatches and small insertions/deletions.

Modes

Conservative

- Bridge quality cutoff high
- Low risk of misassembly
- Least likely to produce a complete genome
- Use when high accuracy needed

Normal

- Bridge quality cutoff intermediate

Bold

- Lower quality bridges used
- Greater risk of error
- Use when completeness more important than accuracy

Performance evaluation of Unicycler

1. **Simulated read** sets from 12 **reference** genomes

2. **Real** read sets from *E.coli* K-12

3. **Novel** *Klebsiella pneumoniae* isolate genome (no reference genome)

- Comparison to other pipelines and tools

Metrics

Misassemblies

- cases where a contig aligns to the reference genome in multiple pieces, not as a single continuous alignment, indicating a structural error in the contig
  - For simulated reads, misassemblies indicate assembler mistake
  - In *E.coli*, false positive misassemblies also possible

Small-error rate

- Mismatches
- Indels

Quast

NGA50

- Length of contig-to-reference alignment
- A correctly assembled contig will have a single, unbroken alignment to the reference
- A misassembled contig will be divided into multiple smaller alignments

Simulated read sets

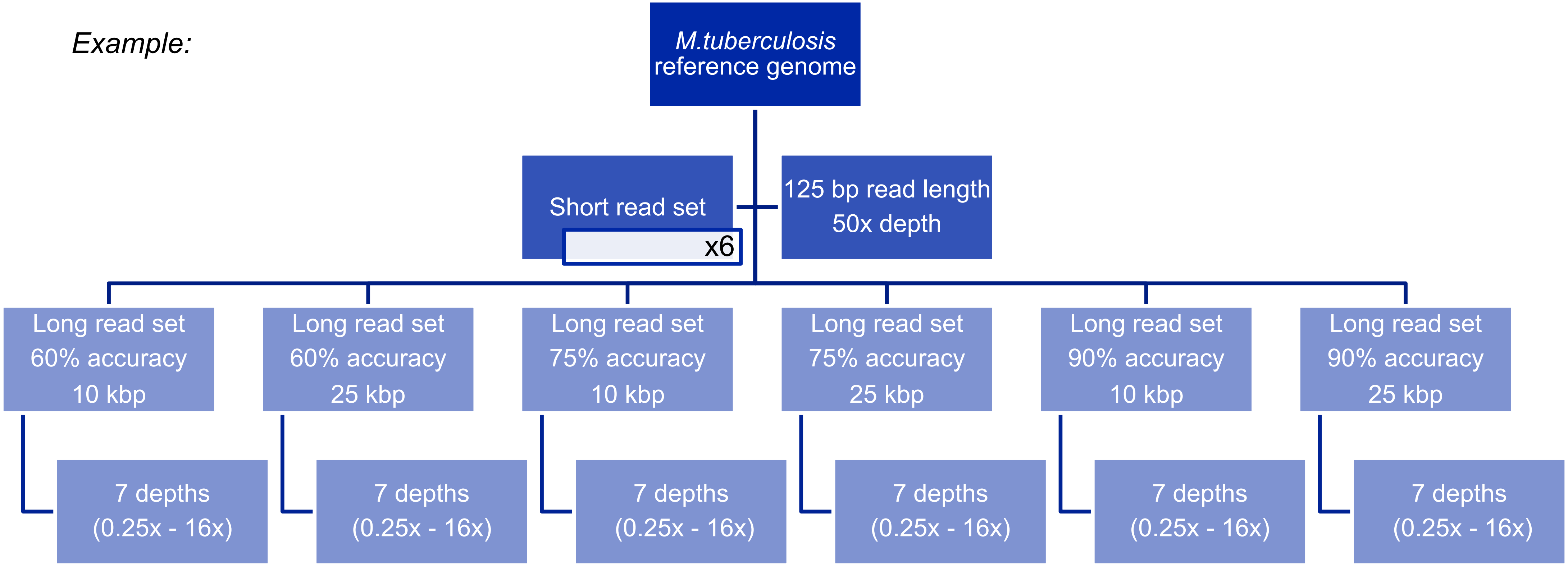
Species	Strain	Genome size (bp)	GC content	Description	MLST sequence type	Other features	GenBank assembly accession
Acinetobacter baumannii	A1	3,917,739	39.3%	Circular chromosome, one plasmid	231	Large, repetitive biofilm-associated protein gene	GCA_000830055.1
Acinetobacter baumannii	AB30	4,335,793	39.0%	Circular chromosome	758	Large, repetitive biofilm-associated protein gene	GCA_000746645.1
Escherichia coli	K-12 MG1655	4,641,652	50.8%	Circular chromosome	10		GCA_000005845.2
Escherichia coli	O25b: H4-ST131 EC958	5,249,449	50.8%	Circular chromosome, two plasmids	131		GCA_000285655.3
Klebsiella pneumoniae	30660/ NJST258_1	5,540,936	57.2%	Circular chromosome, five plasmids	258		GCA_000598005.1
Klebsiella pneumoniae	MGH 78578	5,694,894	57.1%	Circular chromosome, five plasmids	38		GCA_000016305.1
Klebsiella pneumoniae	NTUH-K2044	5,472,672	57.4%	Circular chromosome, one plasmid	23		GCA_000009885.1
Mycobacterium tuberculosis	H37Rv	4,411,532	65.6%	Circular chromosome		High-copy-number PE and PPE genes	GCA_000195955.2
Saccharomyces cerevisiae	S288c	12,157,105	38.1%	16 linear chromosomes, circular mitochondrial sequence		Eukaryote	GCA_000146045.2
Shigella dysenteriae	Sd197	4,560,911	51.0%	Circular chromosome, two plasmids	146	High insertion sequence content	GCA_000012005.1
Shigella sonnei	53G	5,220,473	50.7%	Circular chromosome, four plasmids	152	High insertion sequence content	GCA_000283715.1
Streptococcus suis	BM407	2,170,808	41.0%	Circular chromosome, one plasmid	1		GCA_000026745.1

https://doi.org/10.1371/journal.pcbi.1005595.t001

- Variety of genome sizes and complexities

Simulated read sets

Example:


M.tuberculosis reference genome
Short read set
x6
125 bp read length
50x depth
Long read set 60% accuracy 10 kbp
Long read set 60% accuracy 25 kbp
Long read set 75% accuracy 10 kbp
Long read set 75% accuracy 25 kbp
Long read set 90% accuracy 10 kbp
Long read set 90% accuracy 25 kbp
7 depths (0.25x - 16x)
7 depths (0.25x - 16x)
7 depths (0.25x - 16x)
7 depths (0.25x - 16x)
7 depths (0.25x - 16x)
7 depths (0.25x - 16x)

- For each strain there are 6 short-read sets and 42 hybrid-read sets

Comparing to different pipelines/tools

Short read assembly evaluation

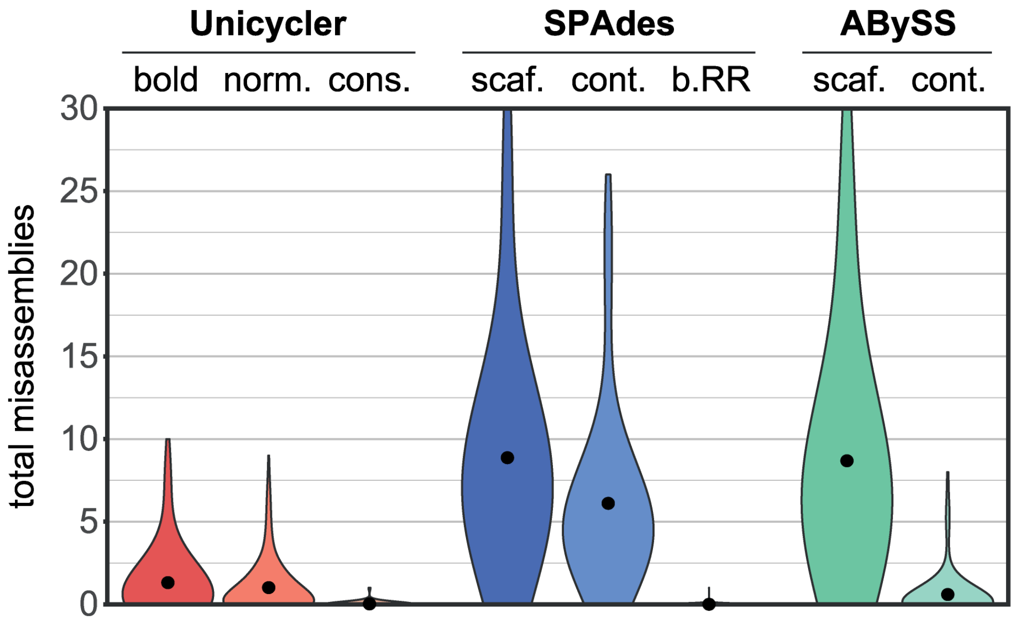
- Unicycler
- SPAdes
- ABySS

Hybrid assembly evalutation

- Unicycler
- SPAdes
- npScarf
- Cerulean

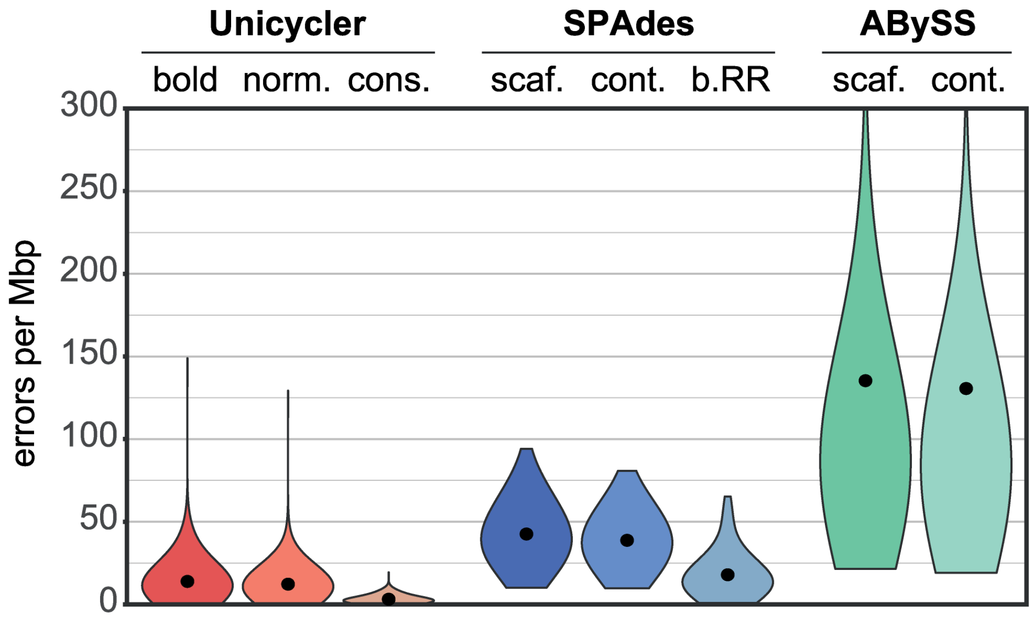
Simulated read sets – short read assembly evaluation

Misassemblies



	Unicycler bold	Unicycler norm.	Unicycler cons.	SPAdes scaf.	SPAdes cont.	SPAdes b.RR	ABySS scaf.	ABySS cont.
mean:	1.3	1.0	0.022	8.9	6.1	0.003	8.7	0.59
median:	1	0	0	7	5	0	7	0
maximum:	10	9	1	31	26	1	31	8

Small errors

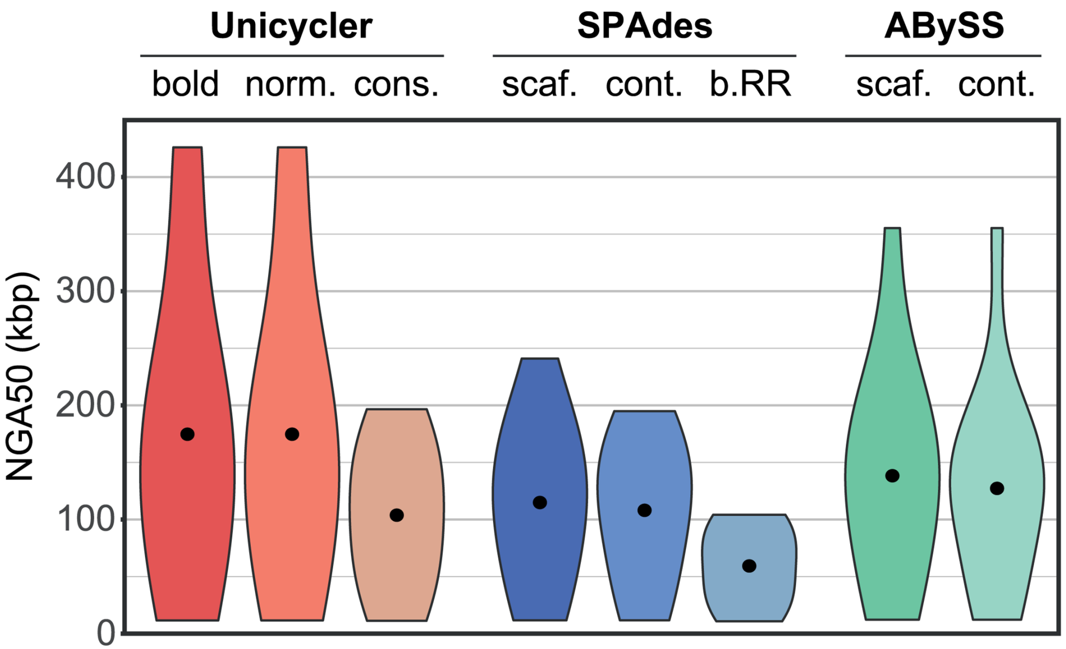


	Unicycler bold	Unicycler norm.	Unicycler cons.	SPAdes scaf.	SPAdes cont.	SPAdes b.RR	ABySS scaf.	ABySS cont.
mean:	14	12	3	43	39	18	135	131
median:	12	11	2	40	37	14	94	91
maximum:	149	129	19	94	81	65	651	649

n=360 per assembler

Simulated read sets – short read assembly evaluation

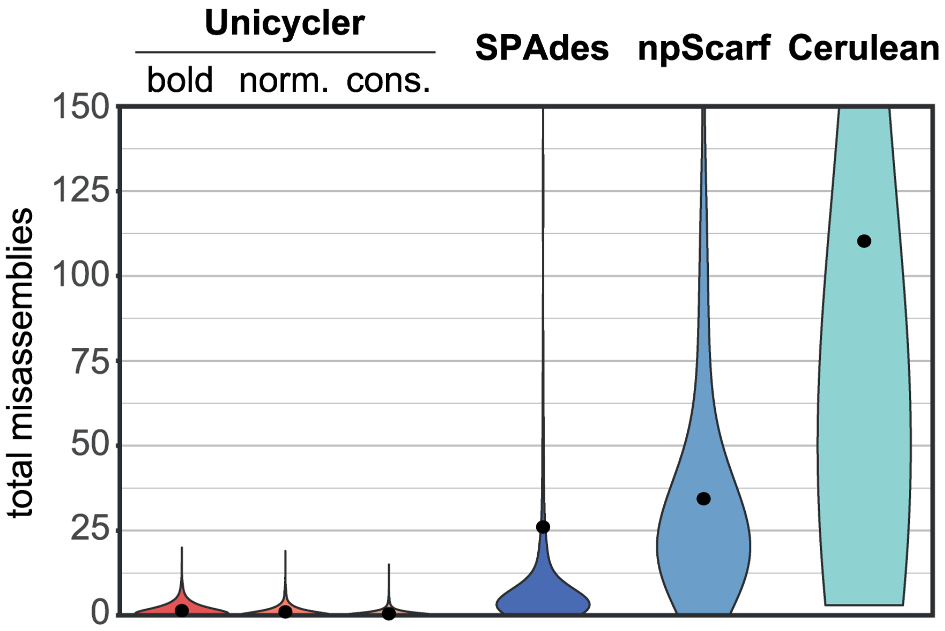
NGA50



mean:	175	175	104	115	108	59	138	127
median:	155	155	101	114	113	62	151	135
maximum:	426	426	197	241	195	104	355	355

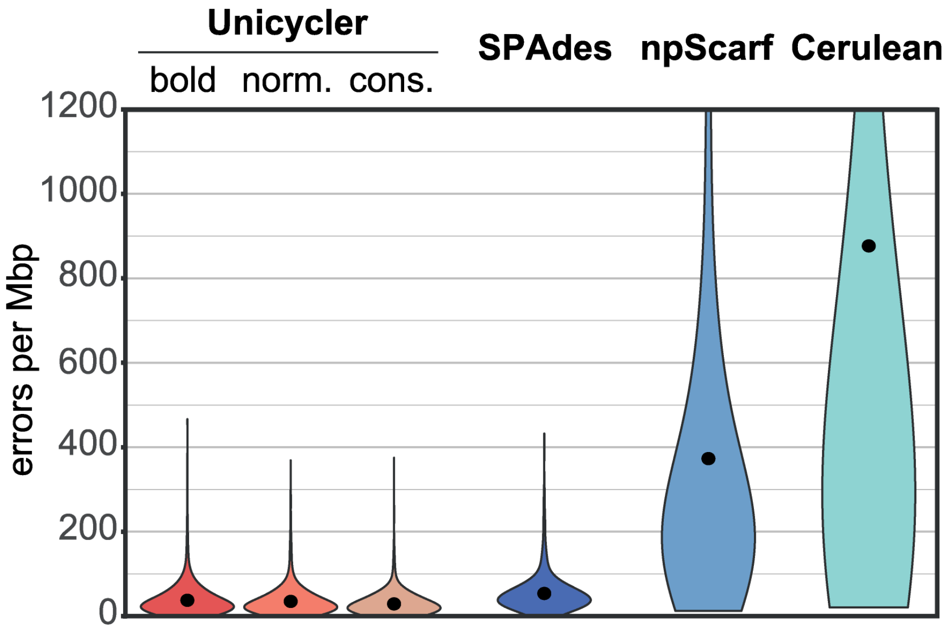
Simulated read sets – hybrid assembly evaluation

Misassemblies



	bold	norm.	cons.	SPAdes	npScarf	Cerulean
mean:	1.4	1.0	0.49	26	34	110
median:	1	0	0	4	24	57
maximum:	20	19	15	791	234	1104

Small errors

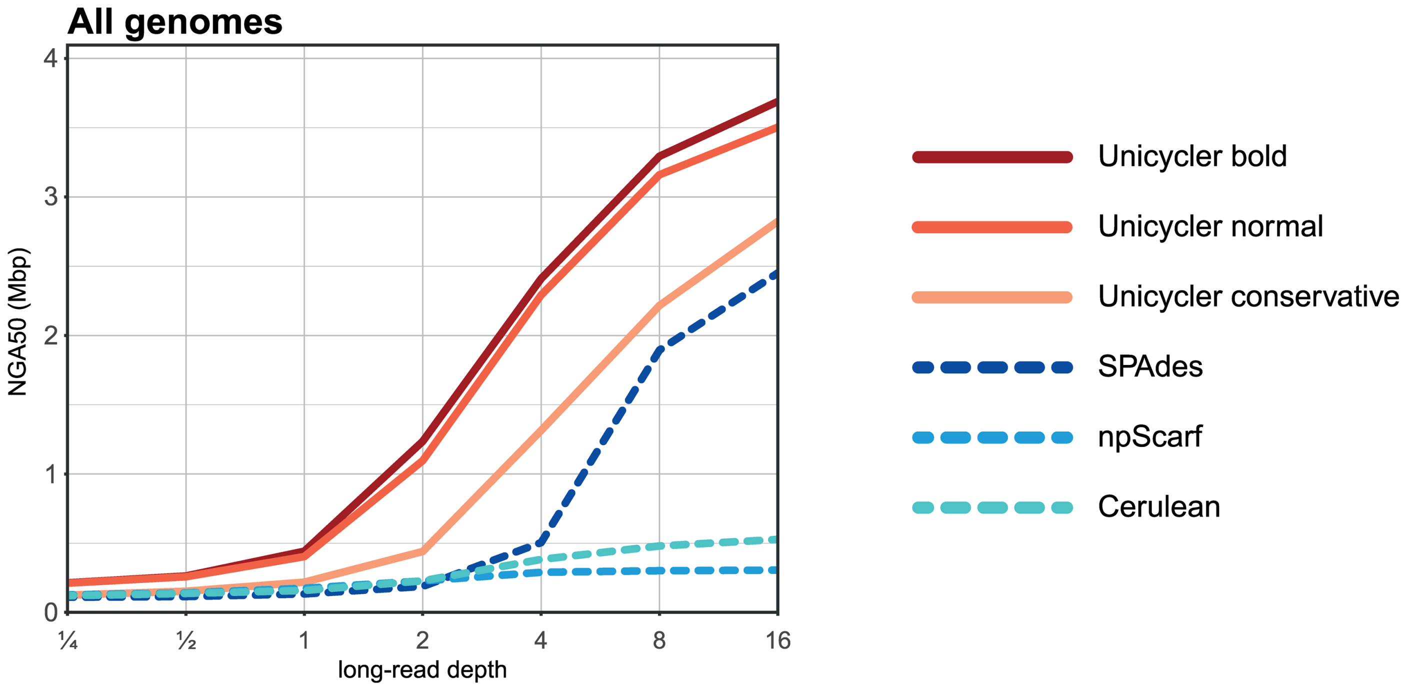


	bold	norm.	cons.	SPAdes	npScarf	Cerulean
mean:	38	35	29	54	373	877
median:	24	24	22	42	219	333
maximum:	467	369	375	432	5021	11916

n=2520 per assembler

Simulated read sets – hybrid assembly evaluation

NGA50

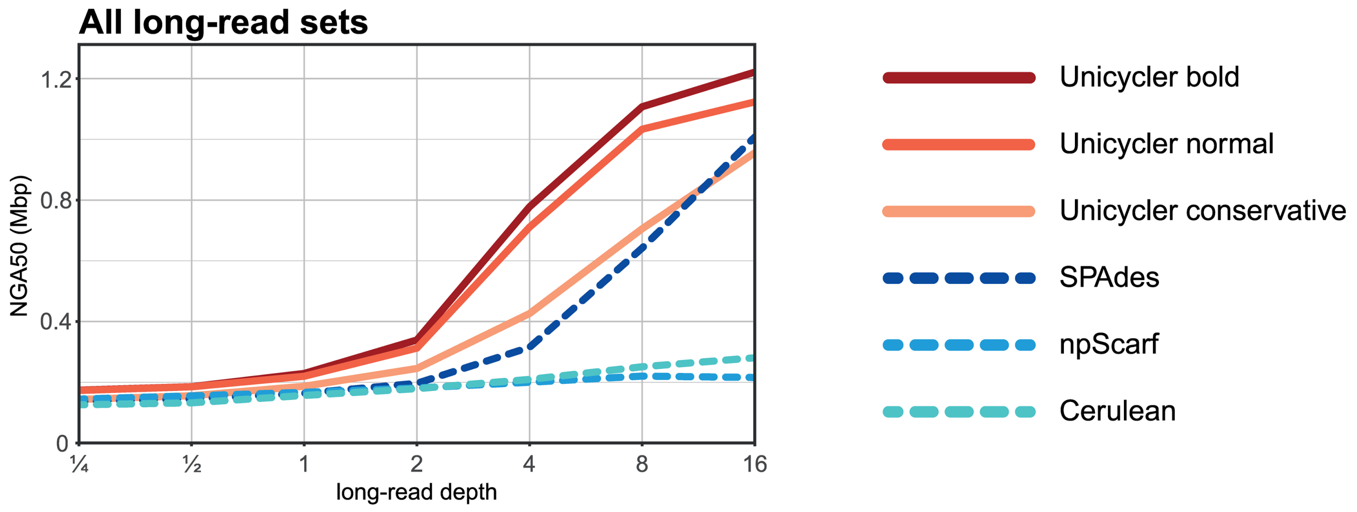


Real *E.coli* K-12 read sets – hybrid assembly

- Short reads produced on Illumina MiSeq
- Long reads from different platforms (ONT, PacBio, different flow cells and chemistries)
- Accuracy assessed by comparison to the reference genome (Sanger-based capillary sequencing at the University of Wisconsin in 1997)

Unicycler produced larger contigs at lower long-read depths than other assemblers

NGA50



All long-read sets

NGA50 overall lower compared to simulated reads

Why?

Klebsiella pneumoniae de-novo assembly

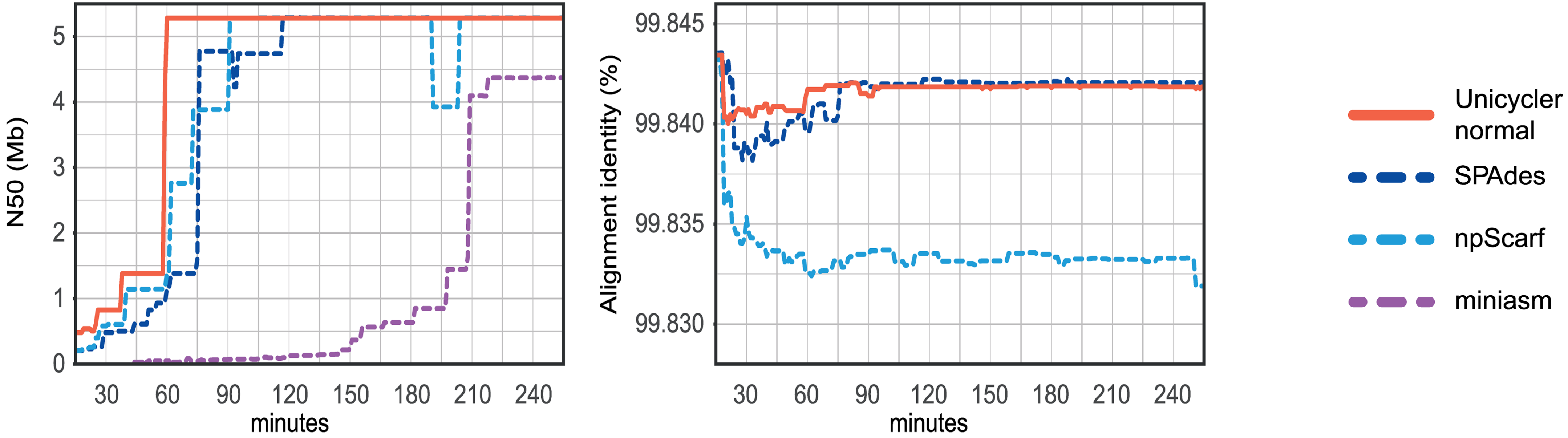
- *K.pneumoniae* isolate INF125
- Virulent strain from urine of a Melbourne hospital patient
- **Real time analysis**
  - Sequencing can be stopped when complete assembly reached

Results:

Overall time?

- 240 subsets of reads (after each minute of sequencing)
  - **Unicycler** completed assebbly with data generated in 45 min (depth = 5.3x)
  - **npScarf** with data generated in 76 min (9x)
  - **SPAdes** with data generated in 102 min (12.1x)
  - **Miniasm** with data generated in 213 min (25.3x)

Klebsiella pneumoniae de-novo assembly



Summary

- Unicycler performed well on both short-read-only and hybrid-read sets
- Larger contigs than other assemblers
- Fewer misassemblies
- Genome sufficiently resolved with low-depth long reads (conserving resources)

Updates

- Unicycler was initially made in 2016, back when long reads were sparse and noisy
- Unicycler was designed to use low-depth and low-accuracy long reads to scaffold a short-read assembly graph to completion, i.e. short-read-first hybrid assembly
- Nowadays, Nanopore sequencing yield is now much higher and **improved accuracy**
- **Long-read-first assembly** is a viable approach (**Trycycler and Polypolish**)
- **Unicycler** still the best tool for **short-read-first hybrid assembly** of bacterial genomes and short-read-only bacterial genome assembly

Future

- Improved accuracy of long read sequencing
- Long read only assembly
- Real time assembly and stopping sequencing when assembly complete
- Ultimately lowered costs

Thank you

Questions?