Toward improved MAGs characterization

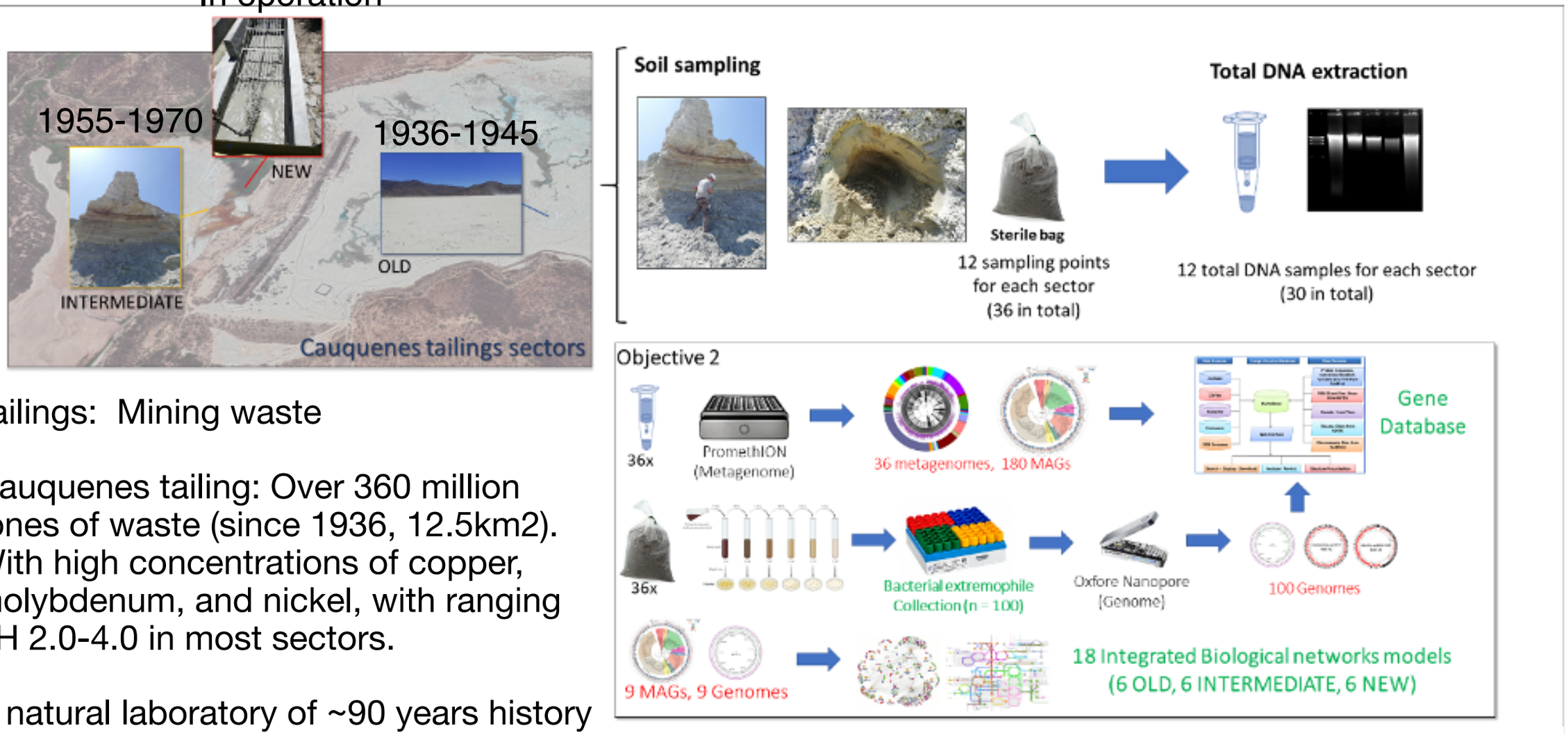
Alex Di Genova
Associate professor
Universidad de O' Higgins

16/05/2023

Systemix center

Anillo ACT210004

In operation

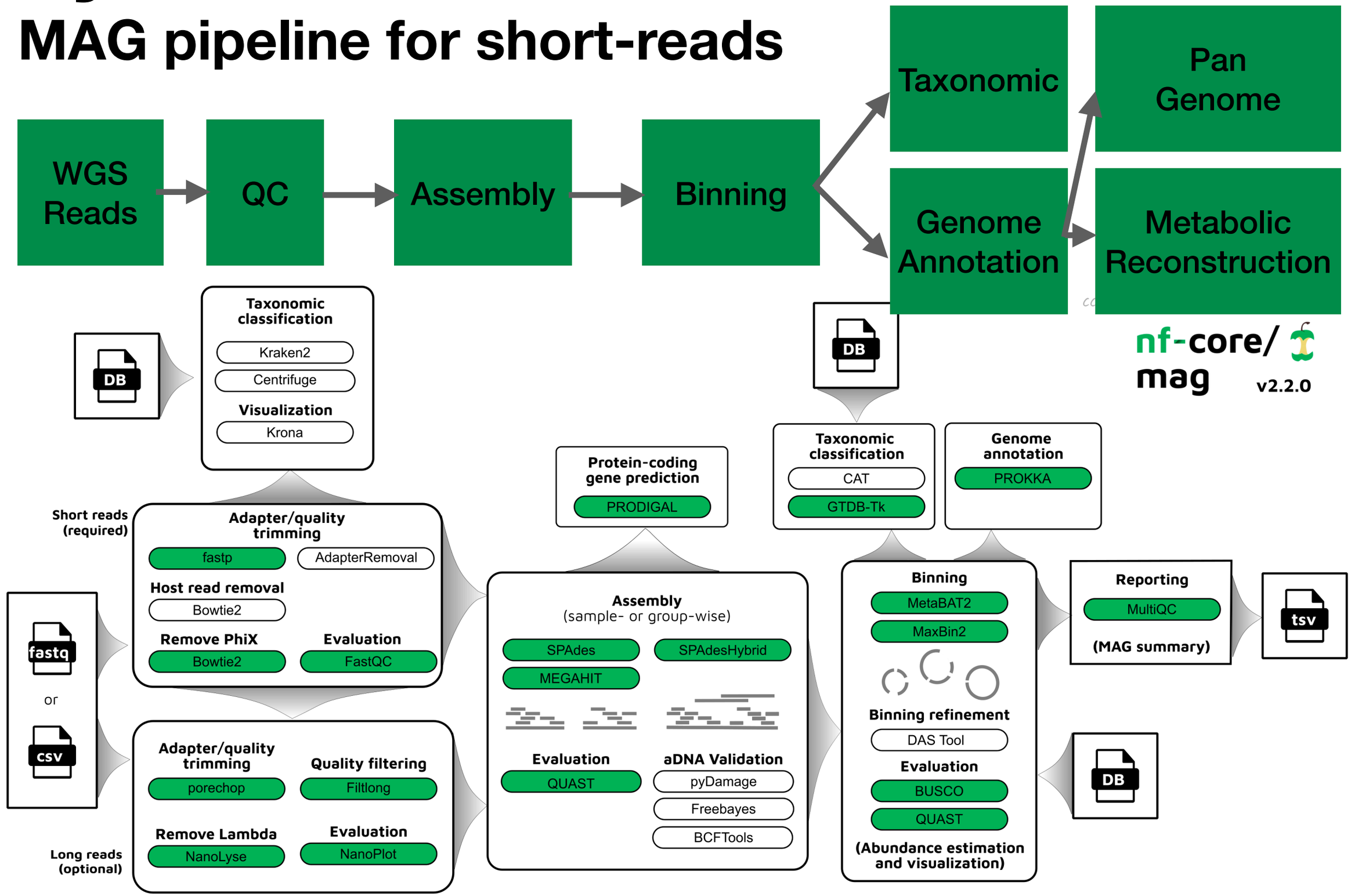


- Tailings: Mining waste
- Cauquenes tailing: Over 360 million tones of waste (since 1936, 12.5km2). With high concentrations of copper, molybdenum, and nickel, with ranging pH 2.0-4.0 in most sectors.
- A natural laboratory of ~90 years history of mine extremophile's bacterial communities.
- Sample and Sequence ~30 metagenomes with short and long reads.
  - 10 with short reads (2x150, 30Gb).

https://systemixcenter.cl/

Systemix - Nexflow

MAG pipeline for short-reads

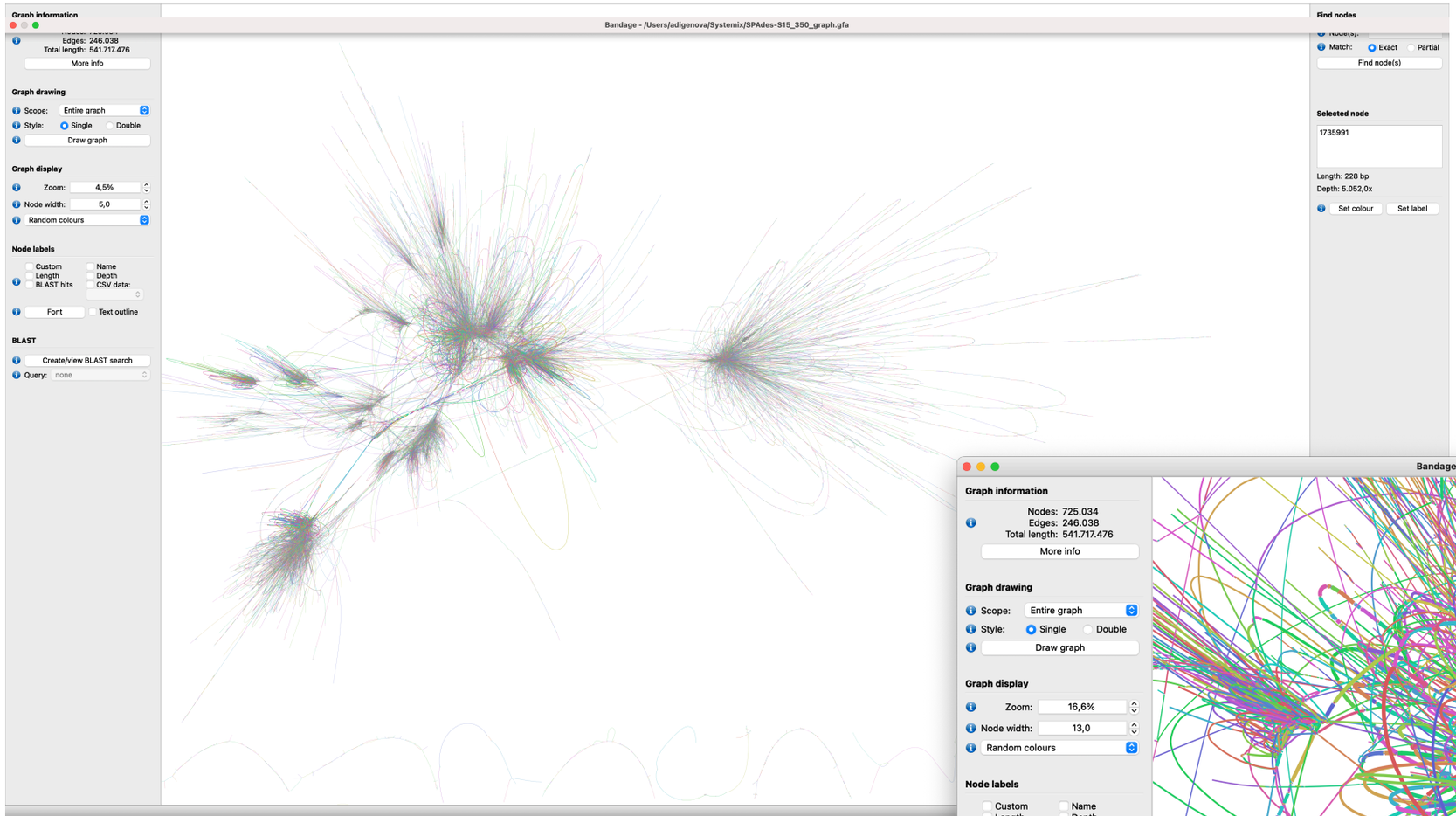


Nextflow : Guarantee reproducibility between HPCs

Di Tommaso, P., Chatzou, M., Floden, E. *et al.* **Nextflow enables reproducible computational workflows**. *Nat Biotechnol* **35**, 316–319 (2017). https://doi.org/10.1038/nbt.3820

Systemix

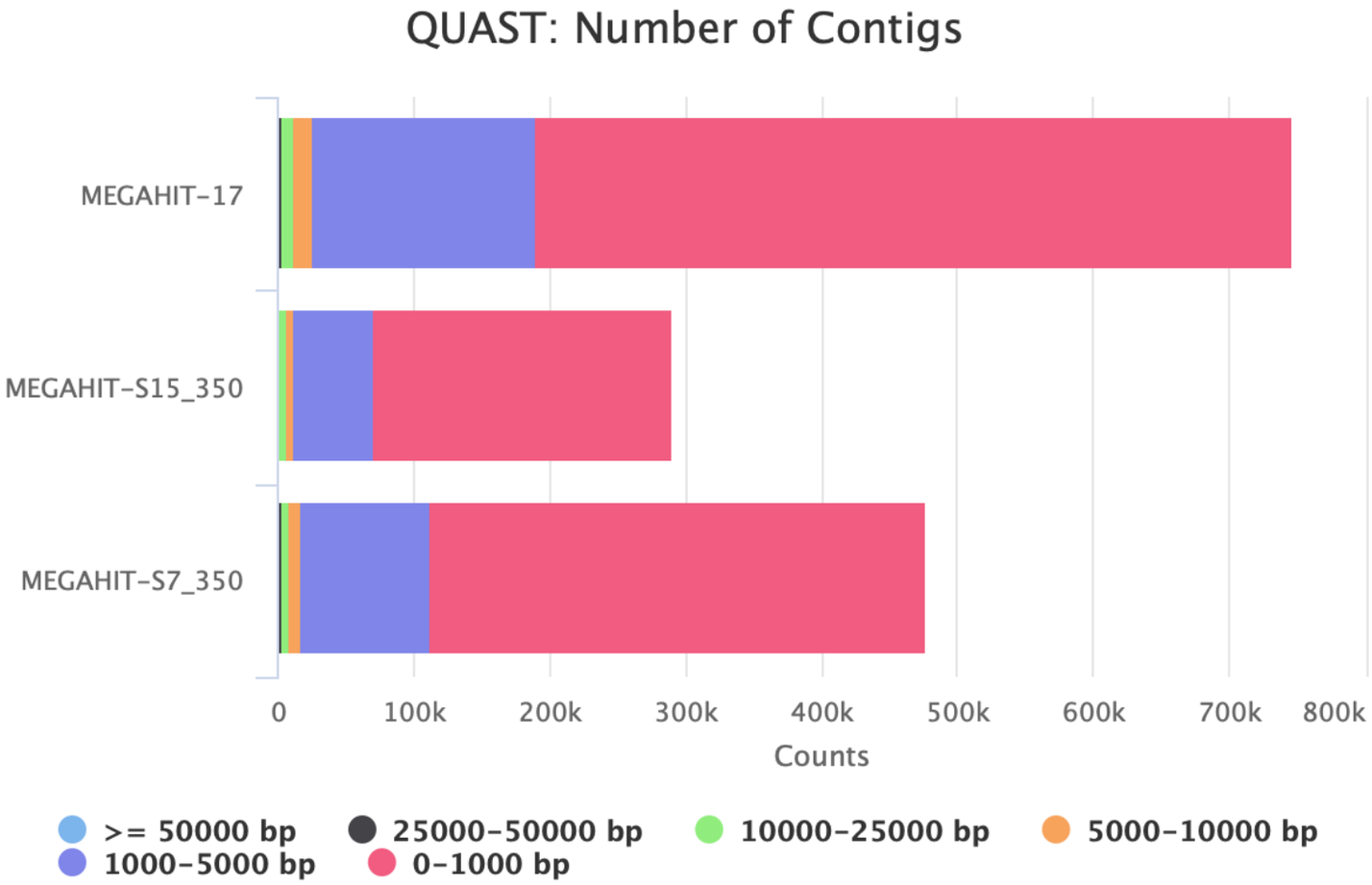
Metagenomes from mining tailings are complex



HairBall

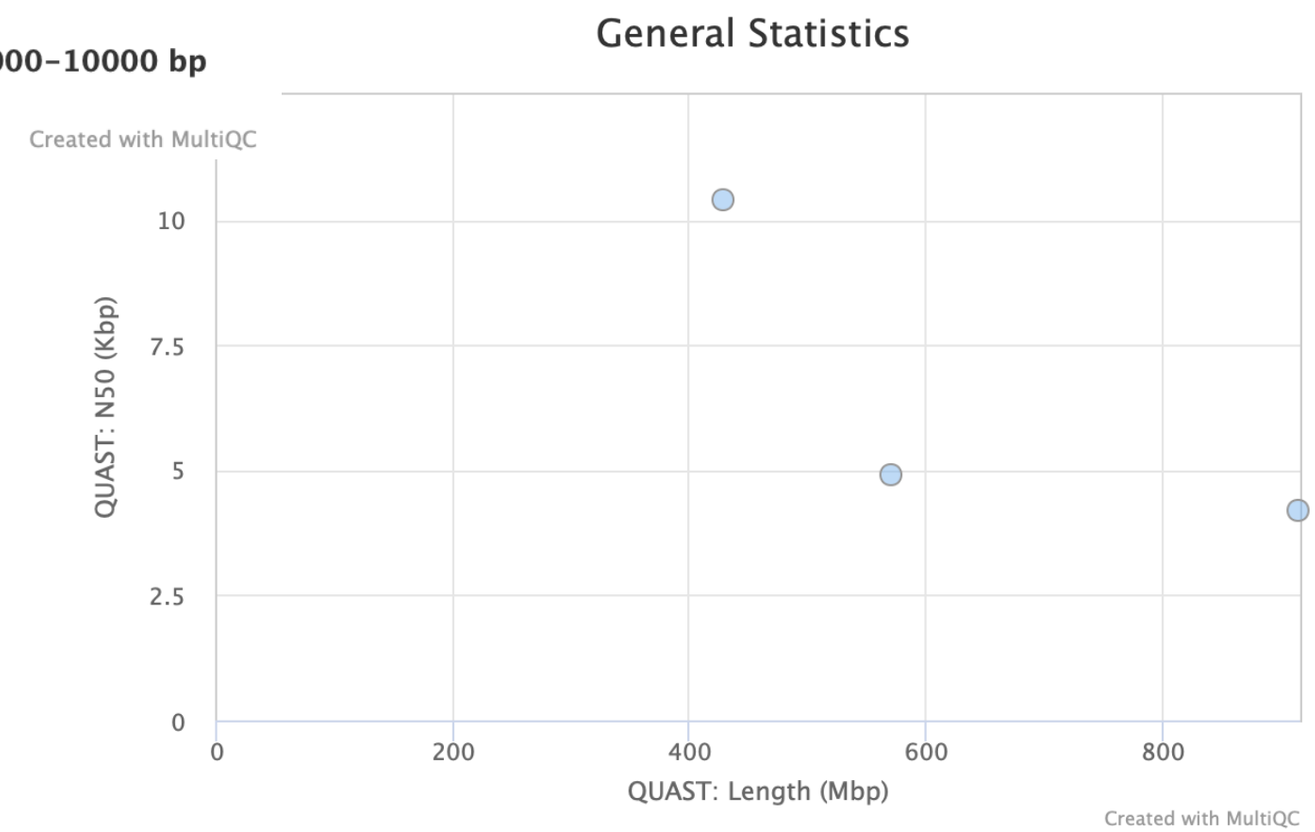
Systemix

MAG pipeline results



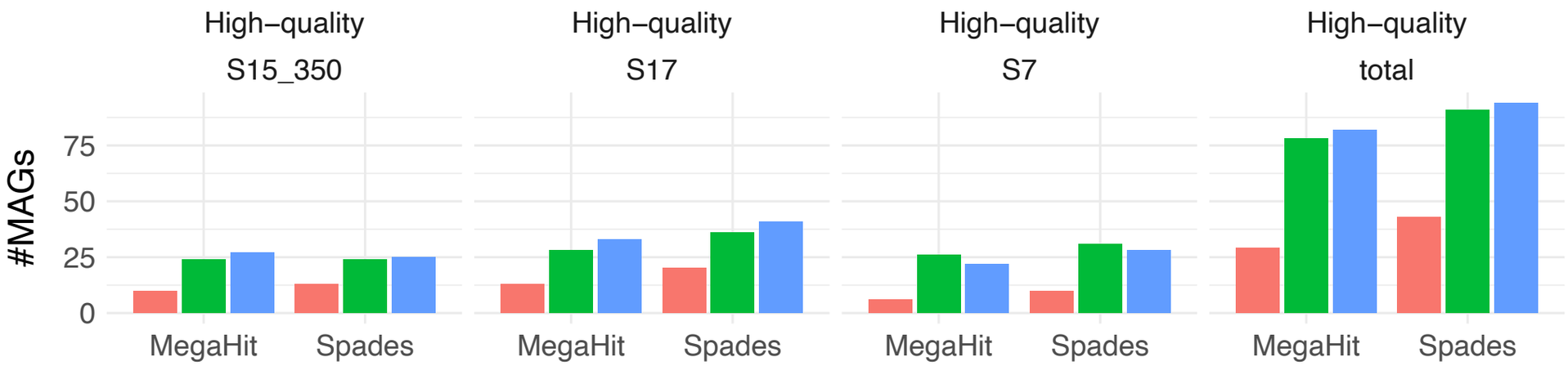
Fragmented assemblies

- > 300k contigs
- N50 < 10kb
- Sizes 400-850Mb



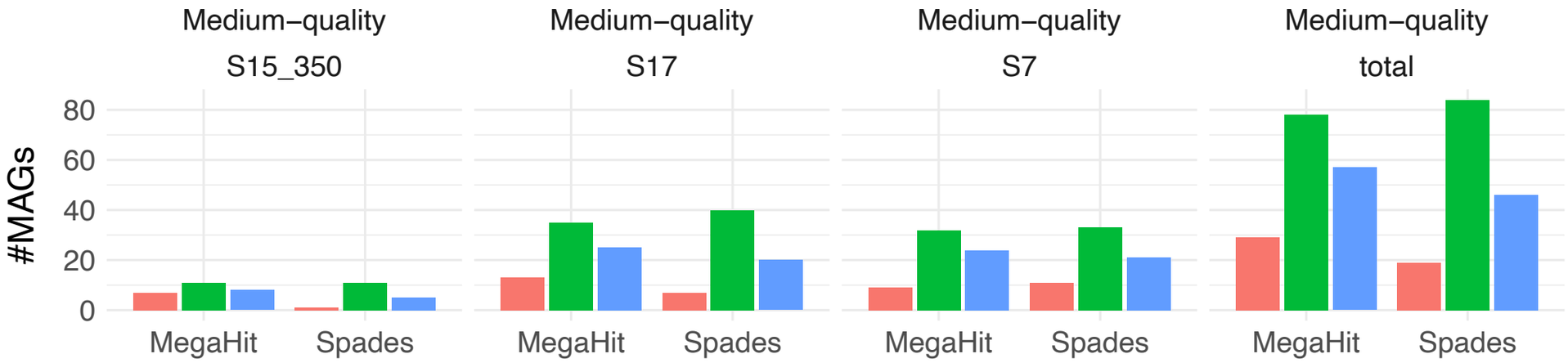
MAG quality

MAG Quality

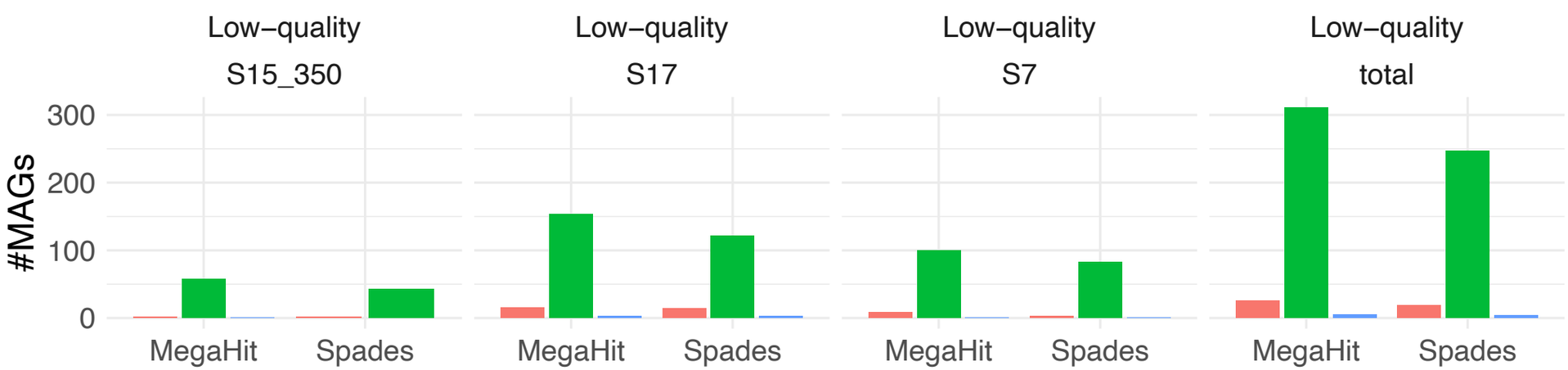


Completeness>=90
Contamination<=5

Presence of the 23S, 16S, and 5S rRNA genes and at least 18 tRNAs.



Completeness>=50
Contamination<=10



Completeness<50
Contamination<=10

binner MaxBin2 MetaBAT2 Refined

Bowers, R., Kyrpides, N., Stepanauskas, R. *et al.* **Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea**. *Nat Biotechnol* **35**, 725–731 (2017). https://doi.org/10.1038/nbt.3893

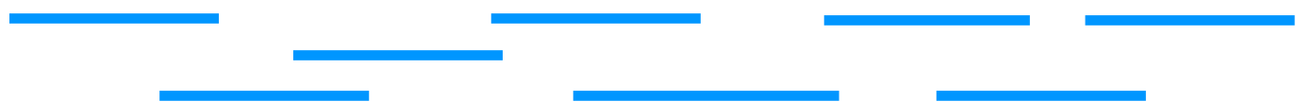
How to improve MAGs quality?

Genome assembly

Genome

Sequencing technology

Reads



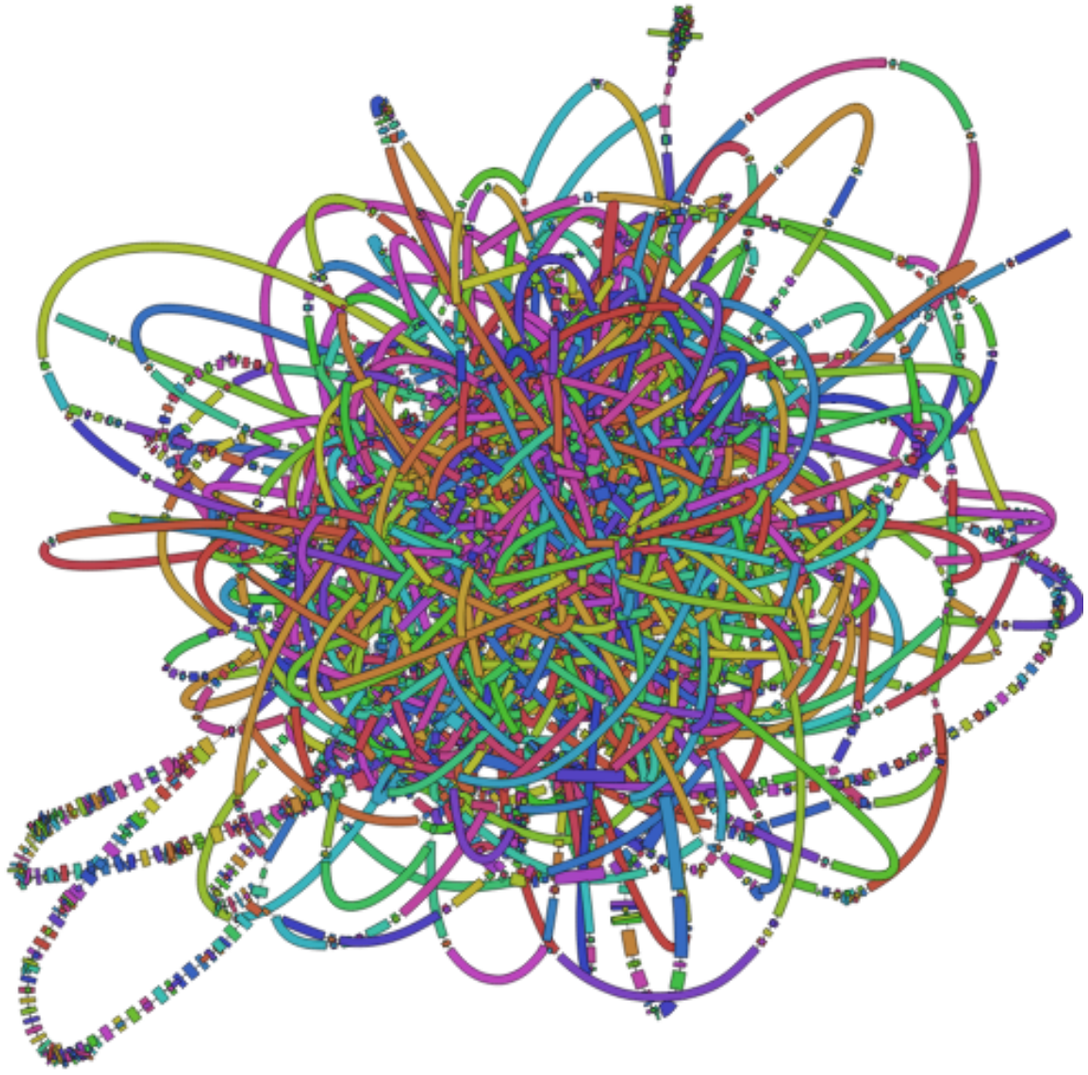
Assembler

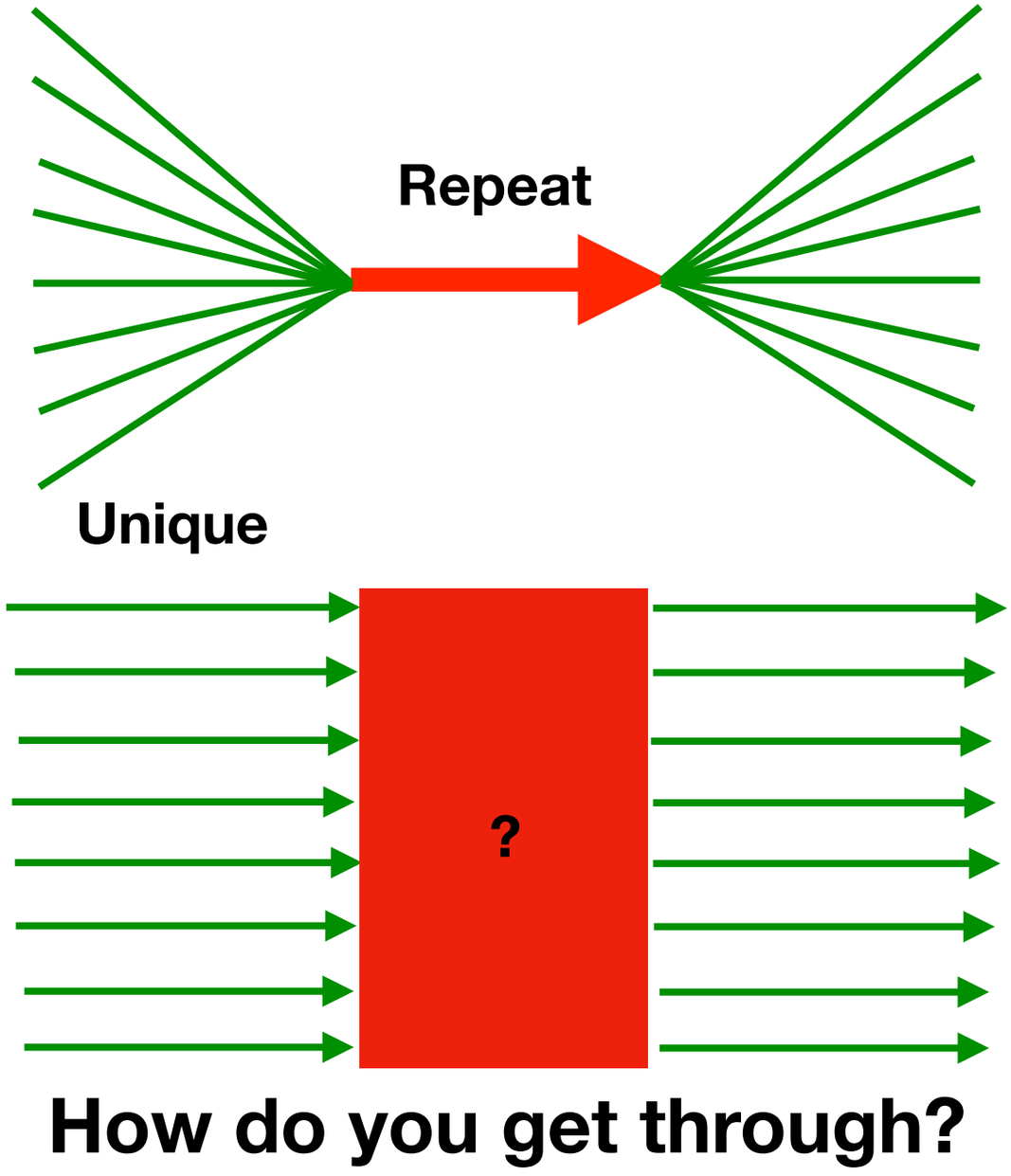
Overlaps

Assembly graph

The genome path (sequence)

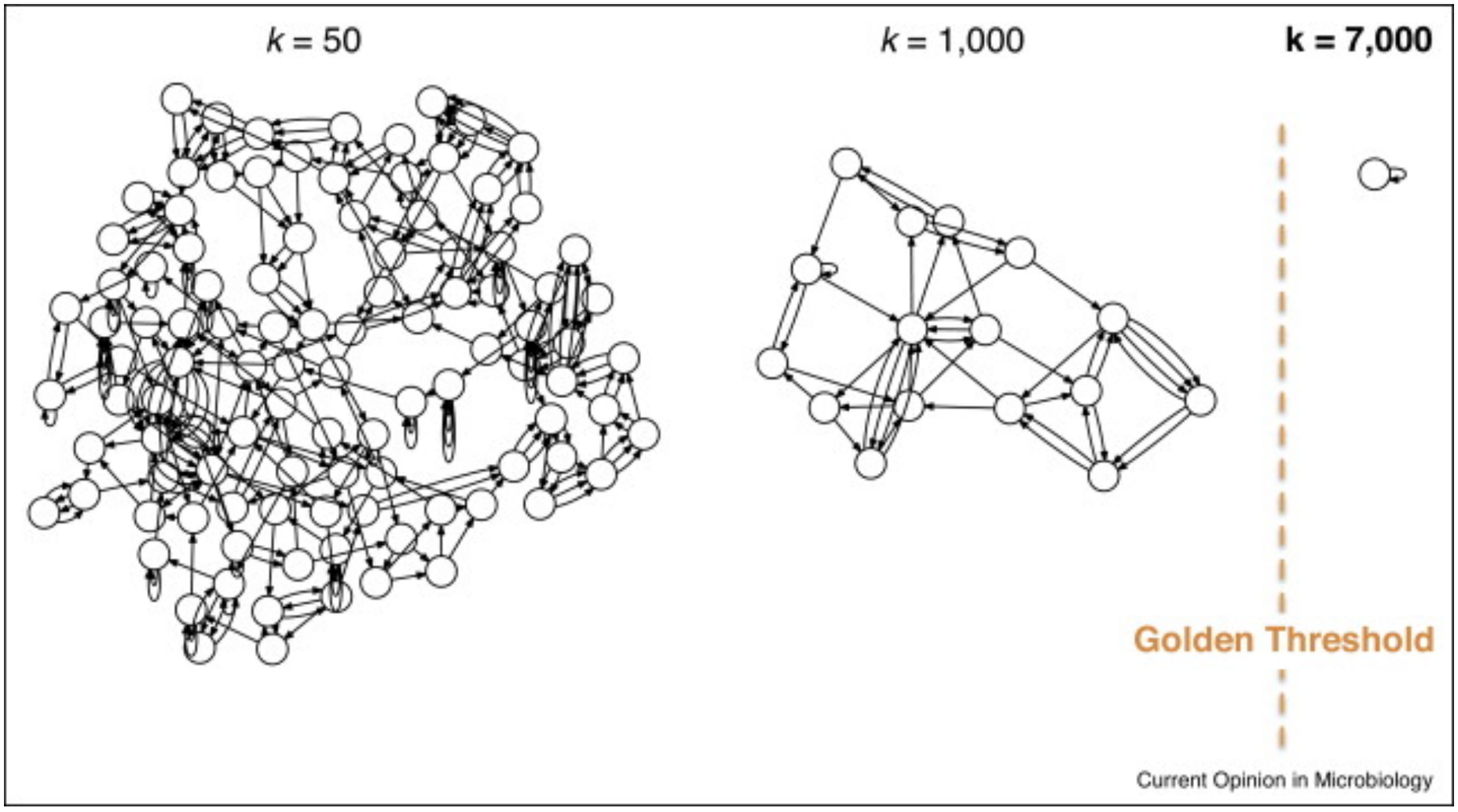
Using short reads there are multiple possible genome assemblies





In practice, it is impossible to find “the” genome path.

Long reads reduce assembly graph complexity



Koren, S., & Phillippy, A. M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current opinion in microbiology*, *23*, 110-120.

Hybrid assembly: How can we combine short and long reads?

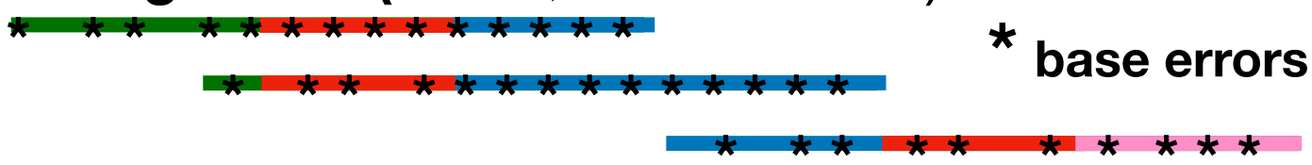
Sequencing technology

Genome

Identical repeats

Short reads (< 300 bp, base error < 0.1%**)**

Long reads (>10 kb, base error <5%)

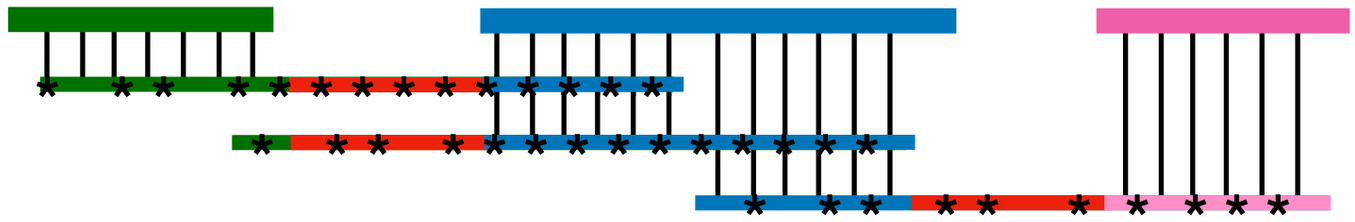


* **base errors**

Wengan Assembler

(1) Assemble short-reads

(2) Scaffold using long reads



(3) Refine repetitive regions (polishing)

The resulting assembly is both *contiguous and accurate*

Wengan: a full hybrid assembler

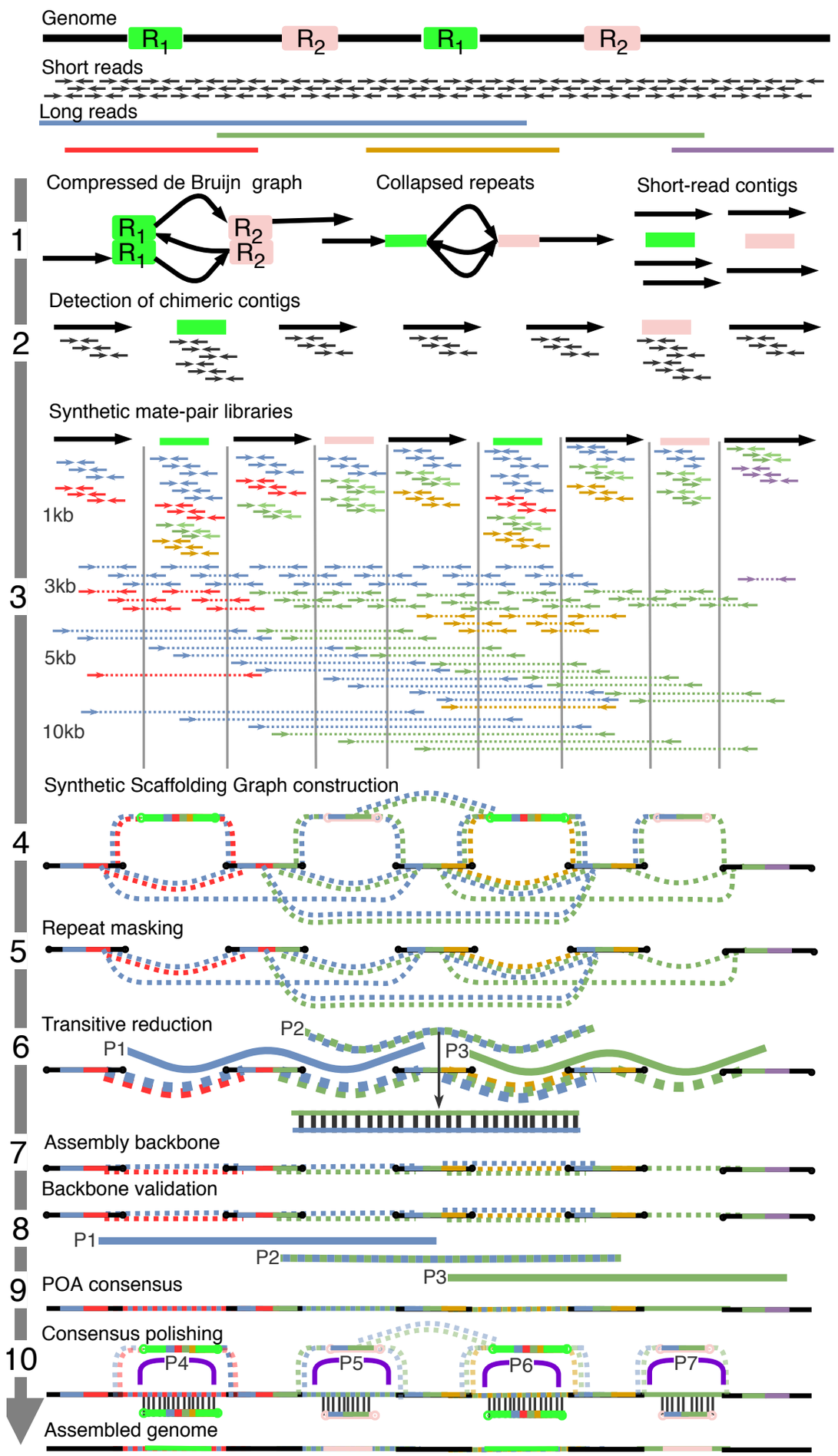
- Avoids entirely all-vs-all read comparisons (**fast**).
- A new assembly graph (*GoogleMaps*).
- 1.5 years of development.
  - ~20k lines of code (C++, PERL)
  - https://github.com/adigenova/wengan

ARTICLES
https://doi.org/10.1038/s41587-020-00747-w
nature biotechnology
Check for updates
OPEN

Efficient hybrid de novo assembly of human genomes with WENGAN

Alex Di Genova[1,2]✉, Elena Buena-Atienza[3,4], Stephan Ossowski[3,4] and Marie-France Sagot[1,2]✉

- **Di Genova, A**. (2018). Fast-SG: an alignment-free. algorithm for hybrid assembly. *GigaScience*, *7*(5).
- **Di Genova, A**. (2021). Wengan: Efficient and high-quality hybrid de novo assembly of human genomes. *Nature Biotechnology*.



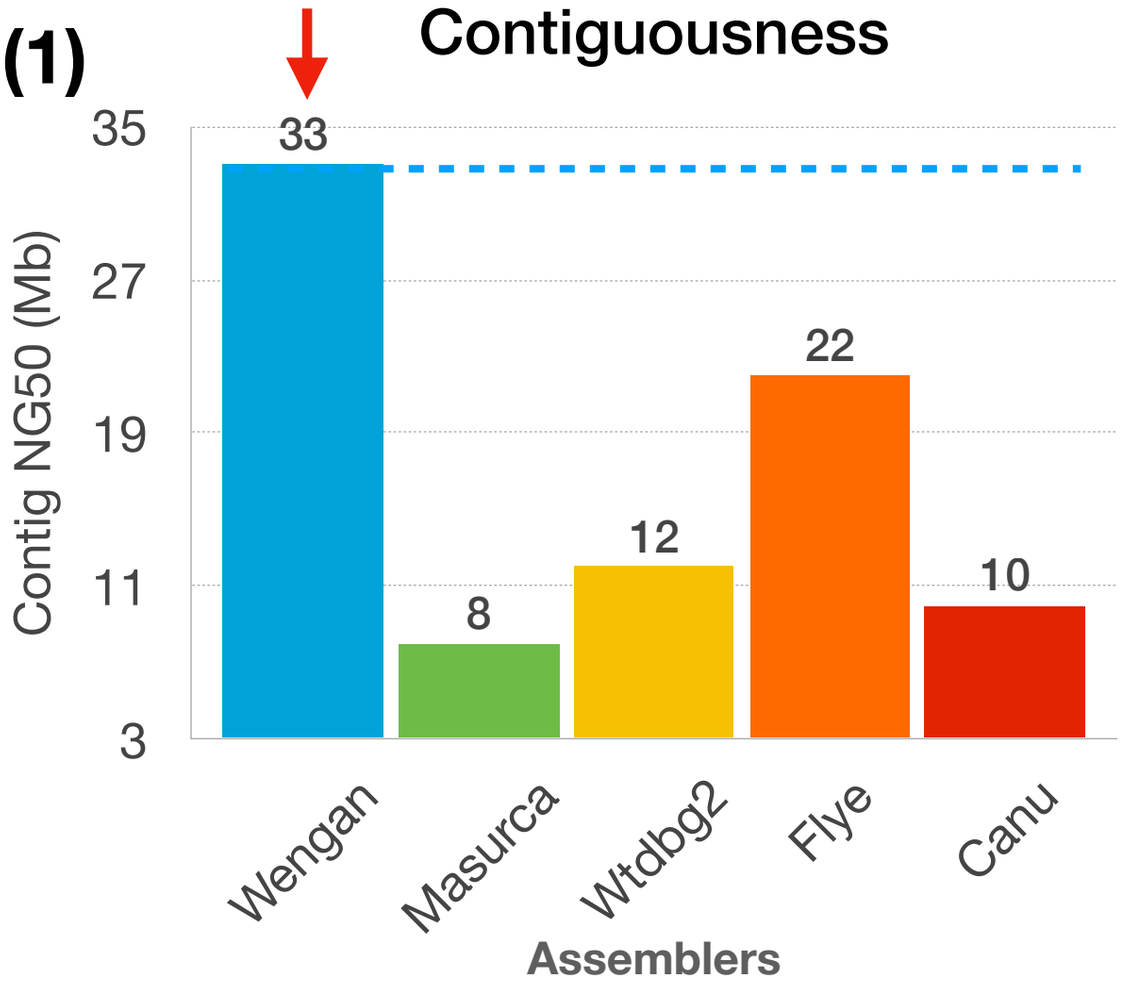
Benchmarking Results

NA12878 data:
- 60X Illumina
- 40X Nanopore
  - **5X > 100kb**

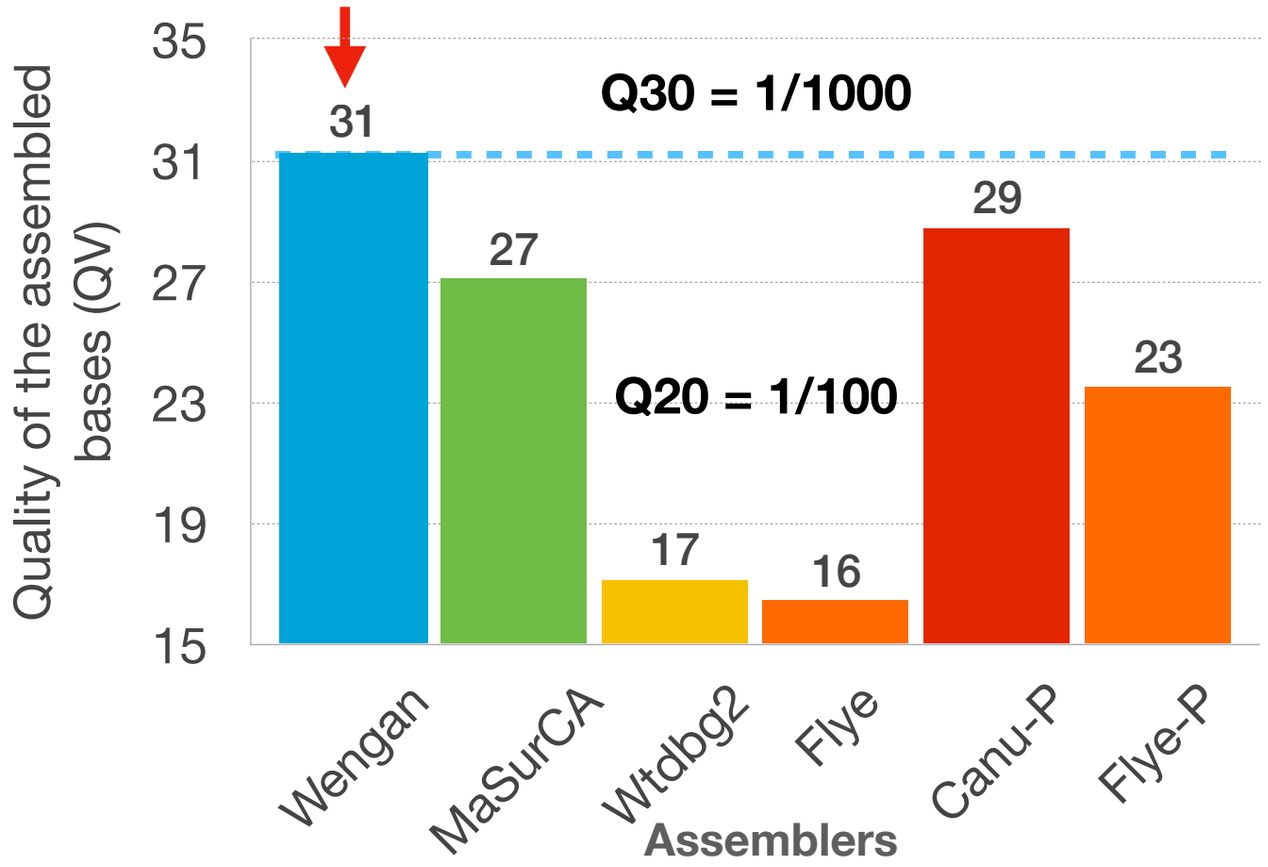
Wengan

NG50: half of reference bases contained in contigs > 33 Mb.

(1) **Contiguousness**



(2) **Consensus quality**



(3) **Computational efficiency**

	Wengan	Masurca	Wtdbg2	Canu	Flye
CPU (h)	**550**	20.000	891	151.000	5.000
Factor	**1**	36	2	275	9

Wengan assembles a human genome in a day

Wengan:
- **High contiguity**
- **High consensus quality**
- **Low computational resources**

Jain, M.(2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. ***Nature biotechnology***, *36*(4), 338.

Is *Wengan ready for* metagenome assembly?

A: No.

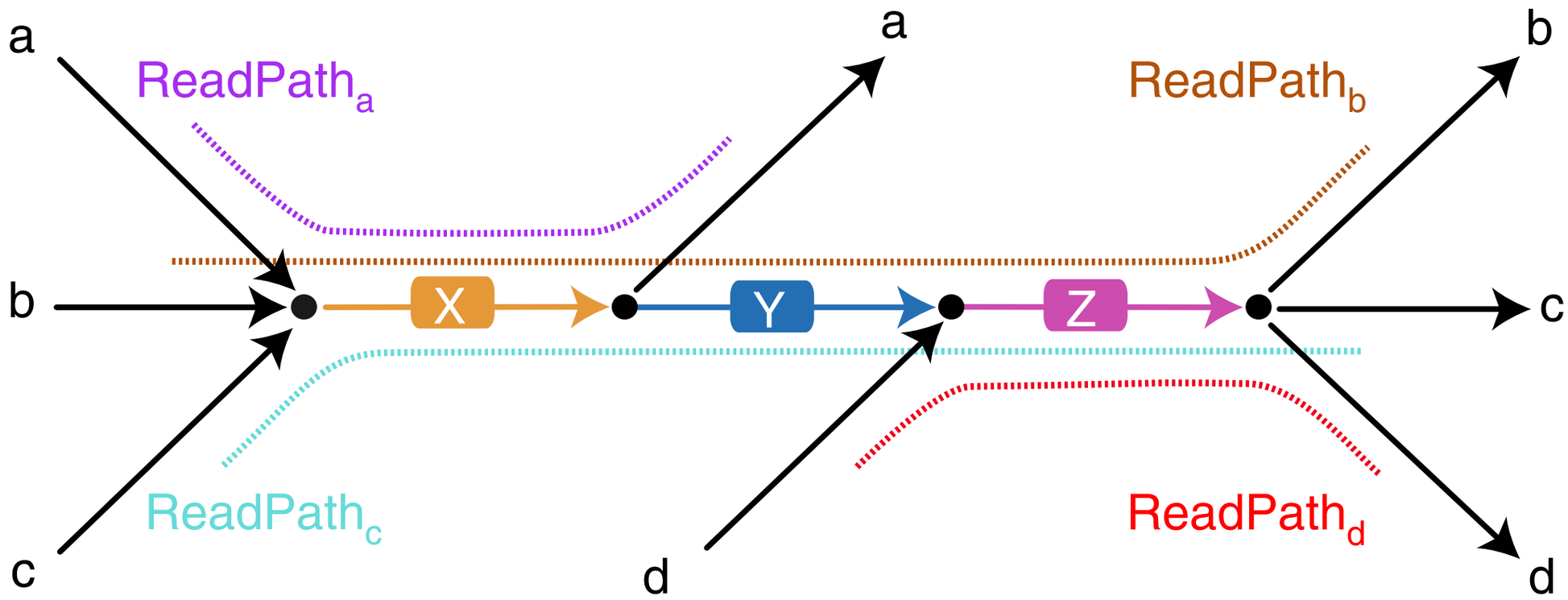
- Some assembly challenge:
  - Heterozygosity and diploid phasing? (trio, graph topologies)
- Metagenomes:
  - Repeat algorithms (coverage is not uniform)
  - binning (integration with the **assembly graph**)
  - Low abundance species (**target sequencing**)
  - Strain-level deconvolution(phasing)

Moises Rojas
Postdoc

From Genome Assembler to Metagenome Assembler

- Flye -> MetaFlye

Identifying repeats

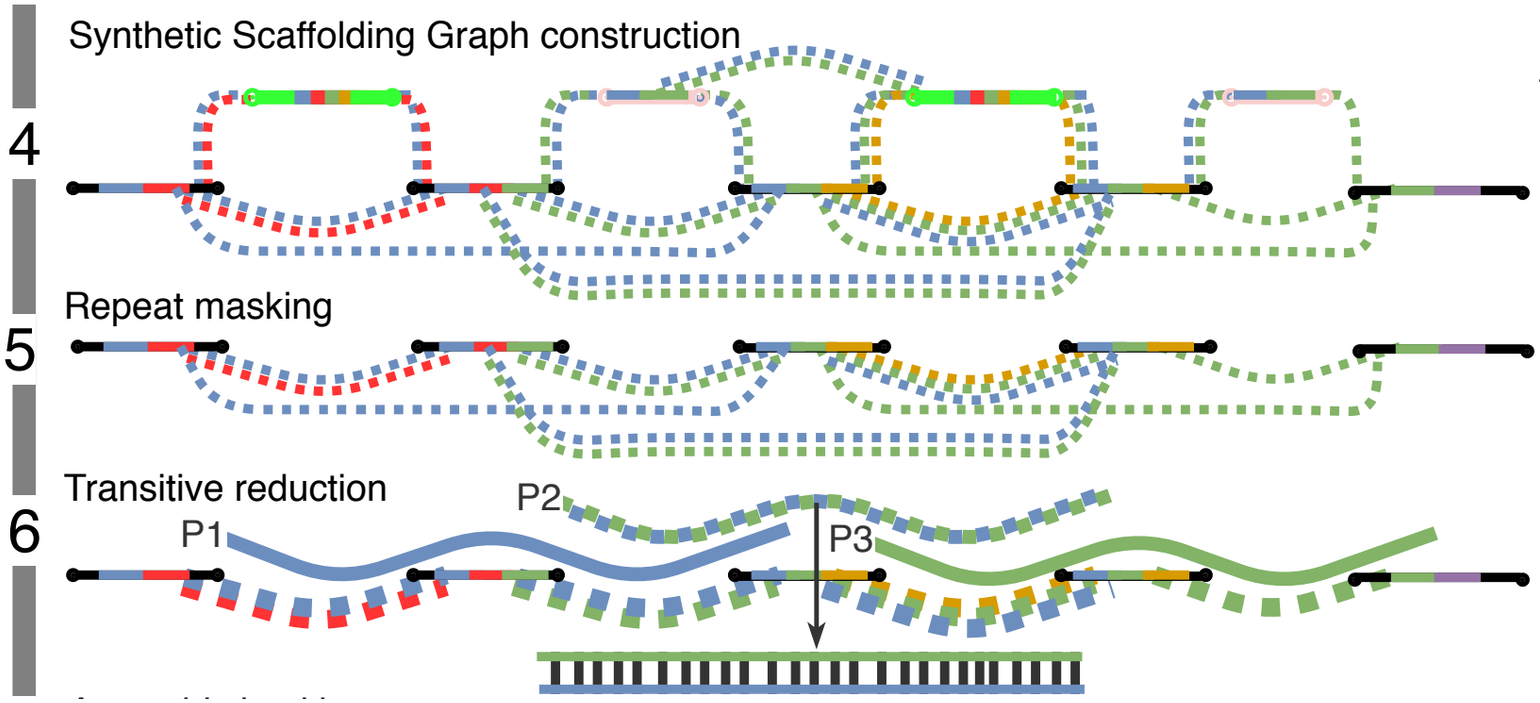


- Hifiasm -> Hifiasm-meta

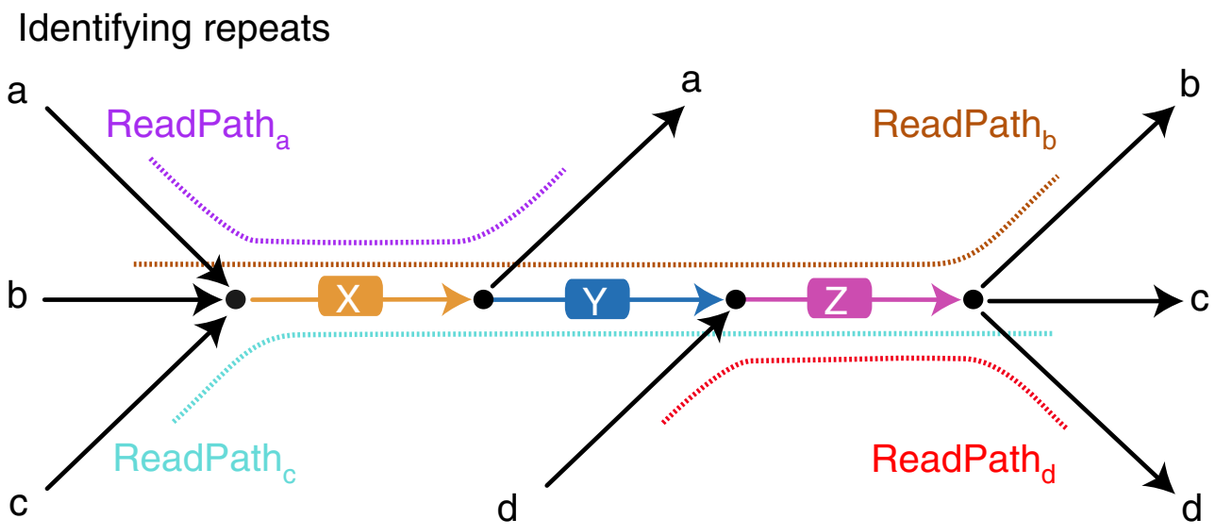
Kolmogorov, M., Bickhart, D.M., Behsaz, B. *et al.* **metaFlye: scalable long-read metagenome assembly using repeat graphs**. *Nat Methods* **17**, 1103–1110 (2020).

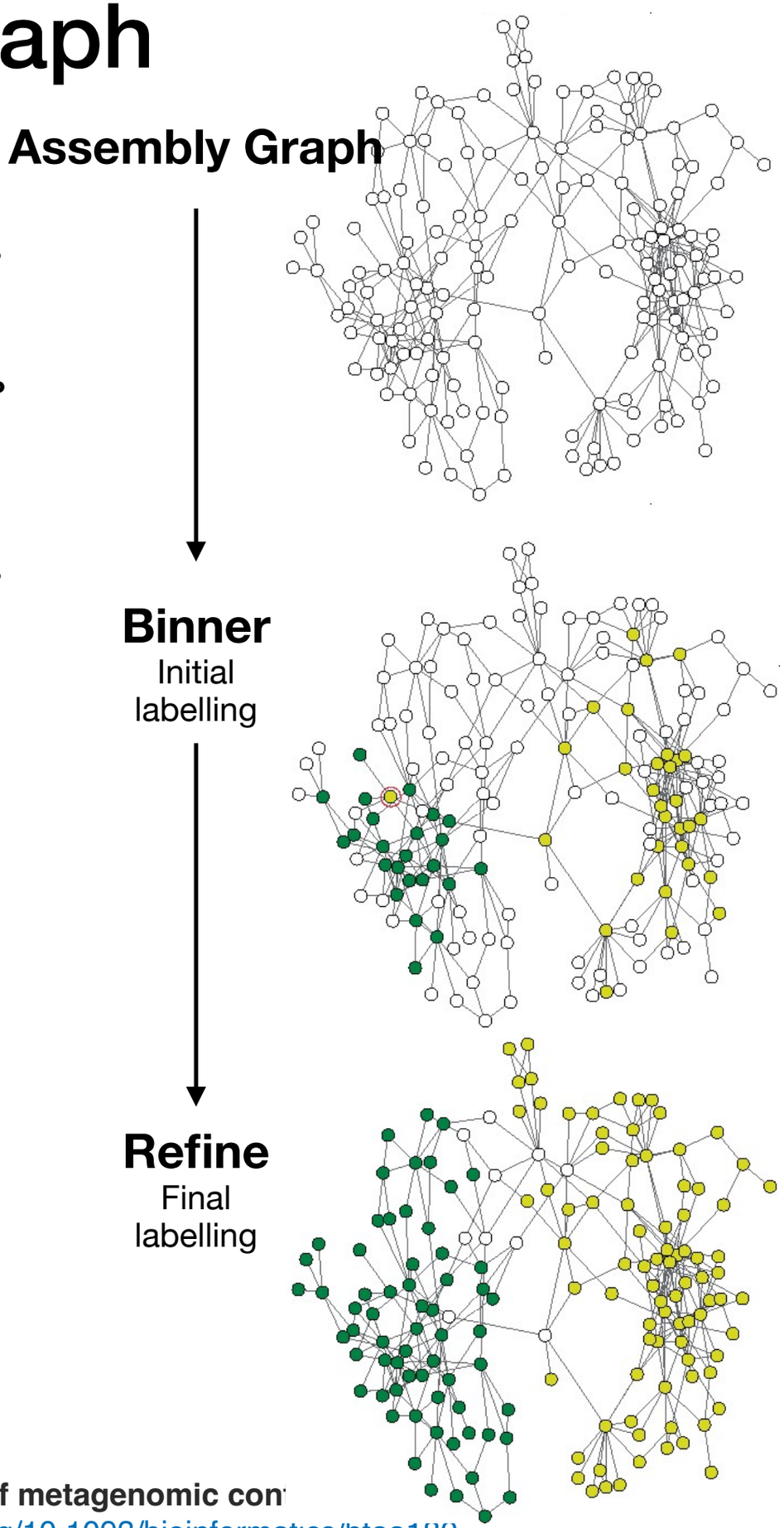
Feng, X., Cheng, H., Portik, D. *et al.* **Metagenome assembly of high-fidelity long reads with hifiasm-meta.** *Nat Methods* **19**, 671–674 (2022).

Binning: Wengan Assembly Graph



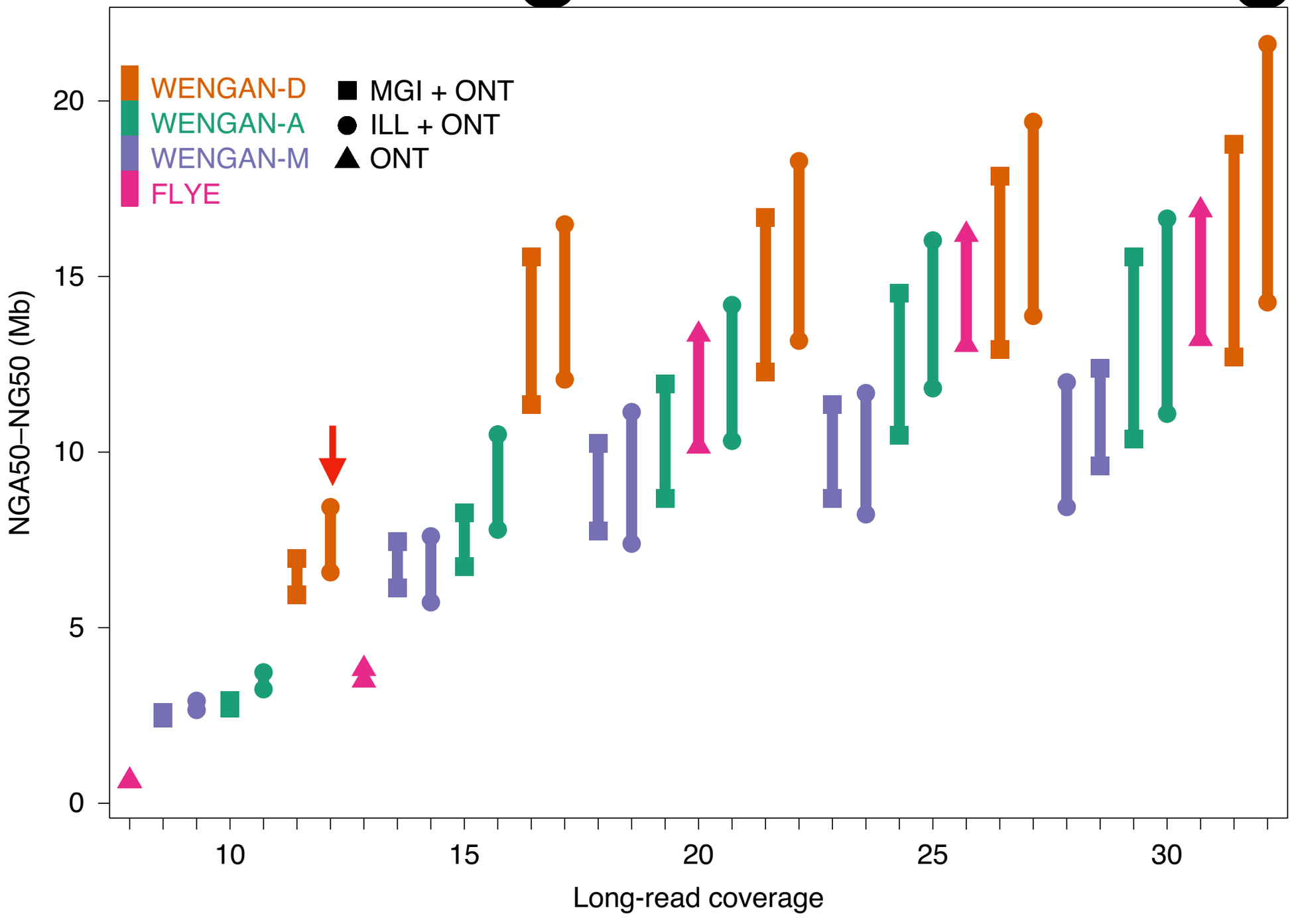
We compute approximate long-read overlaps using synthetic pair-ends as elemental building blocks plus long-read coherent path search on the SSG.





Vijini Mallawaarachchi, Anuradha Wickramarachchi, Yu Lin, **GraphBin: refined binning of metagenomic con**
graphs, *Bioinformatics*, Volume 36, Issue 11, June 2020, Pages 3307–3313, https://doi.org/10.1093/bioinformatics/btaa180

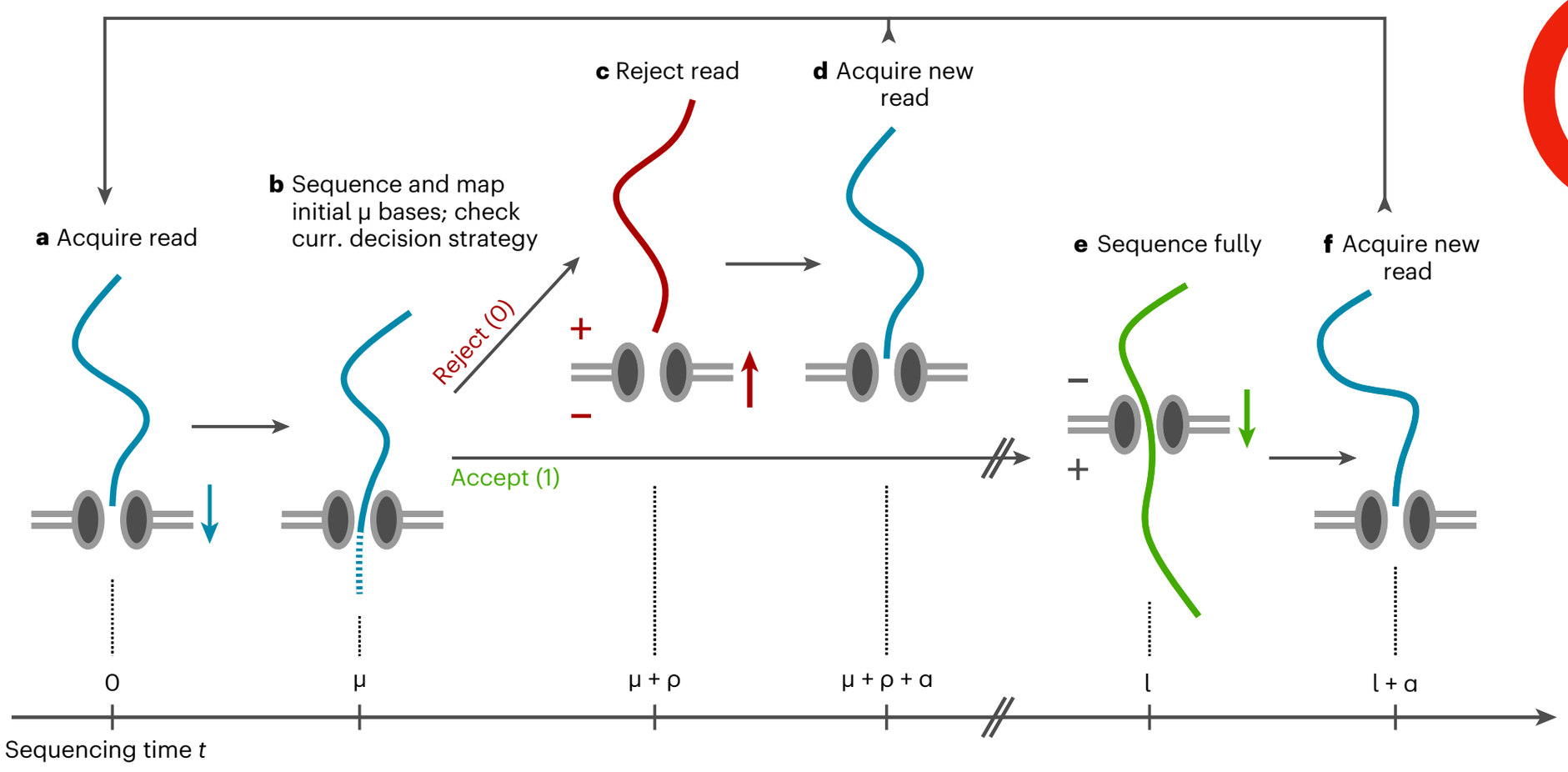
LAM: Wengan is effective at shallow long-read coverage

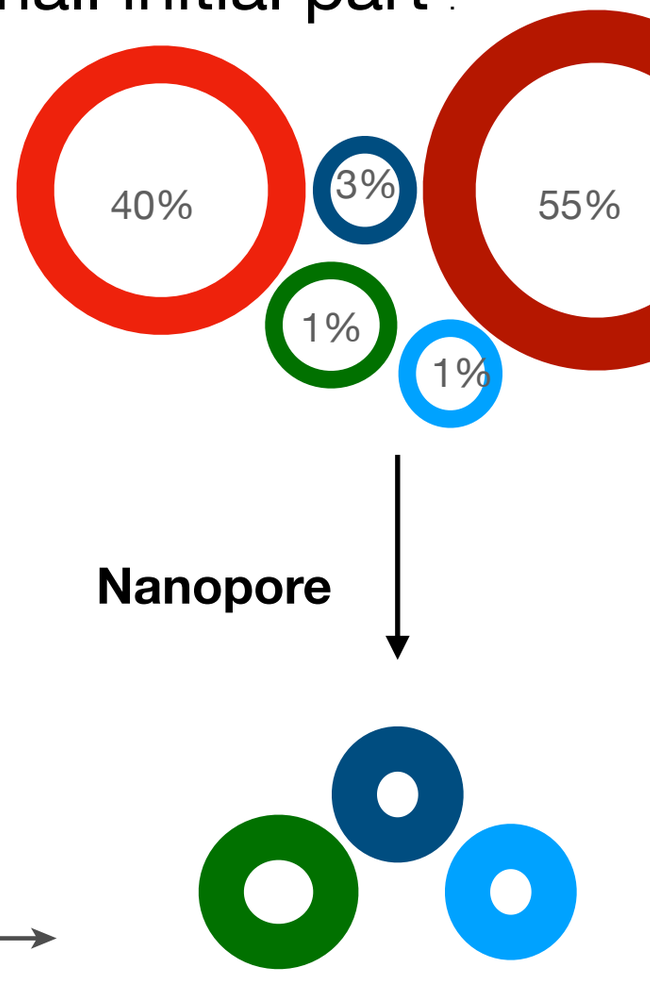


- **10X** of long-read coverage: NG50 5~10Mb

LAM : Read enrichment in real-time.

- Nanopore sequencers can select which DNA molecules to sequence, rejecting a molecule after analysis of a small initial part .





Weilguny, L., De Maio, N., Munro, R. *et al.* **Dynamic, adaptive sampling during nanopore sequencing using Bayesian experimental design.** *Nat Biotechnol* (2023).

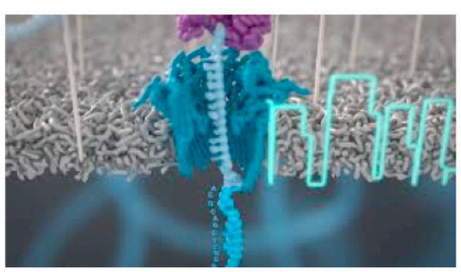
Payne, A., Holmes, N., Clarke, T. *et al.* **Readfish enables targeted nanopore sequencing of gigabase-sized genomes.** *Nat Biotechnol* **39**, 442–450 (2021).

UOH sequencing and HPC platforms (2023)

Universidad de O'Higgins is just 7 years old.

200 Human genomes per year

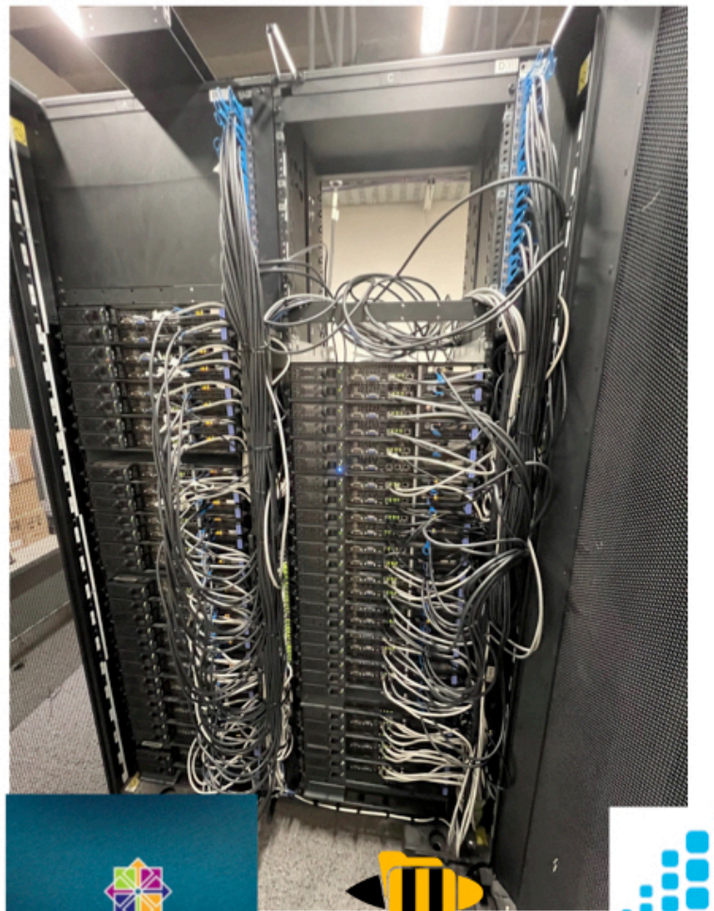
April 2023

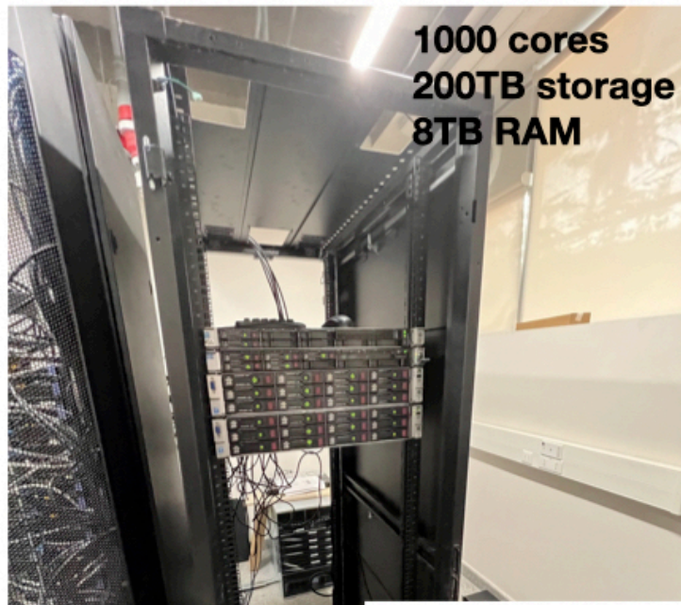


Oxford NANOPORE Technologies

Universidad de O'Higgins

HPC-UOH





CentOS 7

BeeGFS®

slurm workload manager

ANSIBLE

Universidad de O'Higgins

digenoma lab

Algorithms and theoretical analyses for understanding complex biological systems (Cancer)

Omics data

+

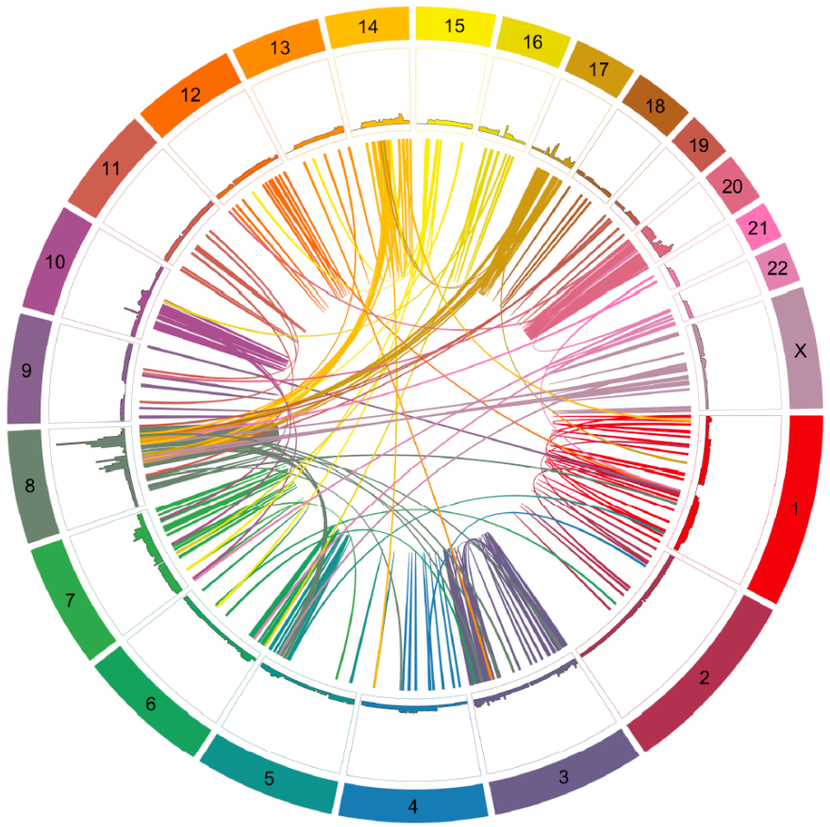
Novel algorithms

+

Genomic heritage (Chile)

+

Students from different backgrounds (Biology, Mathematics, Bio-Informatics, Medicine)



https://digenoma-lab.cl/

digenova@gmail.com
alex.digenova@uoh.cl

Acknowledges

International collaborators

National collaborators

- **Juan Francisco Miquel (UC)**
- **Luis Zapata (ICR)**
- **Mauricio Latorre (UOH)**
- **Jose Manuel Yañez (U.Chile)**
- **…**

Thank you!