JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

Large Scale Metagenome Assembly Using MetaHipMer (MHM2)

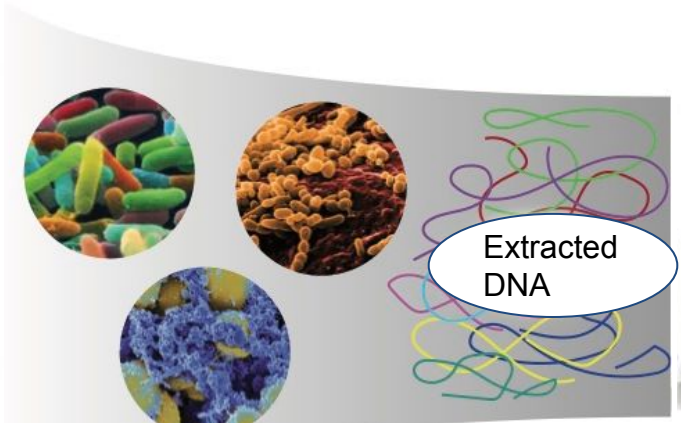
2024 JGI User Meeting
Large Scalable Metagenomics Toolbox Workshop

Rob Egan
04 October 2024

Metagenomics overview

Environments

Microbial communities underlie biogeochemical processes



Extracted DNA

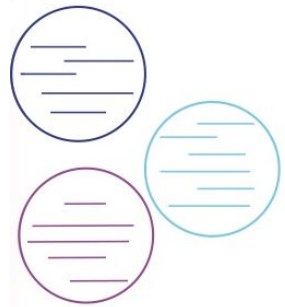
Sequence

Petabases (1×10^{15} base pairs) of metagenome data are an opportunity to characterize environmental microbial communities

Gigabases to terabases of reads

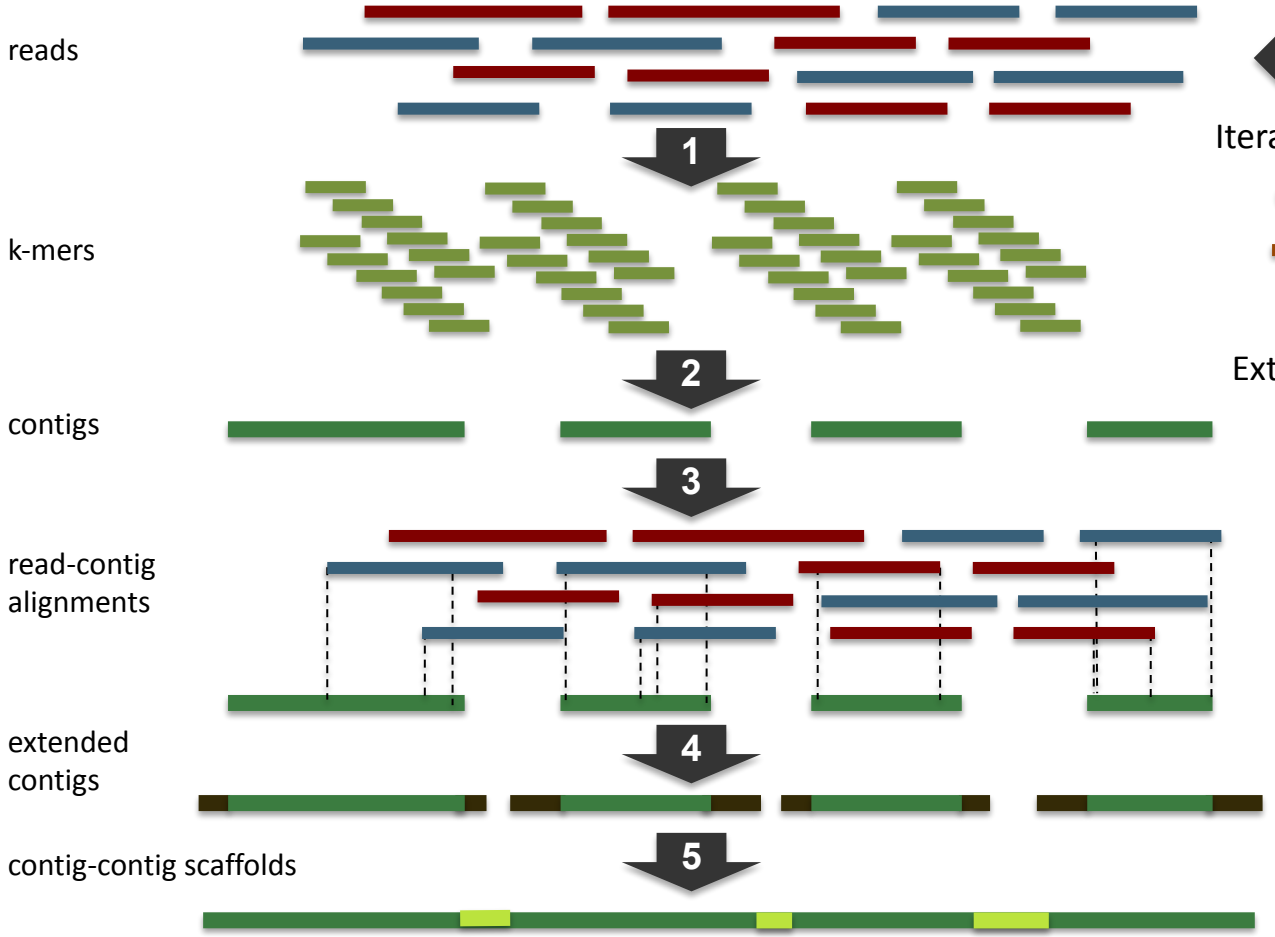
Assemble

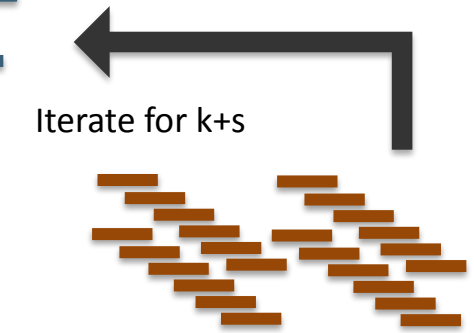
Metagenome-assembled genomes (MAGs)



MetaHipMer (MHM2) via the ExaBiome Project
an Exascale Computing Project (ECP)

MetaHipMer Assembly Pipeline





1) K-mer Analysis (GPU)
K-mer histogram

2) Contig Generation
Walk k-mer graph

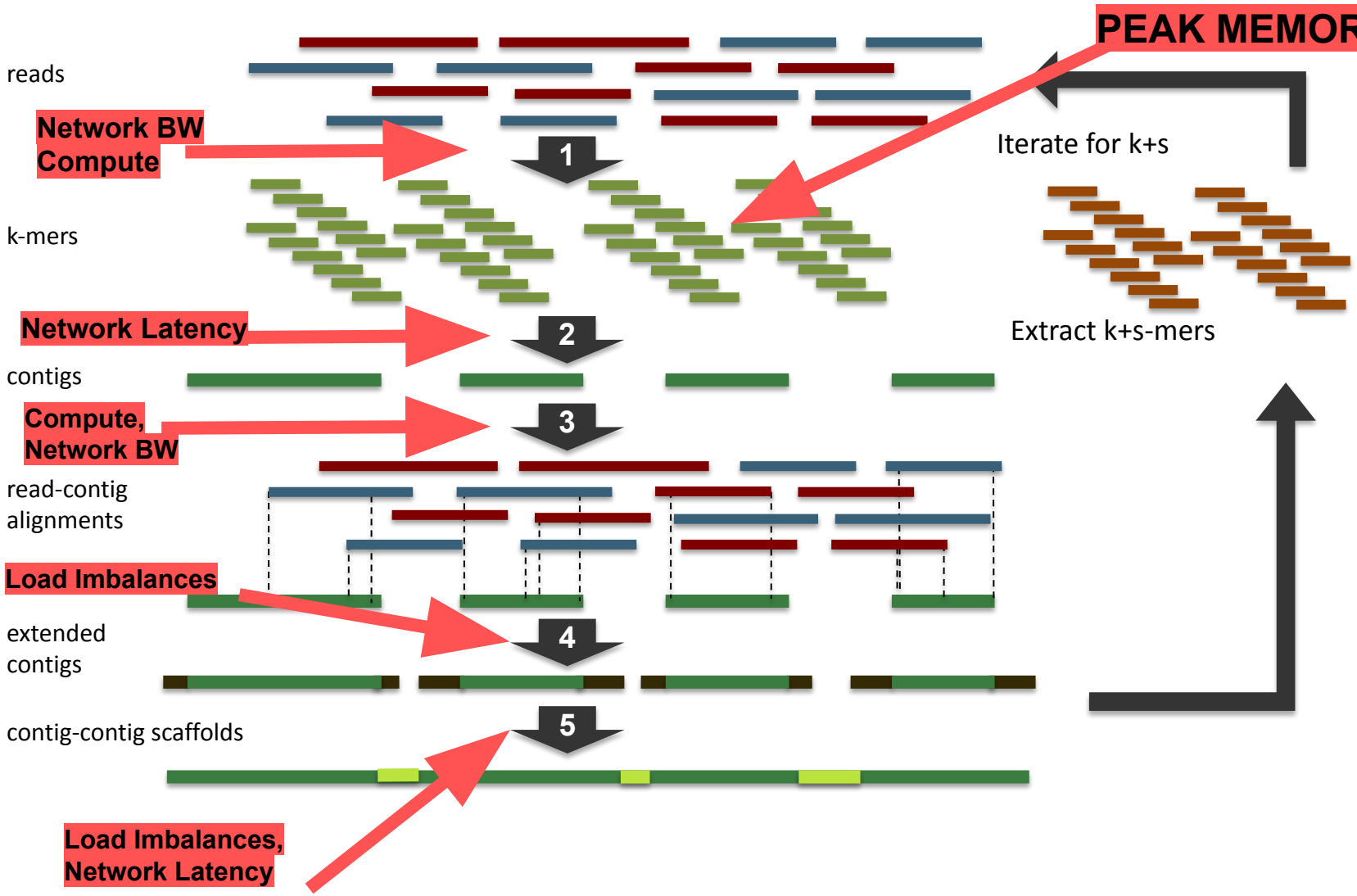
3) Alignment (GPU)
Align reads to contigs

4) Local Assembly (GPU)
Extend ends of contigs

5) Scaffolding
Walk contig graph (iterate)

Actual pipeline is more complex, simplified for purpose of presentation

MetaHipMer Assembly Pipeline



1) K-mer Analysis (GPU)
K-mer histogram

2) Contig Generation
Walk k-mer graph

3) Alignment (GPU)
Align reads to contigs

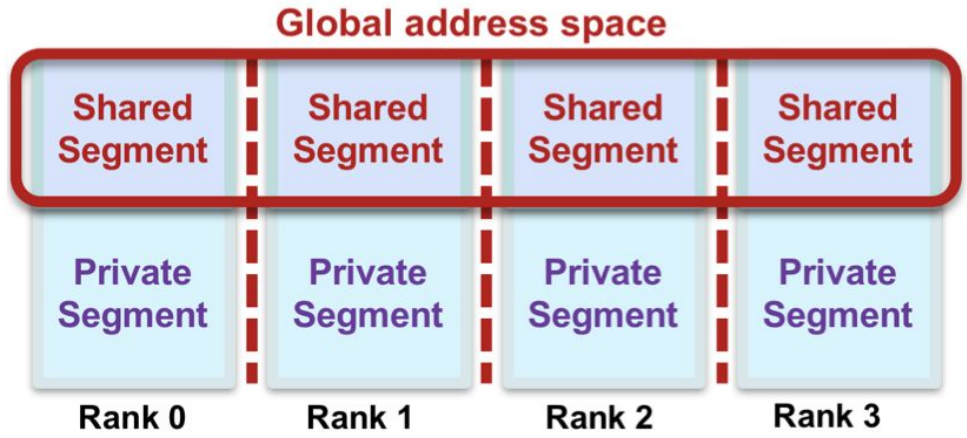
4) Local Assembly (GPU)
Extend ends of contigs

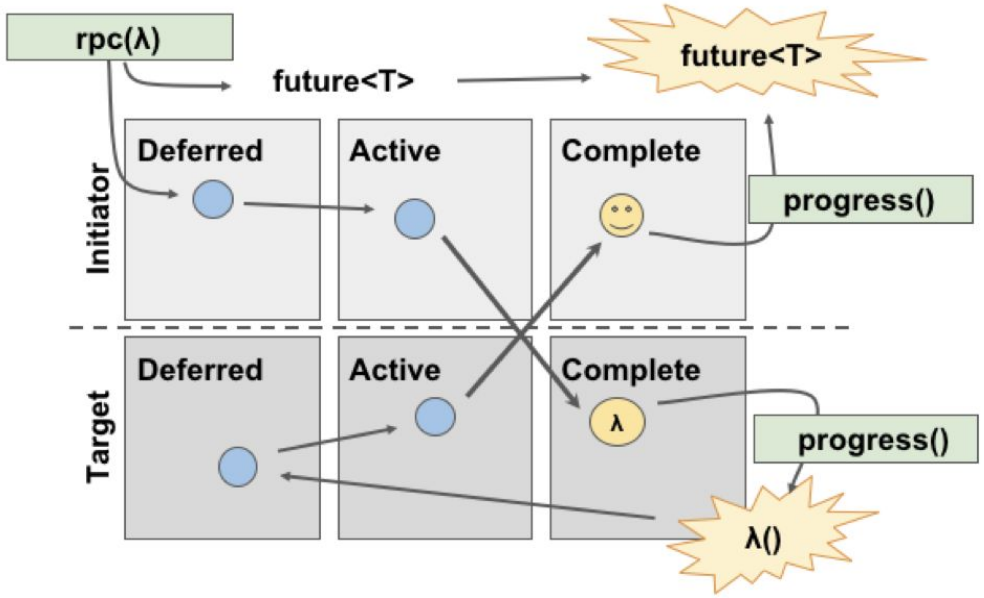
5) Scaffolding
Walk contig graph (iterate)

Actual pipeline is more complex, simplified for purpose of presentation

Built on UPC++ and GASNET

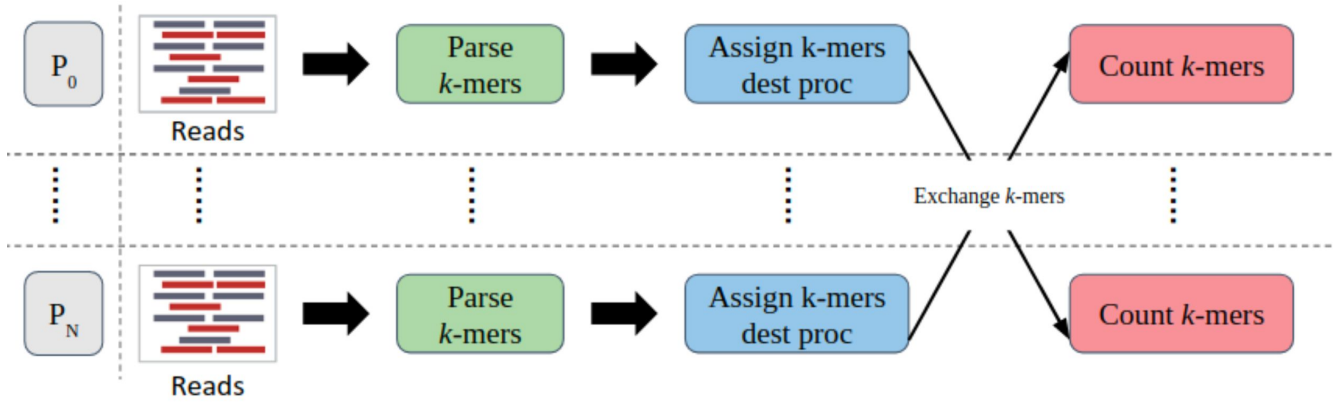
- **Distributed Programming**
  - Collective Reduction, Barriers
  - 1-Sided Gets / Puts
- **Asynchronous and Deferred Execution**
- **Remote Procedure Calls**
  - 1 Sided & Round Trip
  - C++ Lambda Syntax
- **C++ Library / Framework**
  - Built on GASNET
- **Simpler, Smaller & Faster Code**

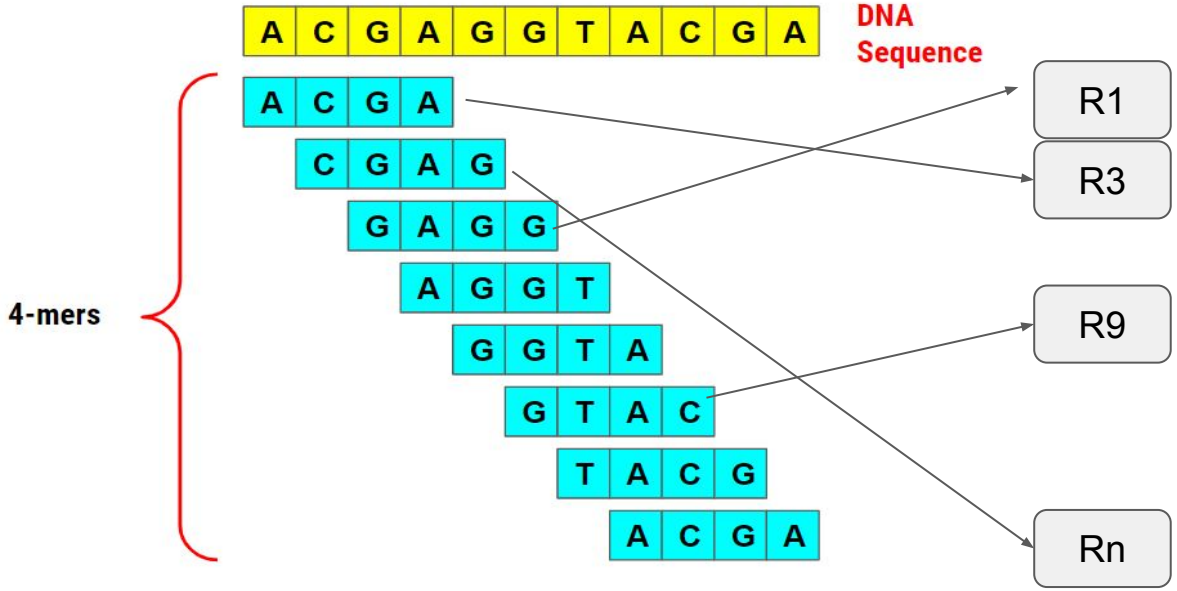


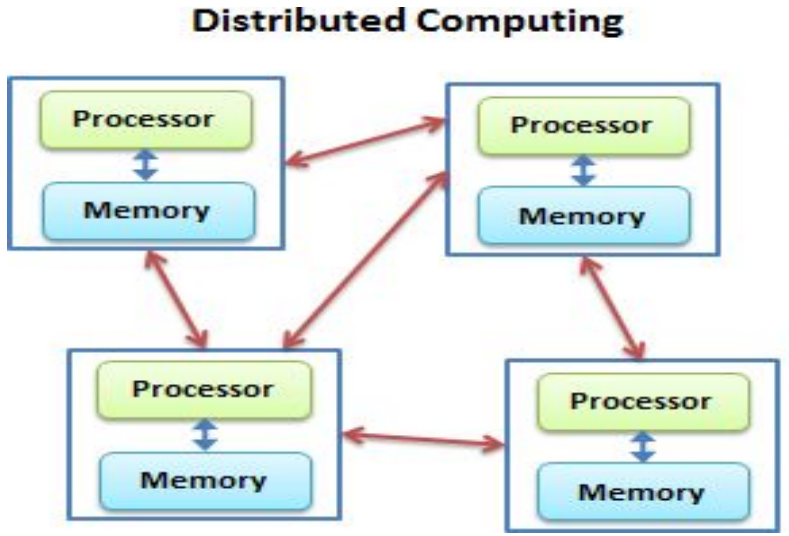


Distributed Hash Tables
Enabled by High Speed Networking

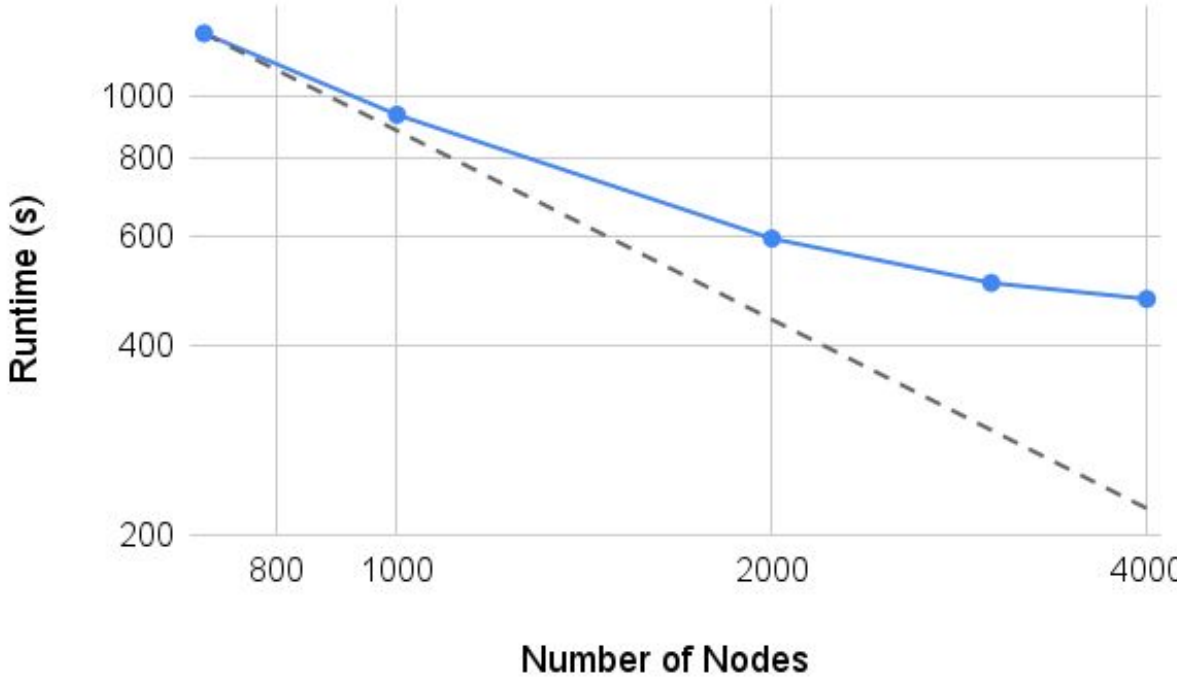
- **Assembly is Limited by Available RAM**
  - Store Data Across All Ranks
- **Distributed Hash Table**
  - High Bits of Hash -> Rank
  - Low Bits of Hash -> Local Bucket

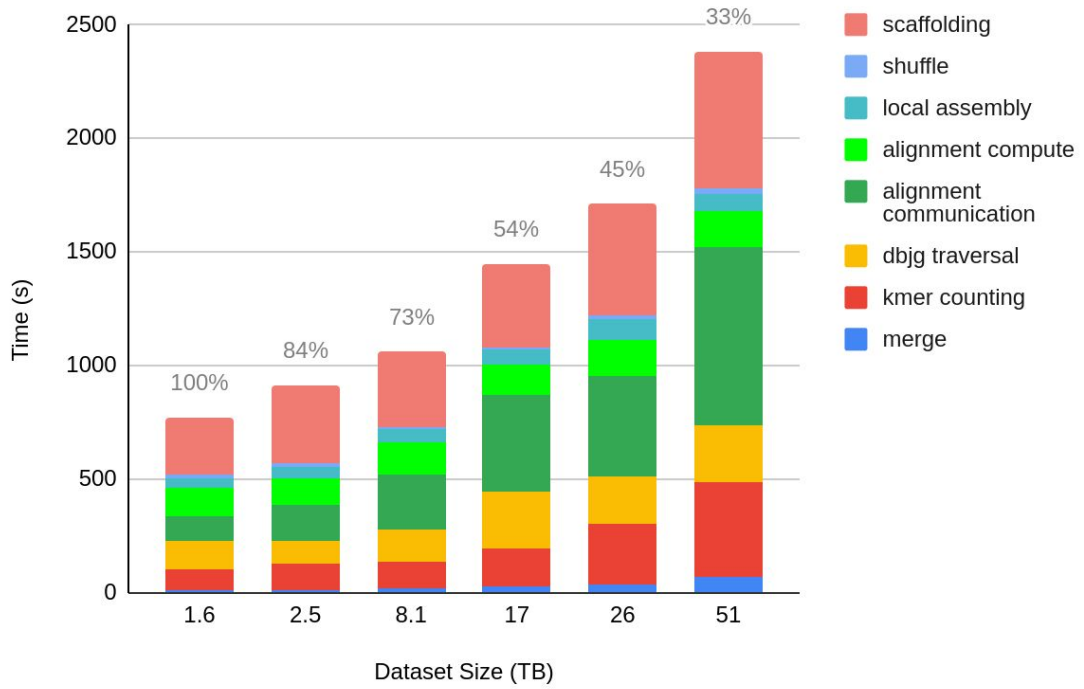






MetaHipMer Performance





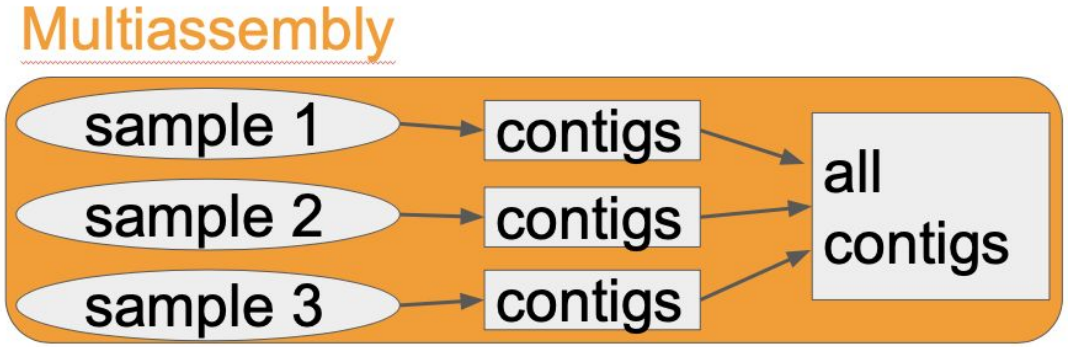
Strong scaling

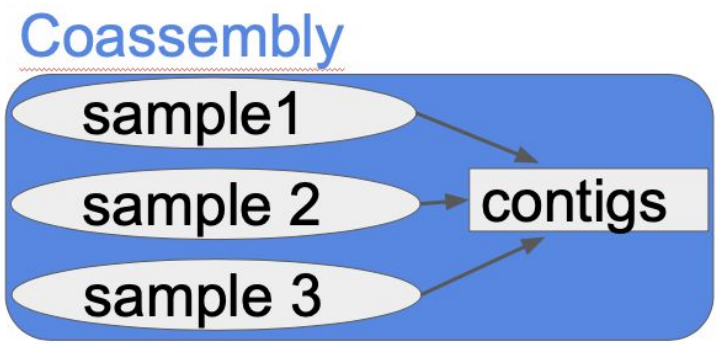
8TB Indian Ocean subset of Tara Oceans

Weak scaling of stages

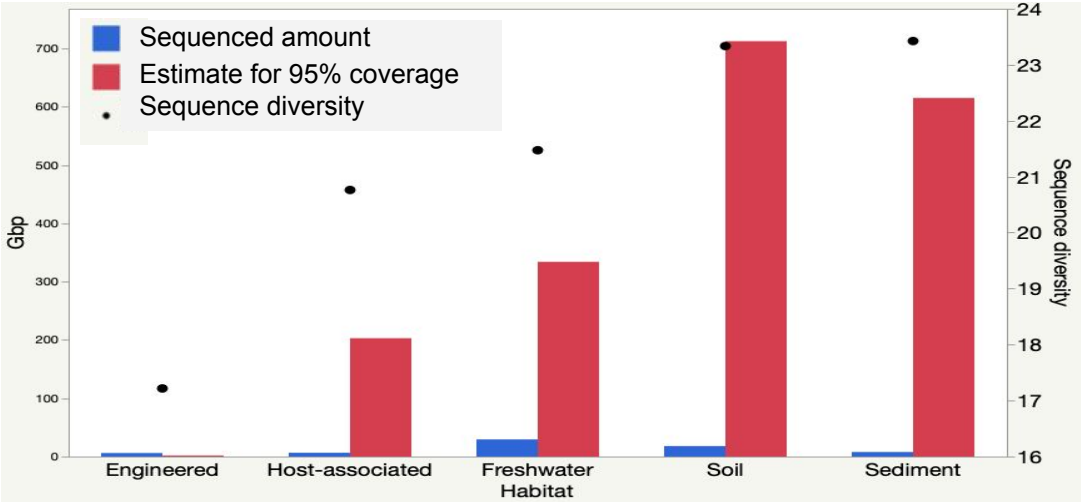
Increasing sizes of Tara Oceans. Efficiency from 200 nodes to 6400 nodes is 33%

MetaHipMer enables coassembly across samples

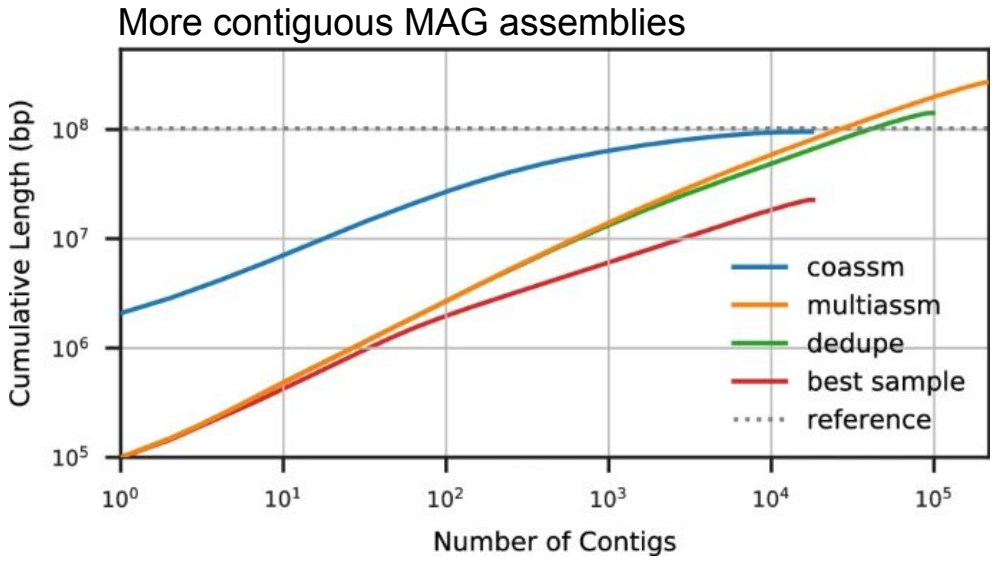




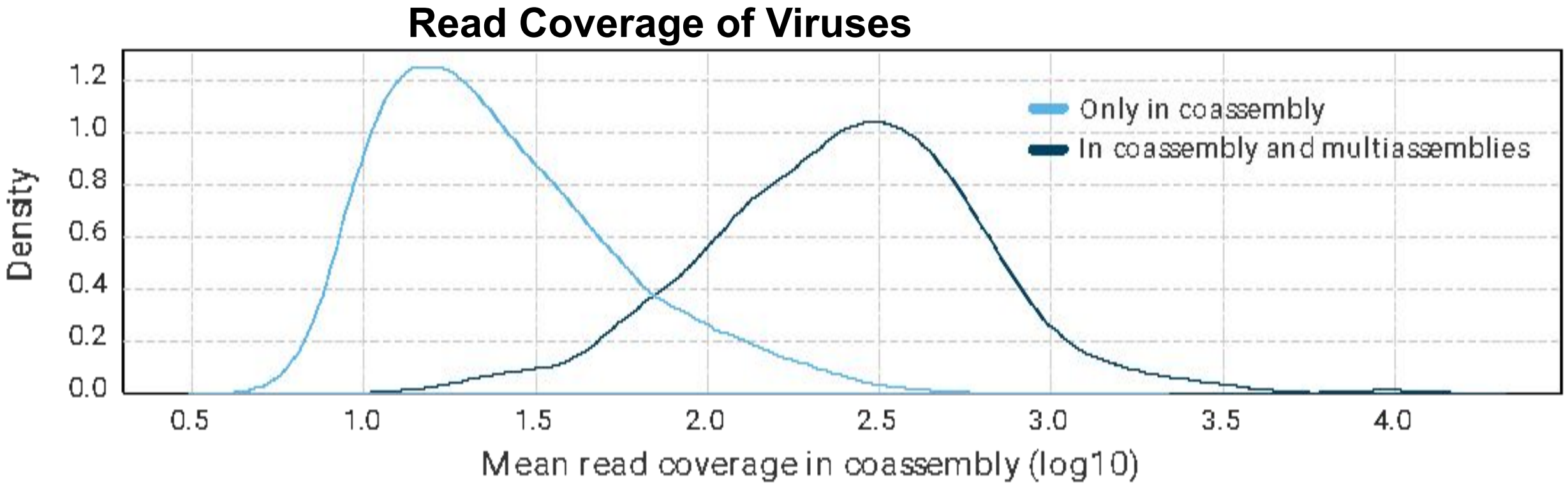
Sequencing depth is limited by cost; assemblies limited by shared memory size



Co-Assembly combines samples from time or space to improve quality



More Viruses in the GRE Coassembly

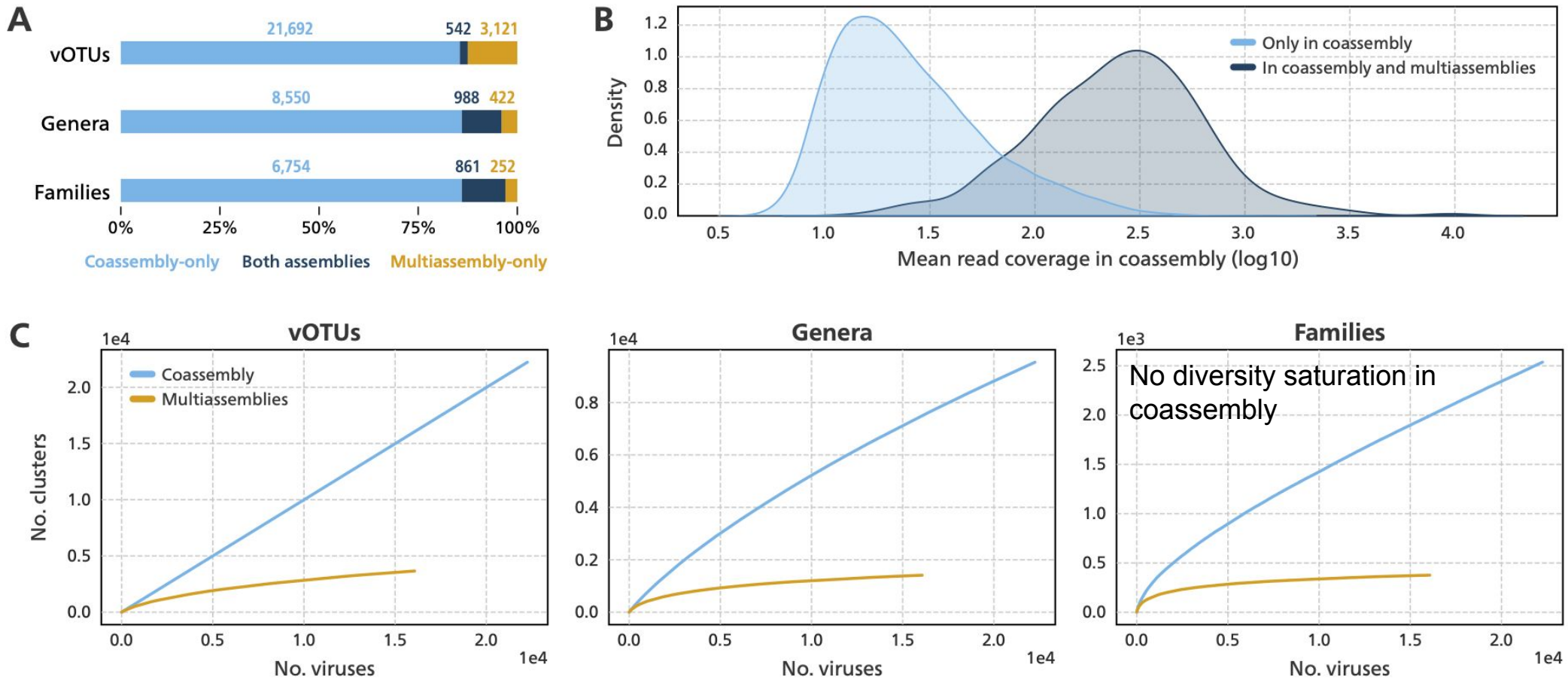


- **Low abundance viruses recovered only in coassembly**
- **Coassembly has more fungi, more eukaryotes, more rare biosphere microbes**

Riley et al. (2023). Terabase-scale coassembly of a tropical soil microbiome. *Microbiology Spectrum*

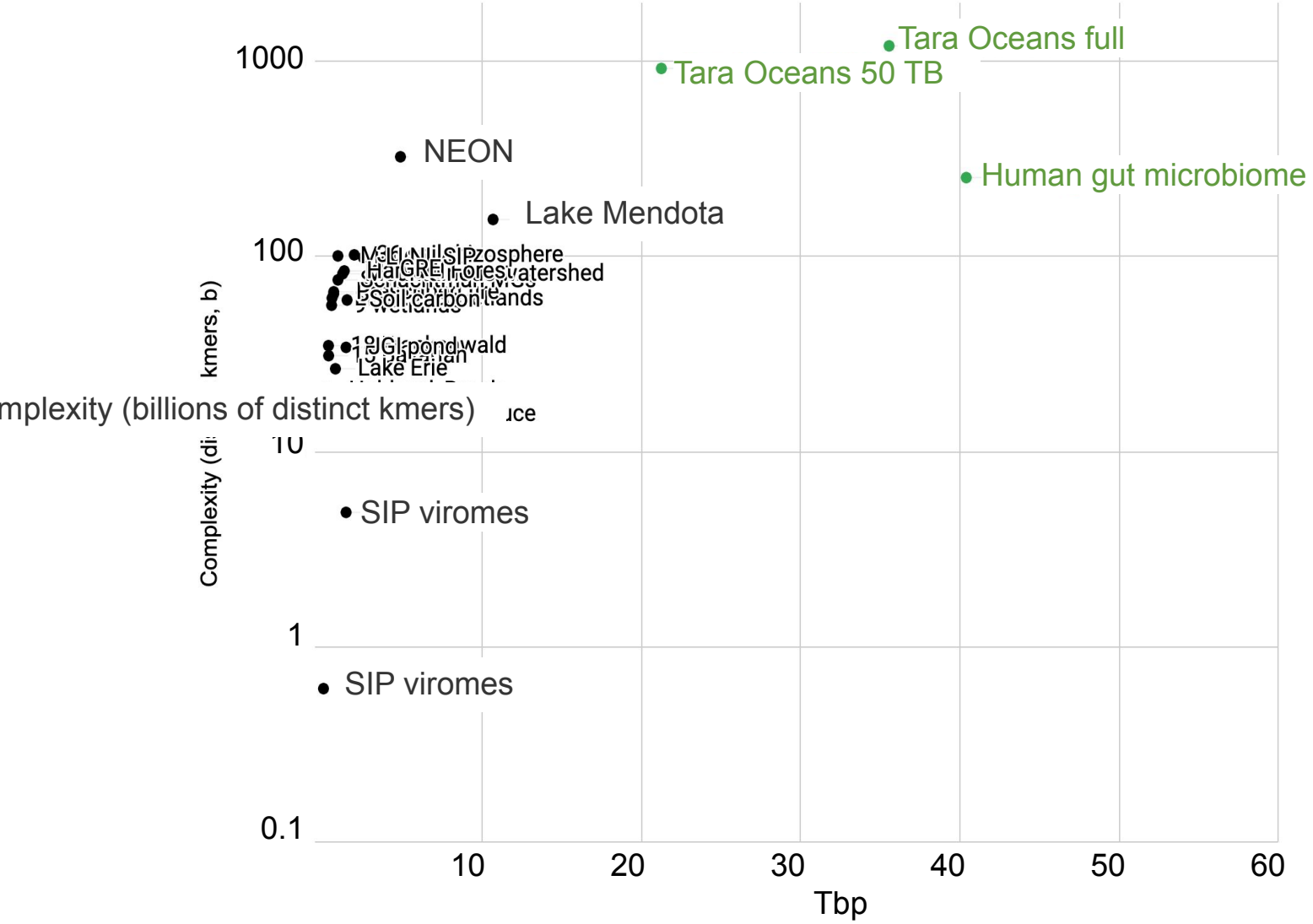
More viruses in the GRE coassembly

22,254 viral sequences were identified using the geNomad pipeline: Camargo AP et al. bioRxiv 2023



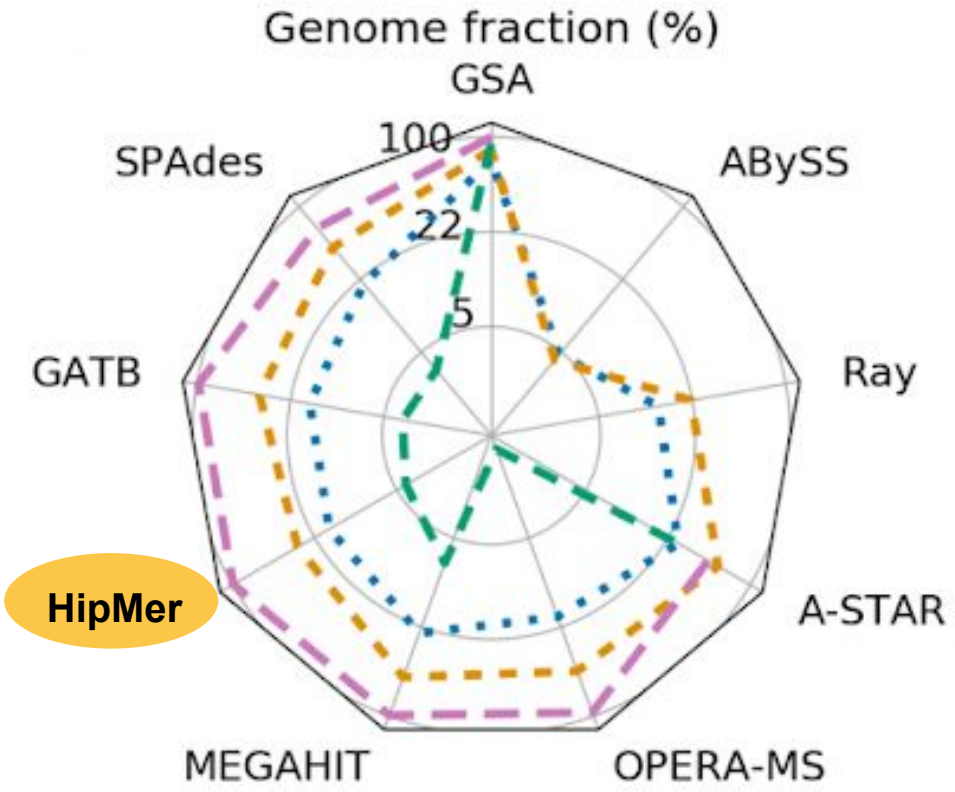
Coassembly enables discovery of more diverse viruses, especially low abundance viruses.
Antonio Camargo

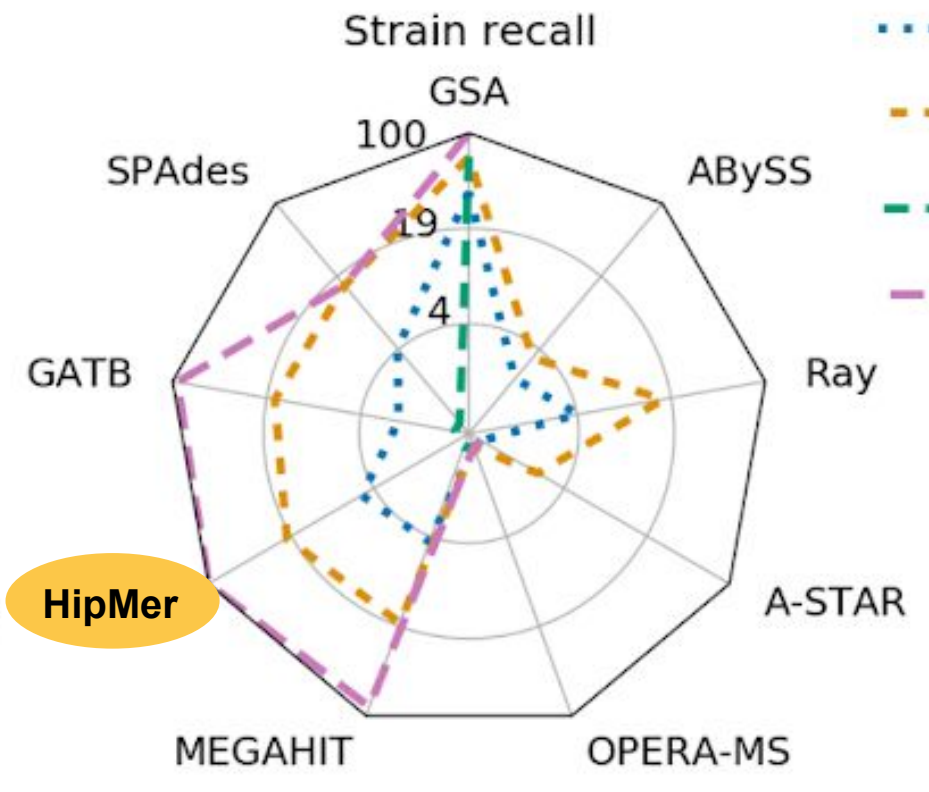
Tara Oceans is most complex; Human gut largest

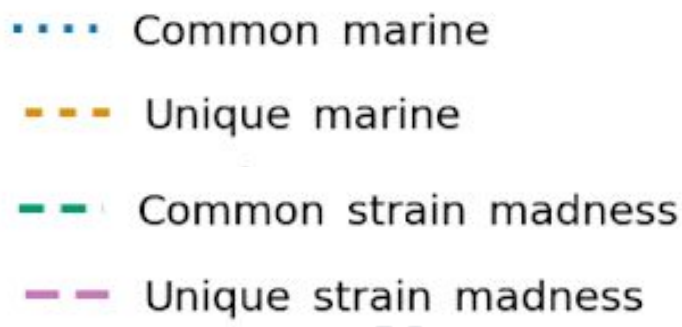


- Complexity drives computing requirement
- Size drive memory need (distributed not shared)
- NEON (soil, plans for 50 Tb) and Lake Mendota (fresh water) are large and complex
- ECP milestone (KPP) based on Tara Oceans with a stretch goal of HMB

MetaHipMer beats others at their own game







"... we analyze 5,002 results by 76 program versions...
The **best ranking method across metrics and all datasets was** [Meta]**HipMer...."**

Okay, I'm sold. How do I start using MHM2?

Build and run on your own cluster
https://bitbucket.org/berkeleylab/mhm2

Use the Docker container (i.e. a 4TB AWS instance)
https://hub.docker.com/r/robegan21/mhm2/

Run your 'Narrative' on KBase
https://www.kbase.us/

Reach out to your JGI contact and have us help or even run it for you at NERSC or Oak Ridge

In the works...

AWS support with Cloud Formation

- Configure, load your data, submit your job (and then pay)

HipMer, the original single large genome assembler

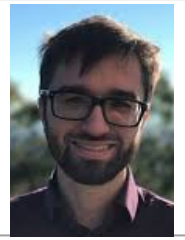
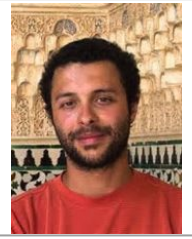
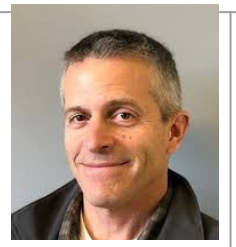
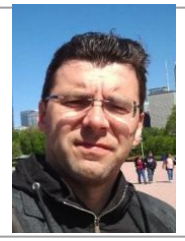
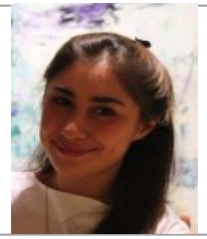
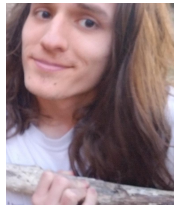
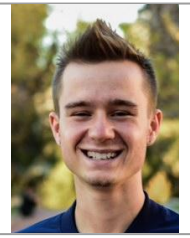
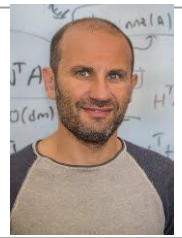
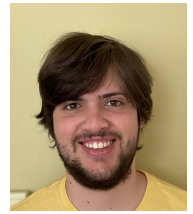
- EOL currently because Berkeley UPC is EOL
- Move HipMer into MHM2 (UPC++) as a single-genome workflow option

AI-powered foundation model for improved scaffolding and binning

- Axiome project

The ExaBiome Team

JGI JOINT GENOME INSTITUTE

								
Muaaz Awan	Ariful Azad	Nick Battacharya	Aydin Buluc	Patrick Chain	Brandon Cook	Alex Copeland	Alicia Clum	Rob Egan
								
Marquita Ellis	E. Georganas	E. Goltsman	Giulia Guidi	Steve Hofmeyr	Taufique Hussain	Richard Lettich	Nikos Kyrpides	J. Madson
								
Hunter McCoy	Russell Neches	Israt Nisa	Lenny Oliker	P. Pandey	Robert Riley	Dan Rokhsar	Gabriel Raulet	Oguz Selvitopi
								
Migun Shakya	Nick Swenson	Andrew Tritt	Kathy Yelick	Brett Youtsey				

BERKELEY LAB

NERSC

Los Alamos NATIONAL LABORATORY EST. 1943

THE UNIVERSITY OF CALIFORNIA · BERKELEY · 1868

upc++

JGI

OAK RIDGE National Laboratory

ECP EXASCALE COMPUTING PROJECT

Lawrence Livermore National Laboratory

INDIANA UNIVERSITY