JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

Metagenome Binning with MetaBAT and GenomeFace

2024 JGI User Meeting
Large Scalable Metagenomics Toolbox Workshop

Rob Egan
04 October 2024

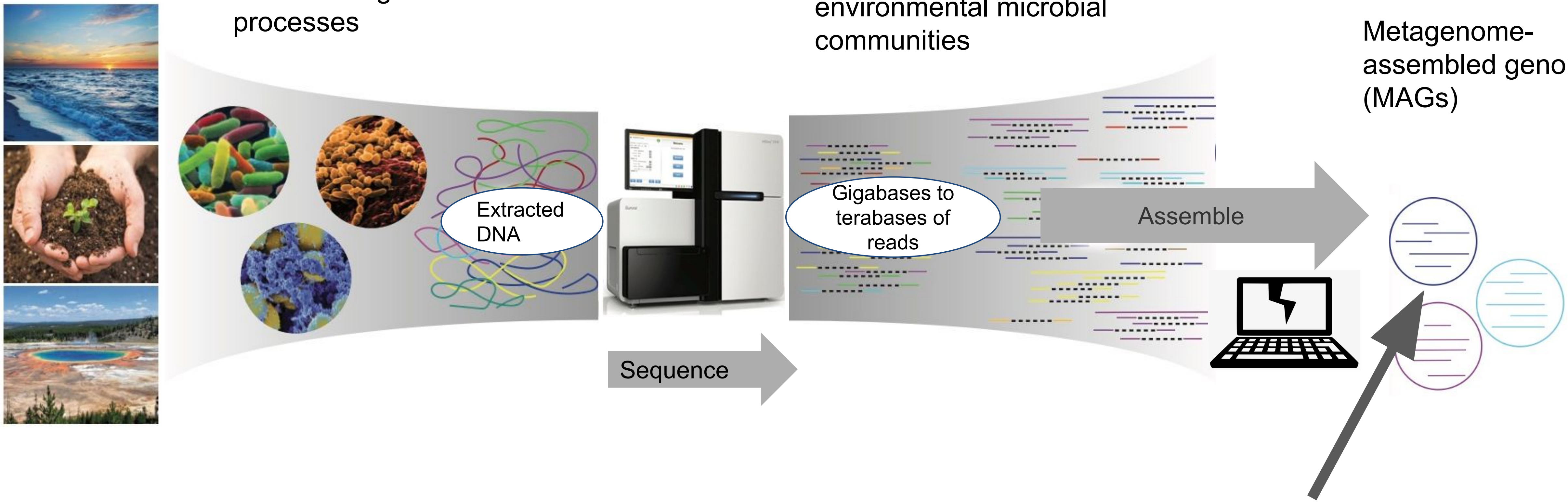
Metagenomics overview

Environments

Microbial communities underlie biogeochemical processes

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

Metagenome-assembled genomes (MAGs)



MetaBAT & GenomeFace

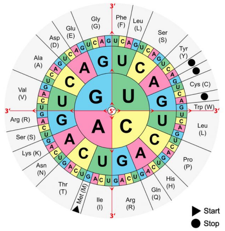
JGI & UC Berkeley / ExaBiome projects

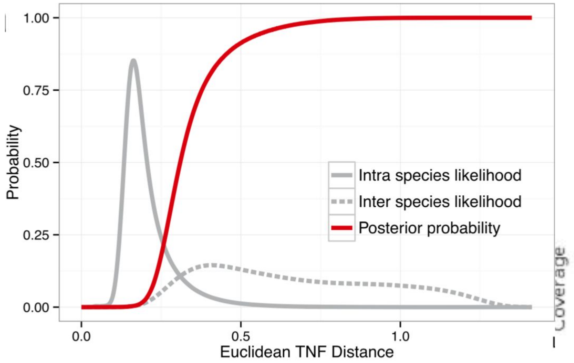
Binning uses extra information hidden in the data

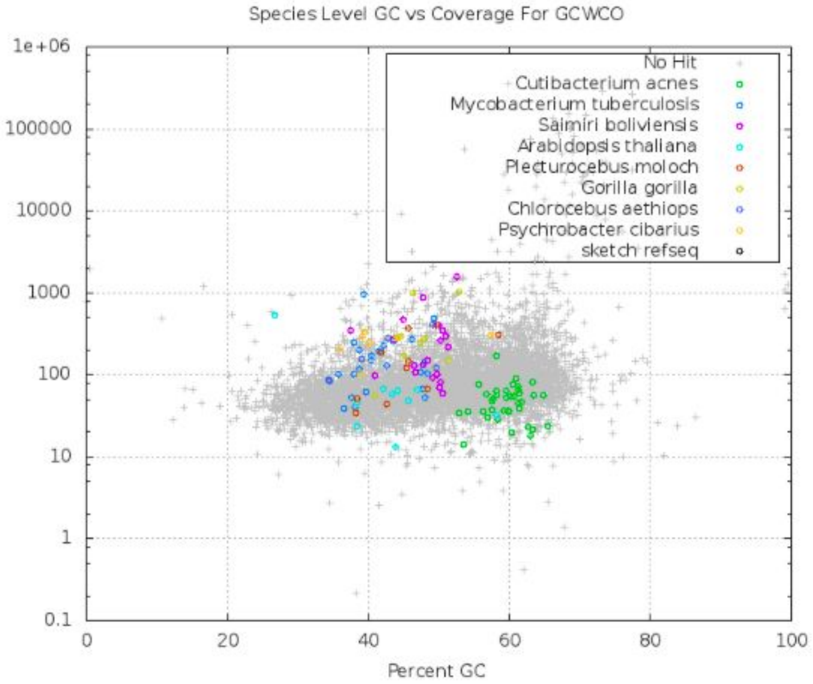
Metagenome assembly is hard to use: large, fragmented and jumbled

Sequence composition

- Codon frequencies
- GC %
- Tetra-nucleotide frequencies
- LLM embeddings?
- Expected core / single copy genes?

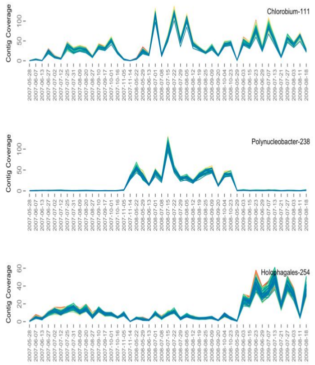






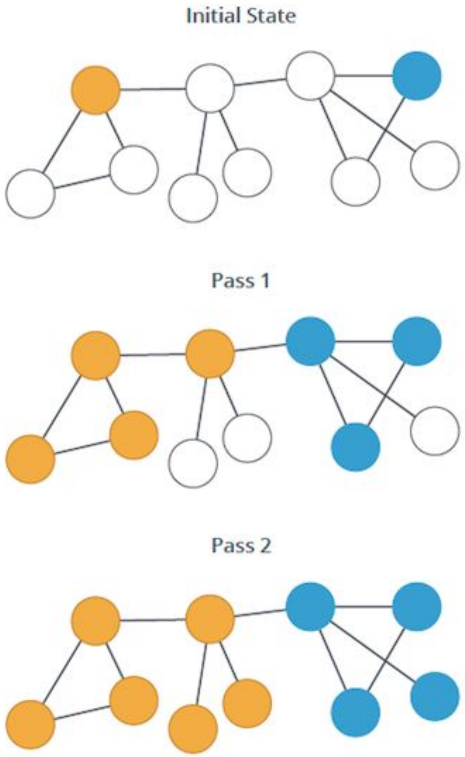
Abundance / coverage

- Expected uniform genomic coverage for each species
- Differential by sample
  - evolution/fitness by time/space/conditions



MetaBAT uses Label Propagation for clustering

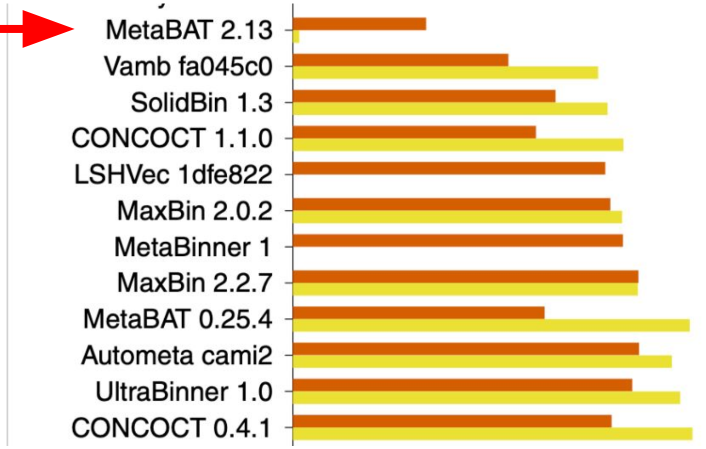
- **Long contigs (default >=2500)**
  - Similarity by sequence composition and differential abundances
  - Generate sparse similarity graph
- **Label Propagation of long contig graph**
  - Initial binning (high threshold)
  - Dissolve small bins
- **All remaining contigs (default >= 1000)**
  - Recruit to existing bins
  - Similarity recalculated to centroid & combined abundances



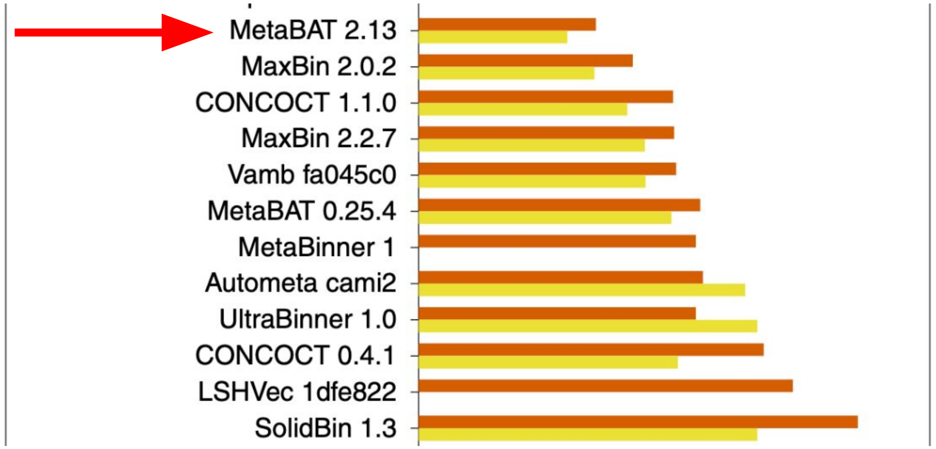
MetaBAT2 - fast, efficient, accurate (from CAMI 2)

Fast & Efficient

Time (log scale)



Memory (log scale)



Very accurate, especially on complex datasets (i.e. assembled strain madness)

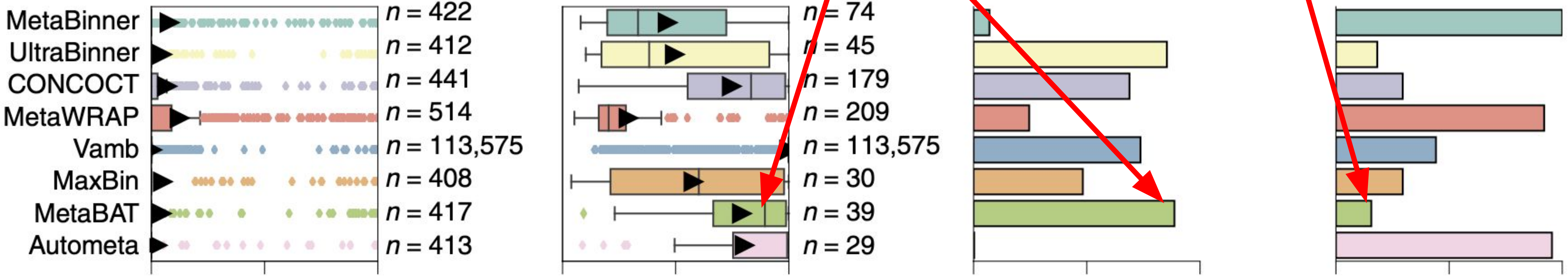
Does miss lower quality bins

Completeness

Purity

ARI (Accuracy)

Binned Bp



Large scale binning (assemblies from MHM2)

MetaBAT MG binning examples in production

8TB Soil -> 75Gbp Assembly

18 hours to bin

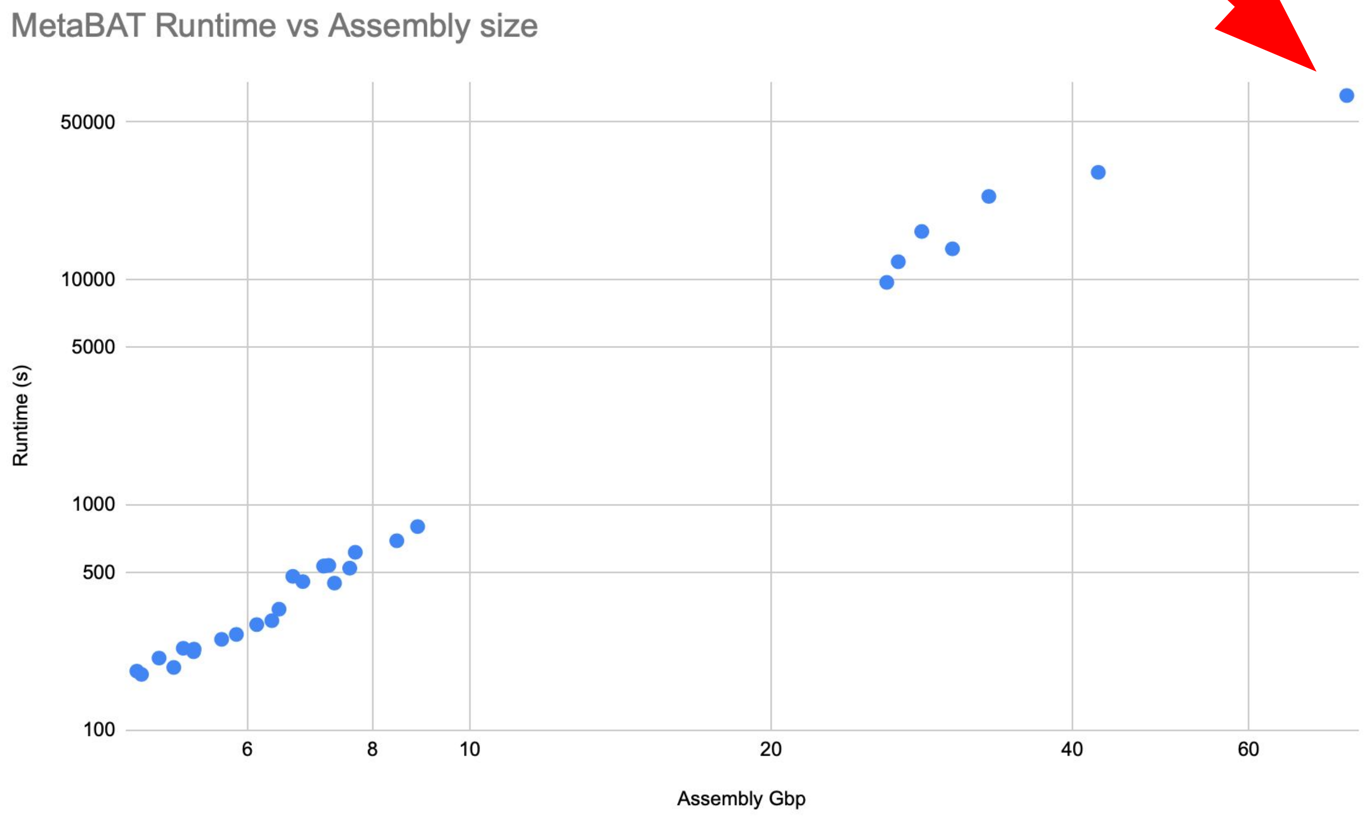
We have also completed much larger projects (not on the graph)

Tara Oceans 72TB -> 323 Gbp

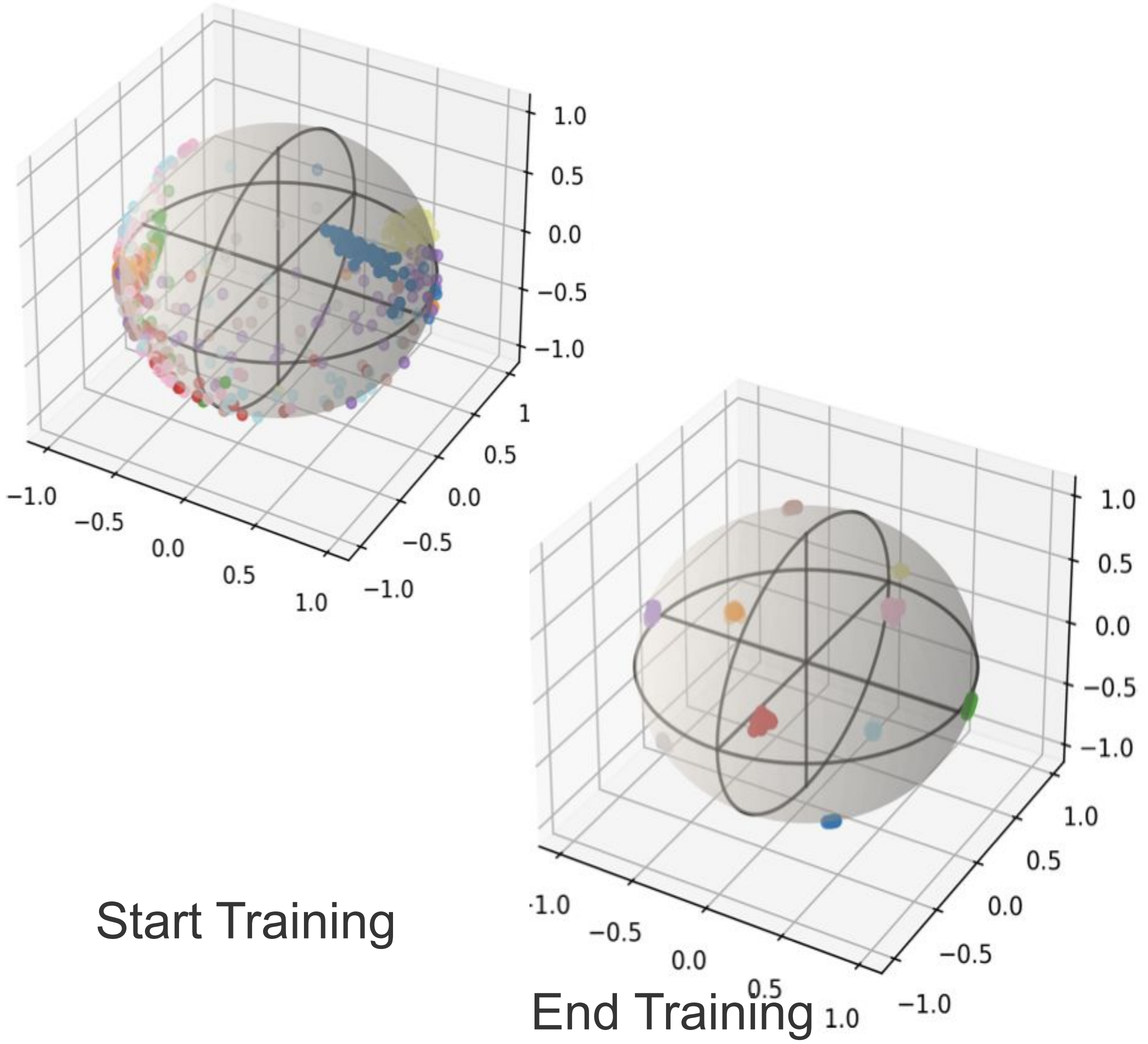
~1 week to bin

HMB 98TB -> 54 Gbp

~1 day to bin



GenomeFace Binner using Machine Learning



Start Training

End Training

Developed new binner using AI/ML

- Inspired by facial recognition
- Trained on 43,000 genomes
- A second neural network captures relative abundance
- Dynamically weight composition vs abundance based on input

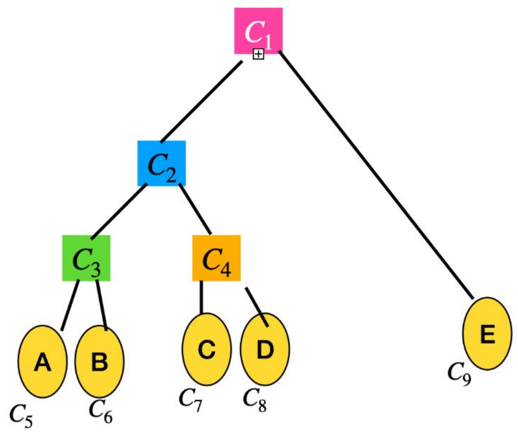
https://doi.org/10.1101/2024.02.07.579326 **GenomeFace: a deep learning-based metagenome binner trained on 43,000 microbial genomes**

Richard Lettich, et al

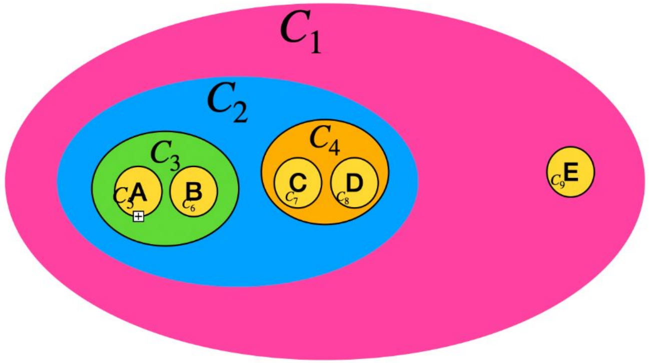
GenomeFace Clustering

- Builds a single hierarchy (minimum spanning tree) based on ML distances
- Uses near-universal marker genes to optimize clusters
- Optimally trades off completeness and contamination
- Uses only two passes over the hierarchy of possible clusters

Dendrogram of $2N - 1$ Nodes

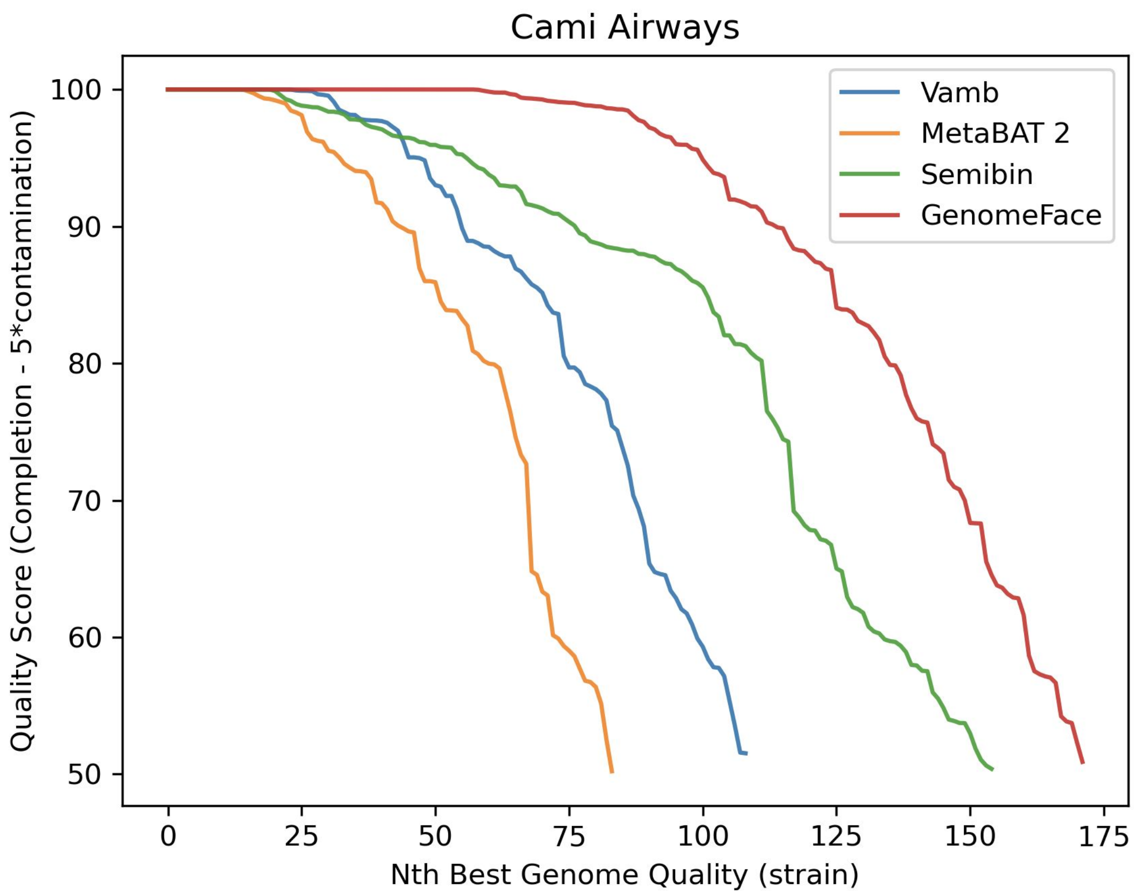


.... Which describes a hierarchy of $2N$ -1 clusters.



Richard Lettich et al

Better quality bins on simulated data



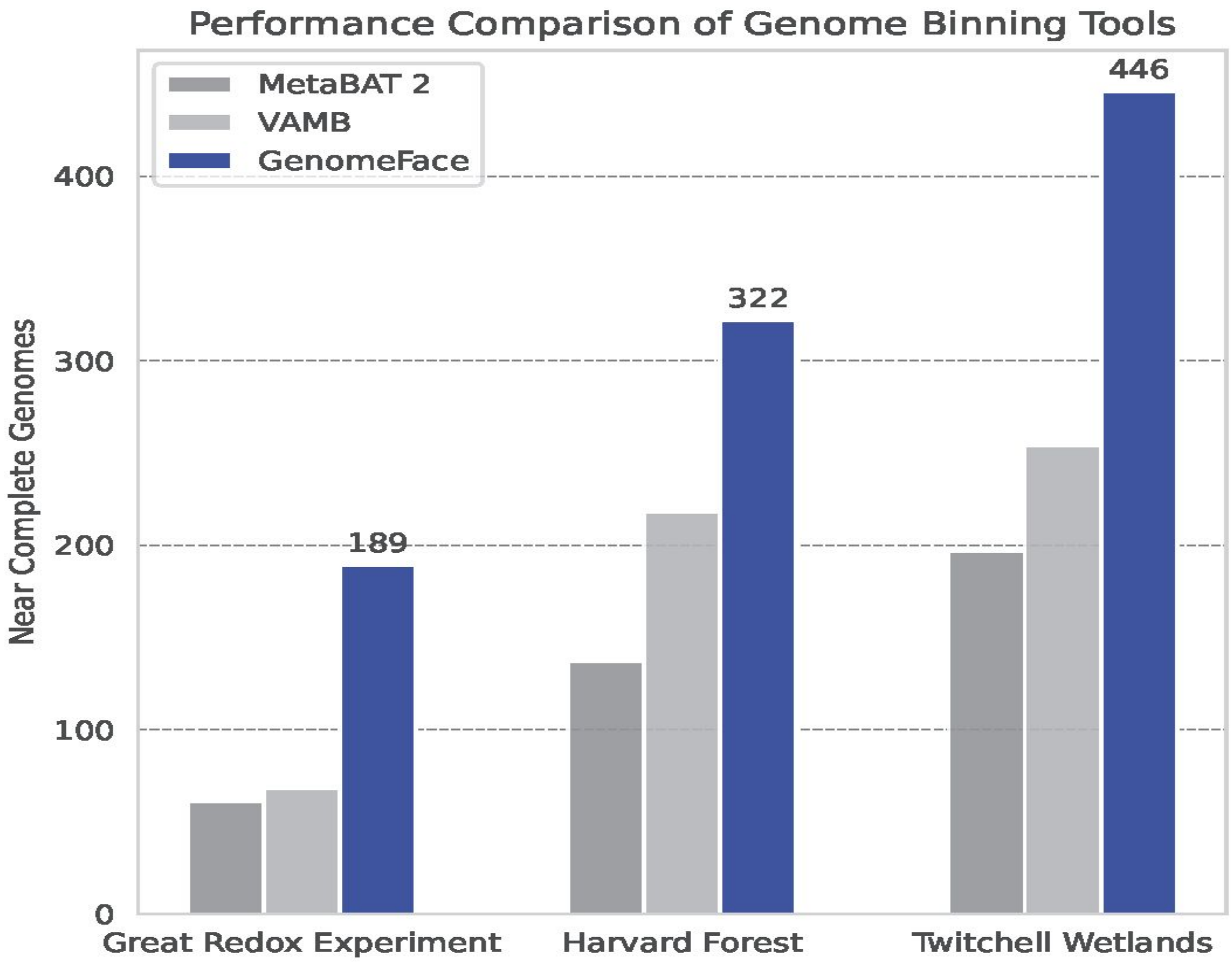
GenomeFace quality

- Outperforms other production binners
- Quality score based on Genome Taxonomy Database score

GenomeFace scalability

- MetaBat 2, Vamb and GenomeFace took under 10 min
- Semibin took 2 hours
- GenomeFace uses GPUs and could extend across nodes

More unique genomes



Bins from 3 of the largest metagenome coassemblies

- 65% more high quality genomes
- 3000 new candidate species (previously uncataloged)

4% expansion of the known bacterial tree of life!

Plans and work in progress…

MetaBAT3

- **Option to use ML for sequence composition metrics**
  - Embeddings (GenomeFace / Genome Ocean / etc)
  - Semi-supervised training on real assemblies
  - Axiome’s Foundation Model
- **New (independent) tool to improve bins with marker genes**
- **Checkpointing**
- **GPU support**
- **Multi-node for the largest problems**

Okay, I’m sold.
How do I start using MetaBAT and GenomeFace?

Build and run on your own computer

https://bitbucket.org/berkeleylab/metabat

Use the Docker container on anyone’s computer

https://hub.docker.com/r/metabat/metabat

Run your ‘Narrative’ on KBase

https://www.kbase.us/

Genome Face

https://richardlett.github.io/gf_instructions.html

Reach out to your JGI contact and have us help!

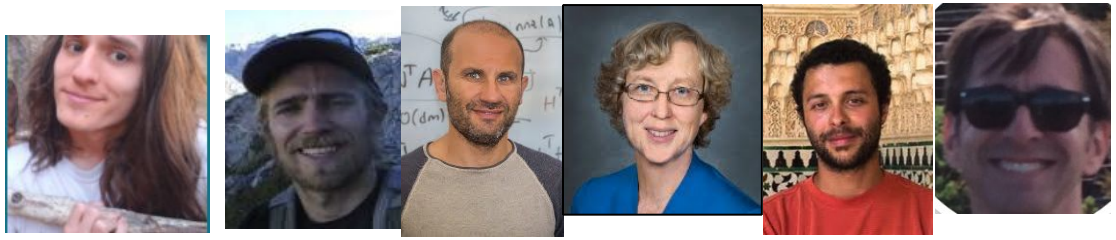
Acknowledgements

MetaBAT contributors

Dongwan Kang
Zhong Wang
Jeff Froula
Feng Li
Ashleigh Thomas
Zhong Wang
Hong An
Volkan Sevim

Genome Face Contributors

Richard Lettich
Robert Riley Andrew Tritt
Lenny Oliker
Kathy Yelick
Aydin Buluç



The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.