Metagenomics, Day 3, Afternoon: Binning and MAGs

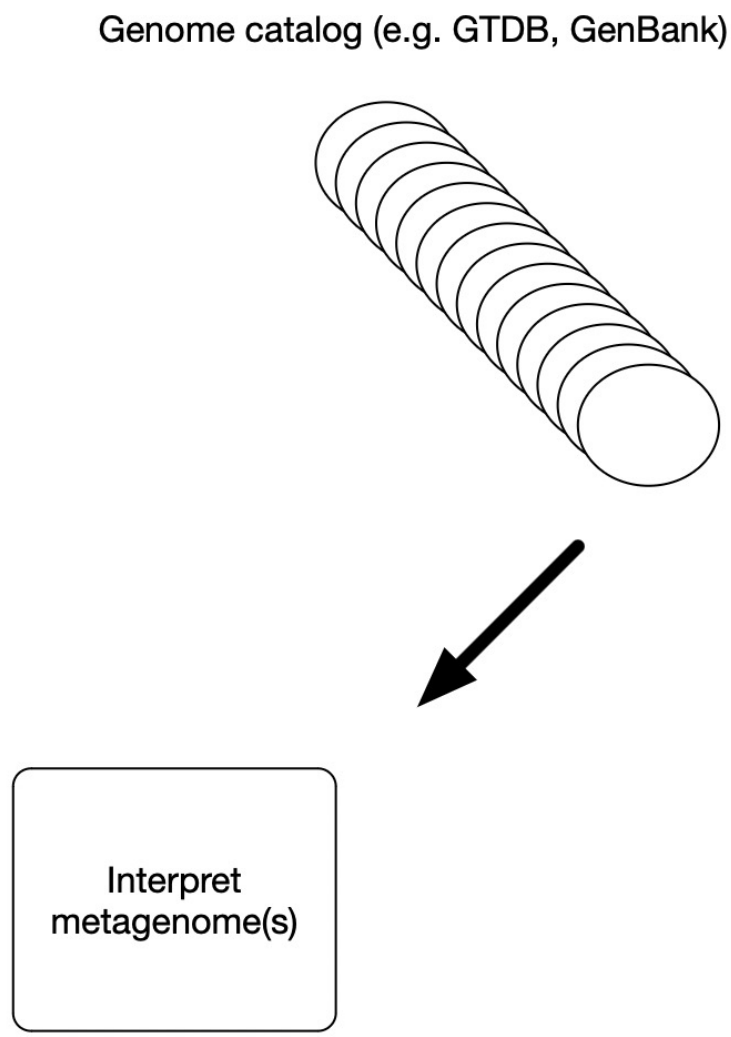
Titus Brown

July 17, 2025

STAMPS 2025

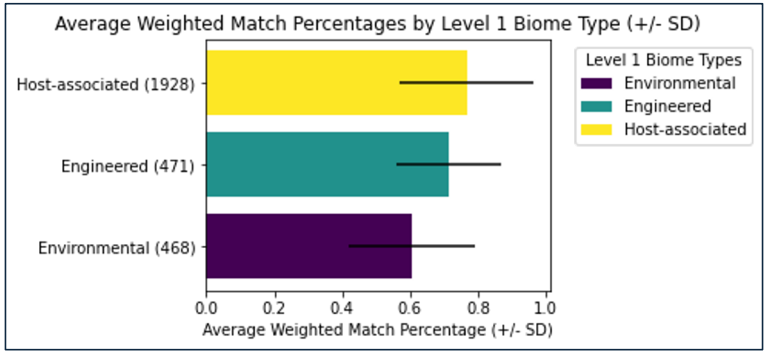
Interpreting metagenomes using a genome catalog is the best way (most sensitive/specific) to interpret metagenome content.

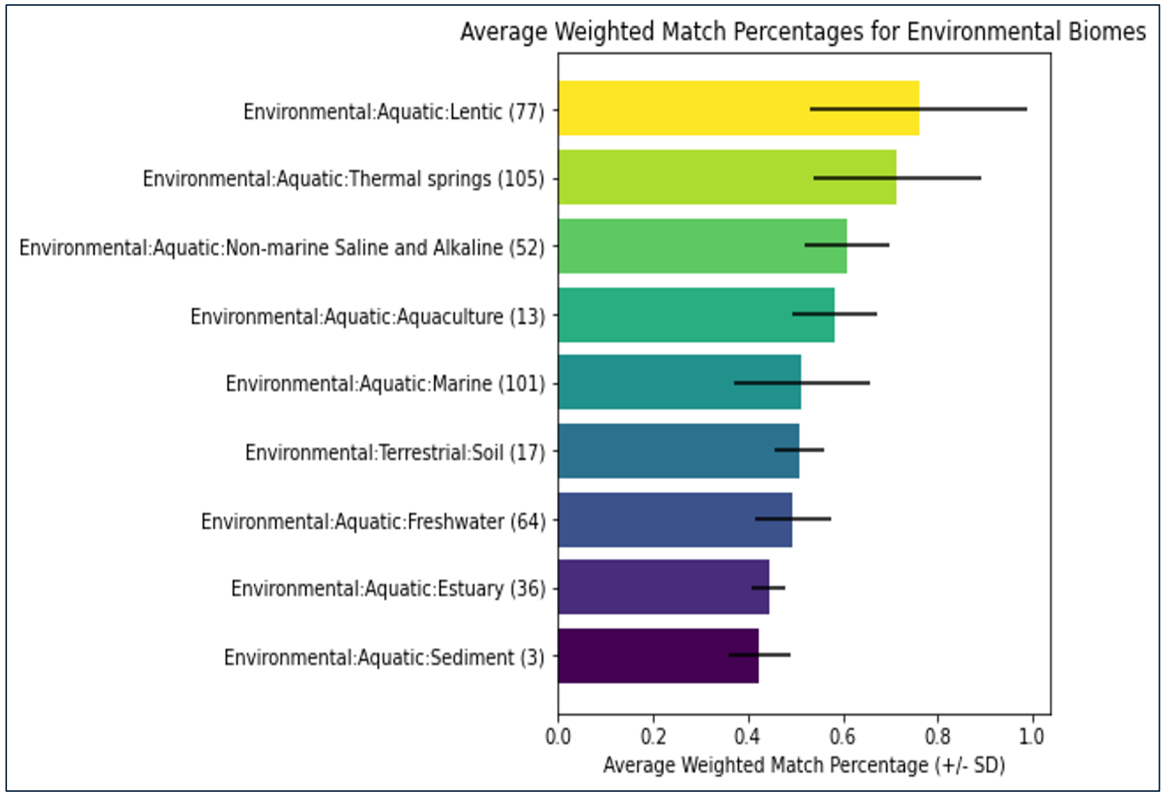
With one big challenge:



Many types of metagenomes have low explainability/assignability

Fraction of metagenomes that match GTDB rs220 + eukaryotes, by metagenome type



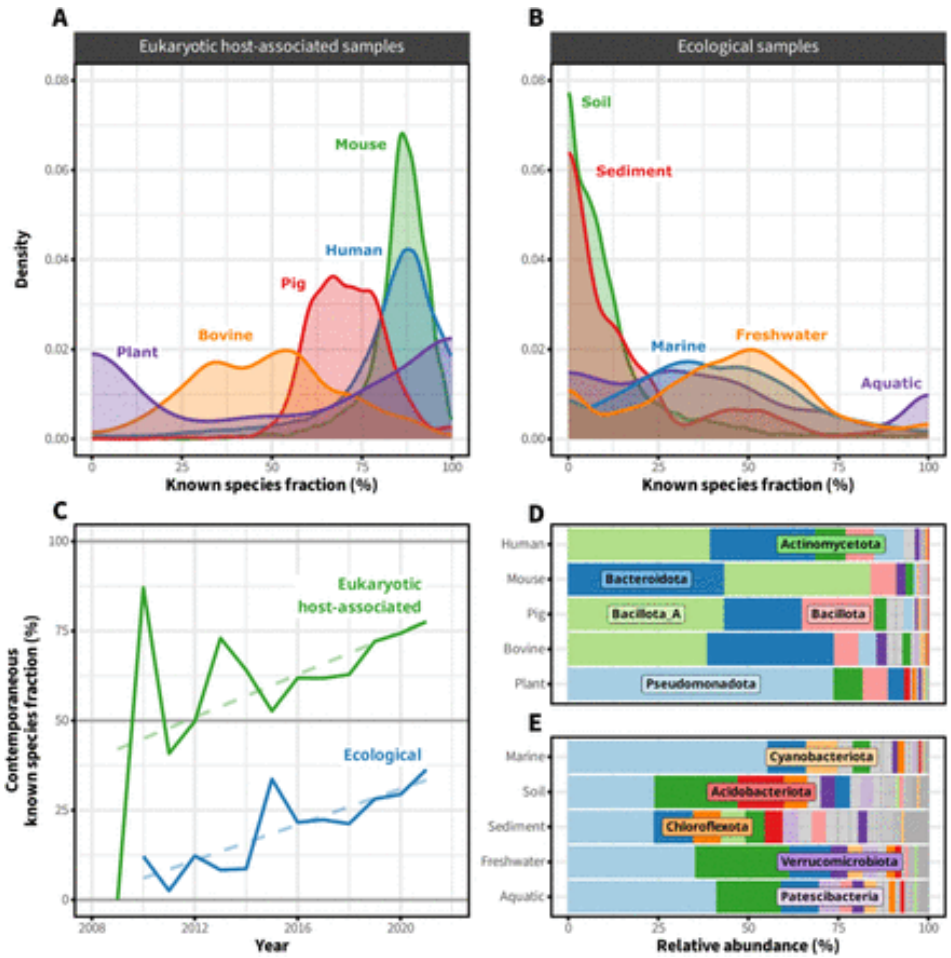


Using ~3000 metagenomes from MGnify

Jean Zhao, UC Davis

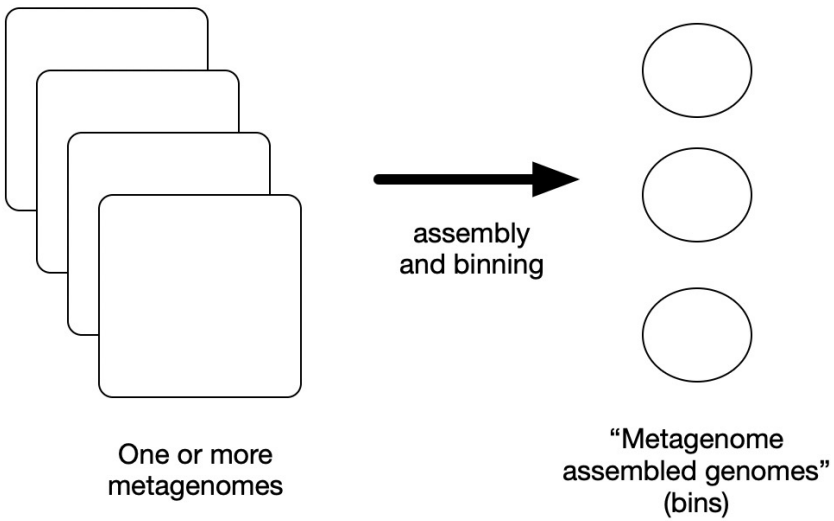
Many metagenomes lack relevant reference genomes.

- Many microbes are hard to culture in isolation;
  - Unknown culture conditions and/or cross-feeding
- Many microbiomes are poorly explored, and/or highly diverse:
  - Soil and sediment are particularly notorious!
- Single-cell microbial sequencing is powerful but does not yet yield complete genomes;



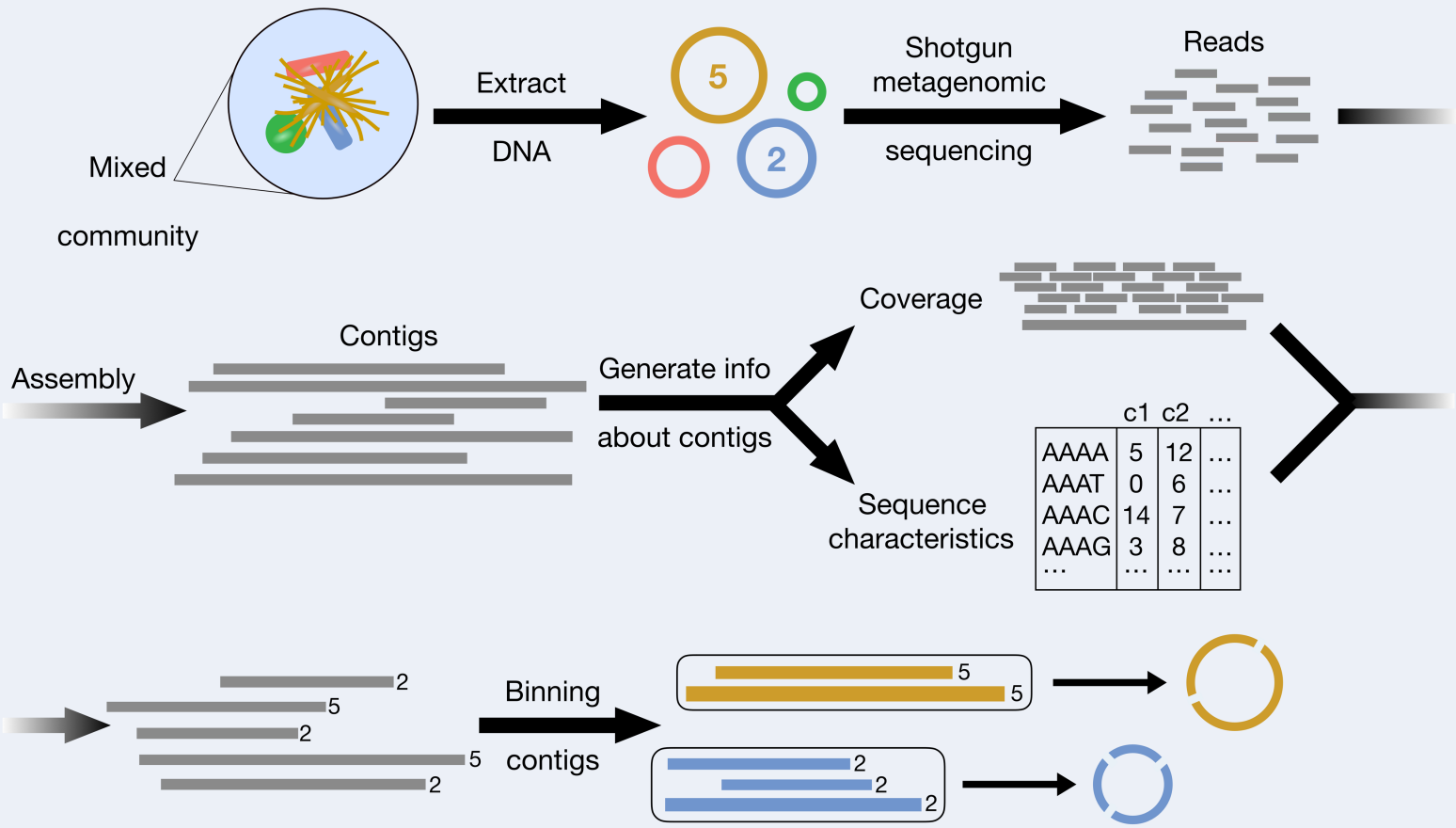
SingleM and Sandpiper: Robust microbial taxonomic profiles from metagenomic data, Woodcroft et al., 2024, bioRxiv

Computationally generate new microbial genomes from metagenomes: “metagenome assembled genomes”, or “genome bins”.



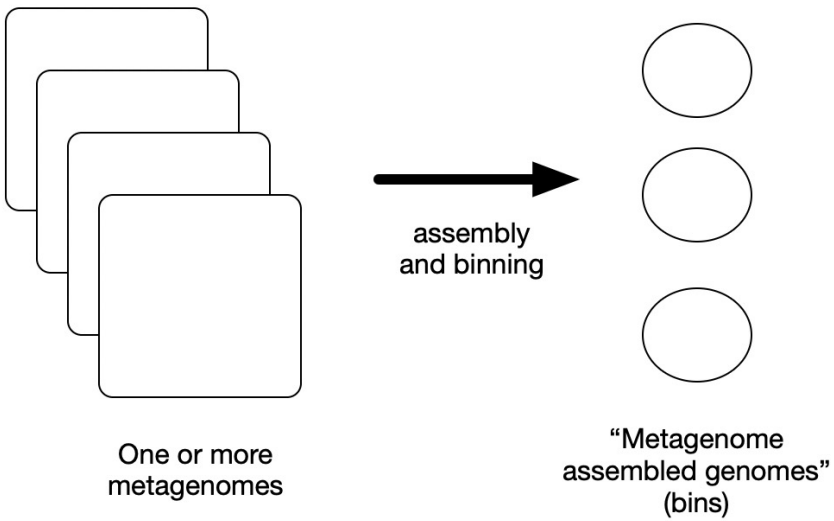
Recovering “genomes” from metagenomes

Illustrated here with 1 sample, but much more powerful with multiple samples

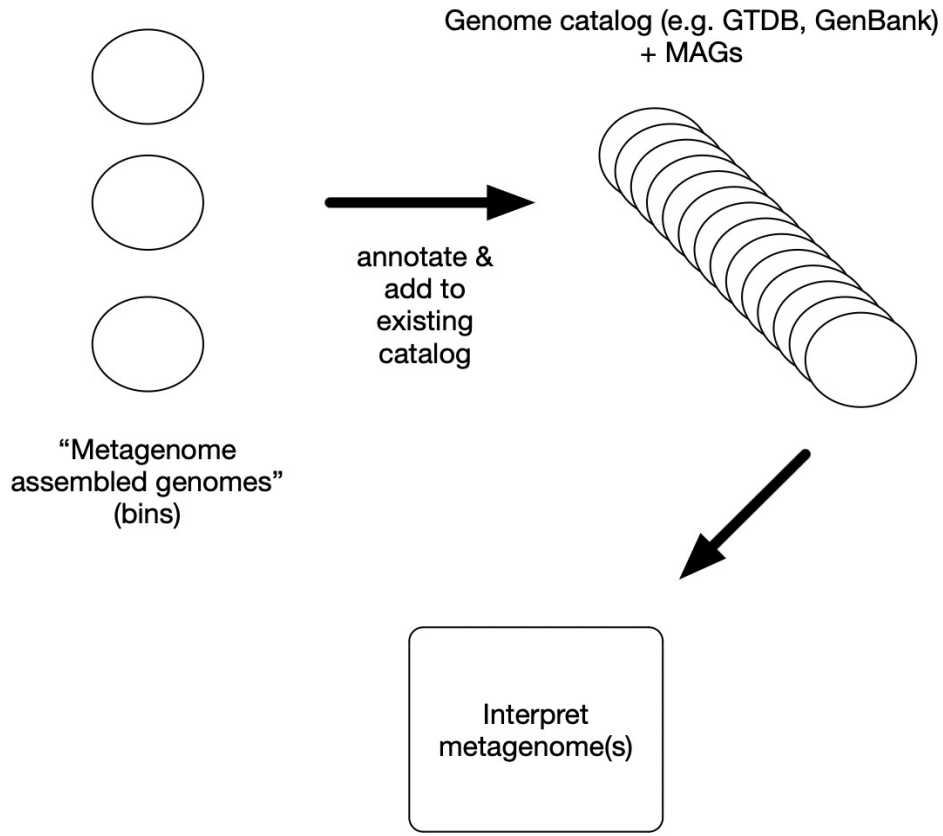


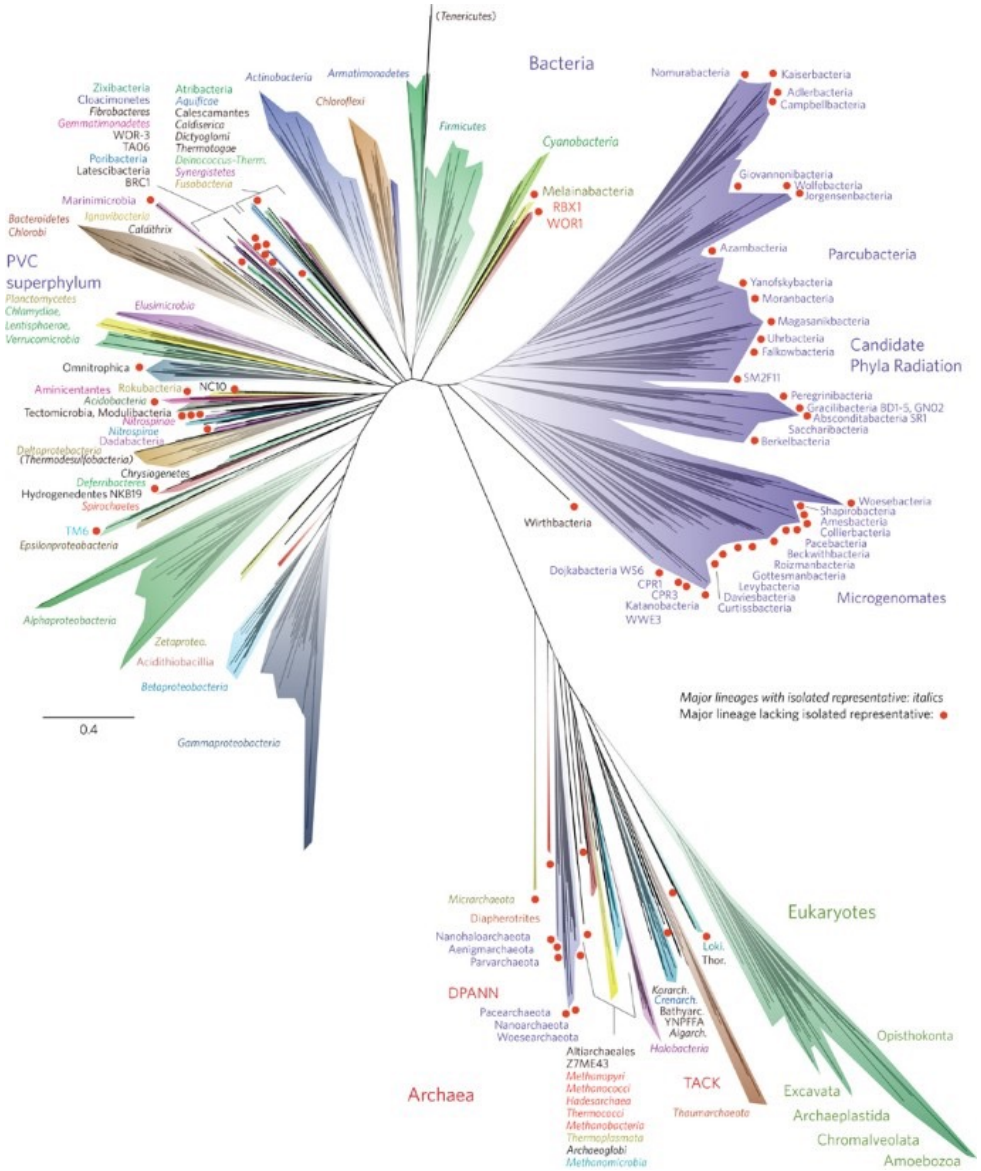
astrobiomike.github.io

Computationally generate new microbial genomes from metagenomes: “metagenome assembled genomes”, or “genome bins”.



Then, add these new genomes to your catalog of *known* genomes and use them to interpret your metagenome(s).





Starting in 2010, the majority of new bacterial and archaeal phyla have been discovered via metagenomics, because they are hard/impossible to culture.

Hug et al., Banfield, 2016.

The majority of genomes in GenBank are now MAGs.

A unified catalog of 204,938 reference genomes from the human gut microbiome

Alexandre Almeida, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, Ekaterina Sakharova, Donovan H. Parks, Philip Hugenholtz, Nicola Segata, Nikos C. Kyrpides & Robert D. Finn

Nature Biotechnology **39**, 105–114 (2021) | Cite this article

Etc. ☺

Article | Open access | Published: 11 November 2023

A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources

Bin Ma, Caiyu Lu, Yiling Wang, Jingwen Yu, Kankan Zhao, Ran Xue, Hao Ren, Xiaofei Lv, Ronghui Pan, Jiabao Zhang, Yongguan Zhu & Jianming Xu

Nature Communications **14**, Article number: 7318 (2023) | Cite this article

Questions to discuss:

- When is a binning effort useful?
- What kinds of sequencing is most appropriate? => long reads
- What environments are “hard” to bin? => highly rich & diverse environments, as well as environments with lots of strain diversity.

Todd Treangen will discuss assembly, and binning, and strain diversity, on Friday!

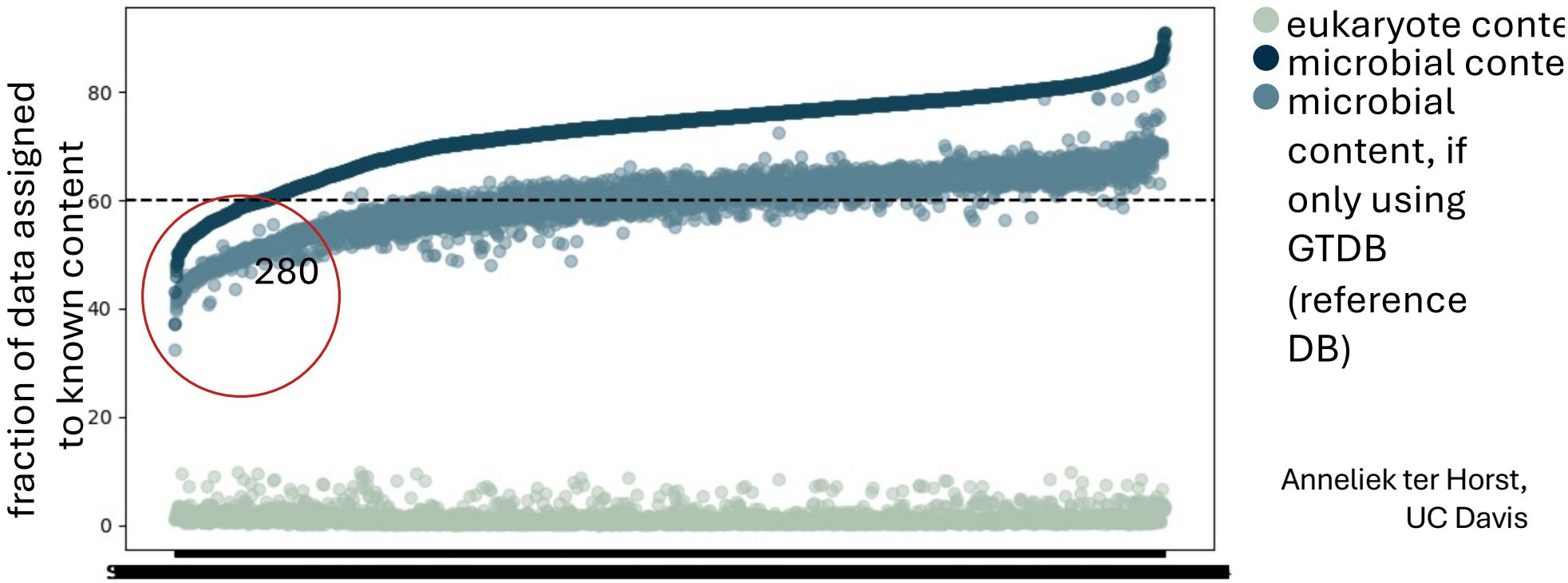
An (emerging) case study: building a genome catalog from pig gut metagenomes

Pig gut metagenomes: worldmap + country pcoa.

Anneliek ter Horst,
UC Davis

Figure: Overview of country of origin for all metagenomes used. Bubbles are log-transformed and relative to the number of metagenomes from that country

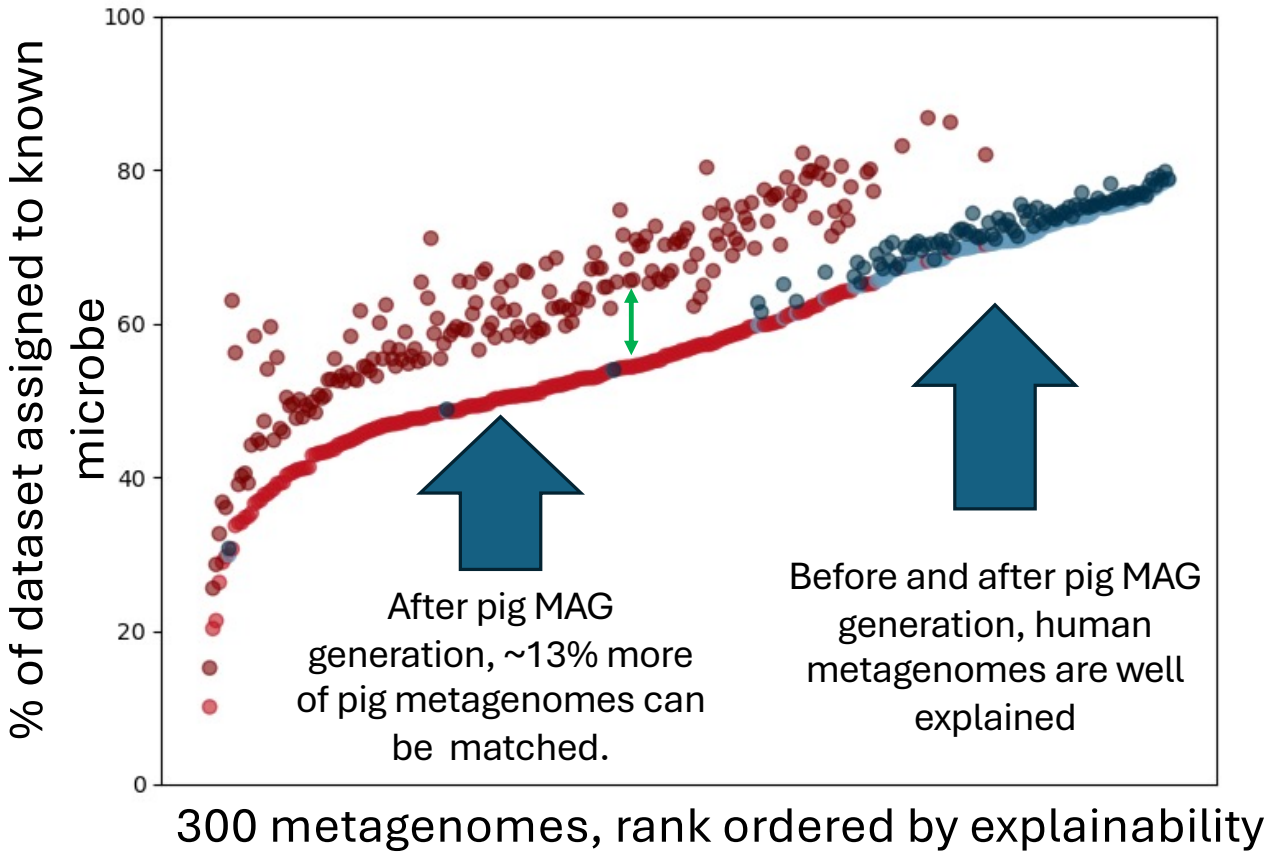
Unknown microbial content decreases when adding host-specific reference genomes by 13% on average



Anneliek ter Horst, UC Davis

3,217 datasets, rank ordered by fraction of data assigned to microbial genomes

MAG generation increases the amount of data in metagenome that can be assigned to a species



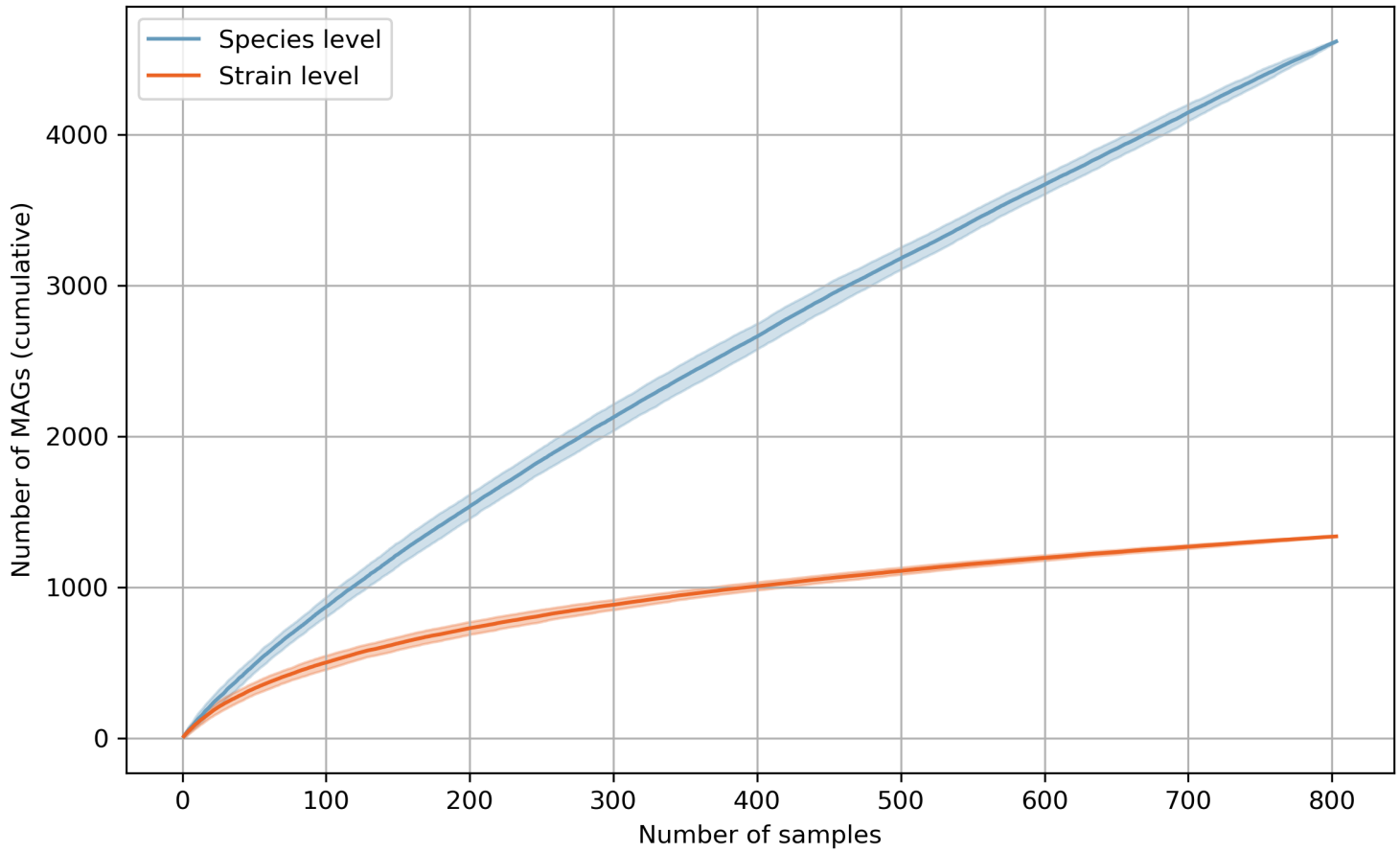
Adding pig specific reference genomes helps assign more data to (now) known microbial genomes.

Anneliek ter Horst,
UC Davis

Anneliek ter Horst,
UC Davis

While species level diversity saturates, strain does not

- **How many new species/strains are found in each new sample?**
- Species level curve suggests sufficient sampling effort
- Strain level curve shows no saturation, so more samples will reveal more strain level diversity



Thoughts on the pig gut metagenome effort -

- We are seeing substantial new strain-level diversity and some surprising species-level diversity that would have gone unrecognized without this effort.
- Seems likely that each new environment will have many new strains and some new species.

- We are developing metrics for sourmash to help determine when a MAG-building effort is a potentially good idea (as well as to figure out when you’re “done” ☺)

Perhaps the most important thing to know about MAG generation and genome binning:

- Assembly and binning has reasonably high precision, but *very* poor recall.
- OR, to put it another way, *most* genomes present in a metagenome ***will not assemble+bin*.**
- This can be for many technical reasons; ask Todd for details on Friday ☺

- HOWEVER, bins generated from a particular metagenome can help you interpret **other** metagenomes from that environment.

Perhaps the most important thing to know about MAG generation and genome binning:

- When exploring a new environment, apply assembly and binning to ***all*** your metagenomes.
- Build an "environment specific" catalog...
- ...and then ***combine*** that with your other catalog(s), and use the combined catalog to interpret your metagenomes.

- Do *not* assemble and bin a metagenome, and then characterize the bins, and then claim that you've characterized the entire metagenome. ☺