Metagenomics

BIOC 6102 Special Topics

September 13, 2022

Michael S. Robeson II, Ph.D.
University of Arkansas for Medical Sciences
Department of Biomedical Informatics

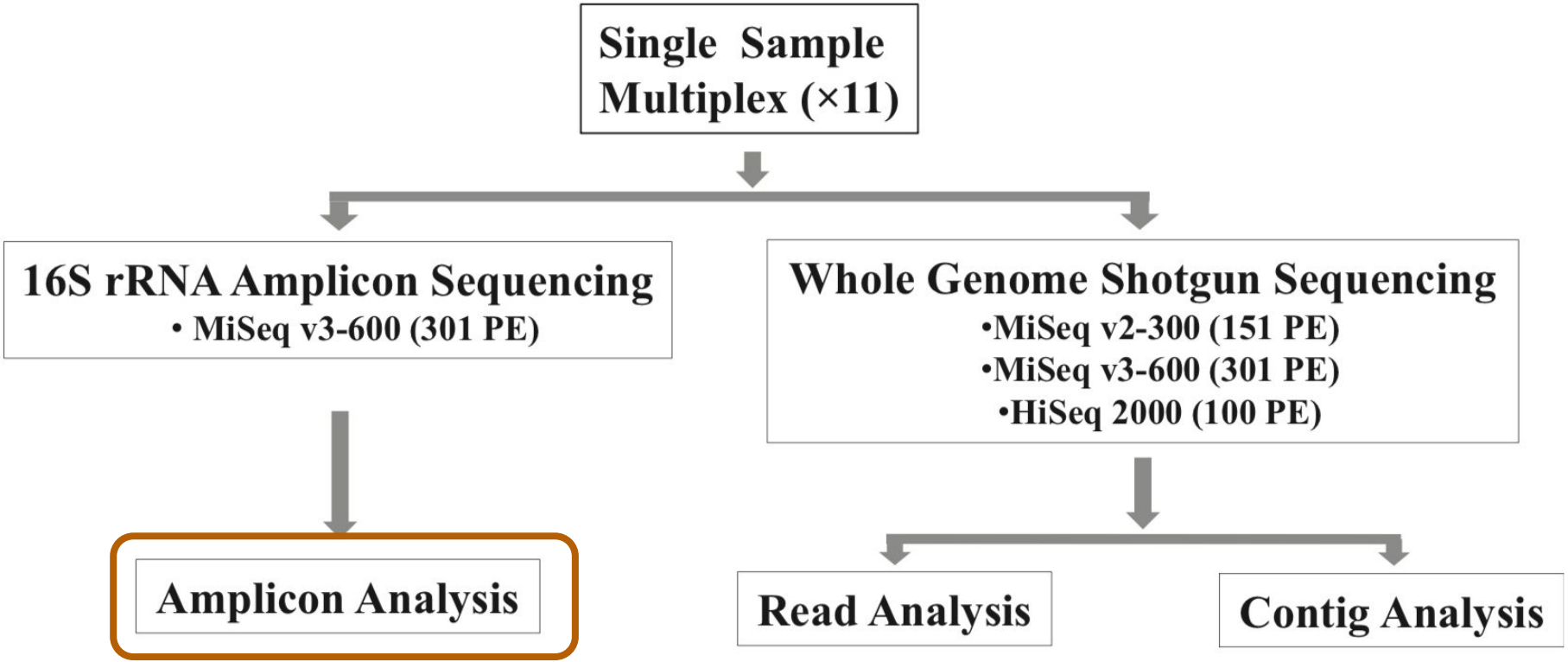
Metagenome

“The collection of genomes and genes from the members of a microbiome.”

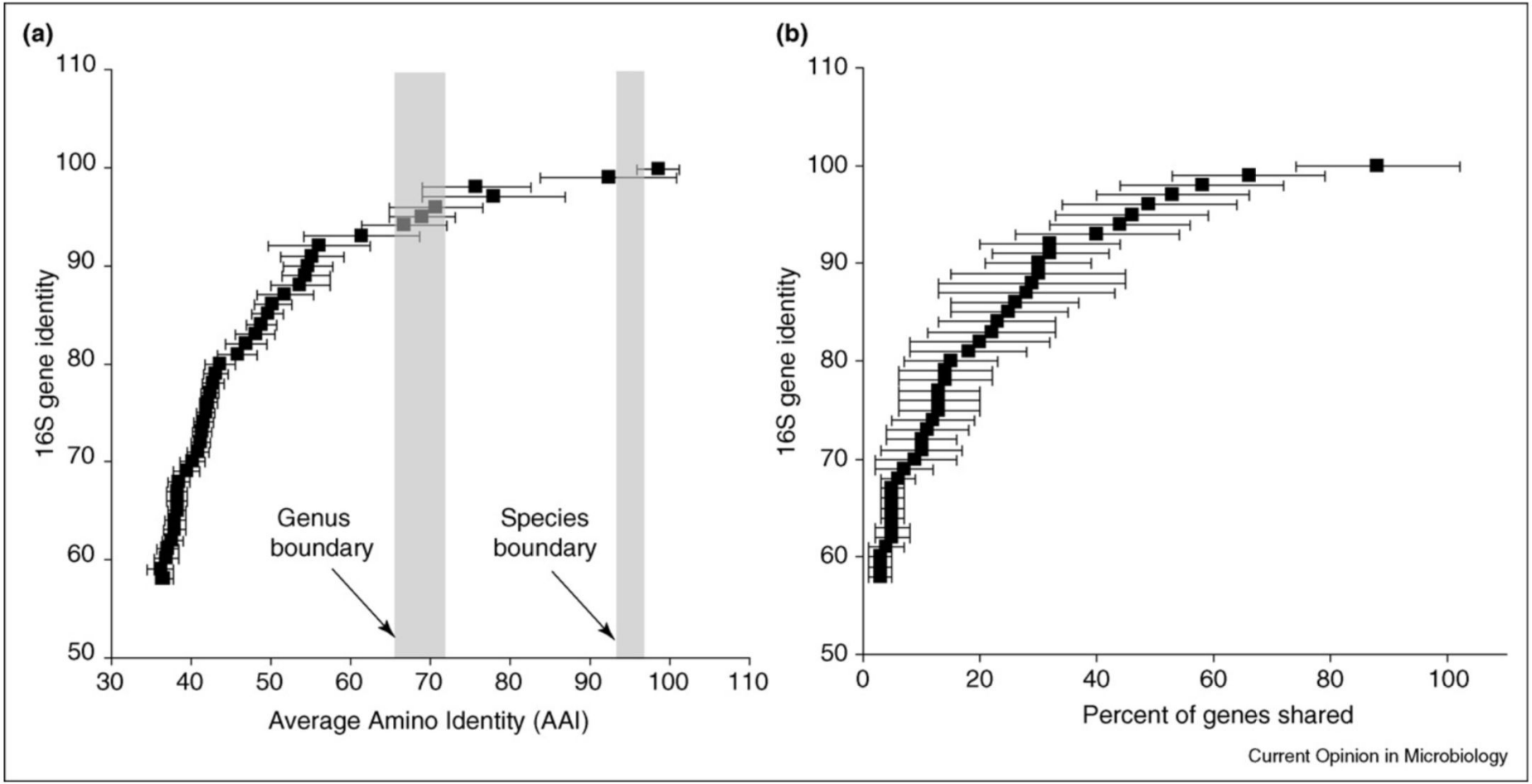
Obtained through shotgun sequencing of DNA extracted from a sample

Marchesi, J.R., and Ravel, J. 2015. The vocabulary of microbiome research: a proposal. Microbiome 3: 31. doi:10.1186/s40168-015-0094-5.

Experimental Strategy



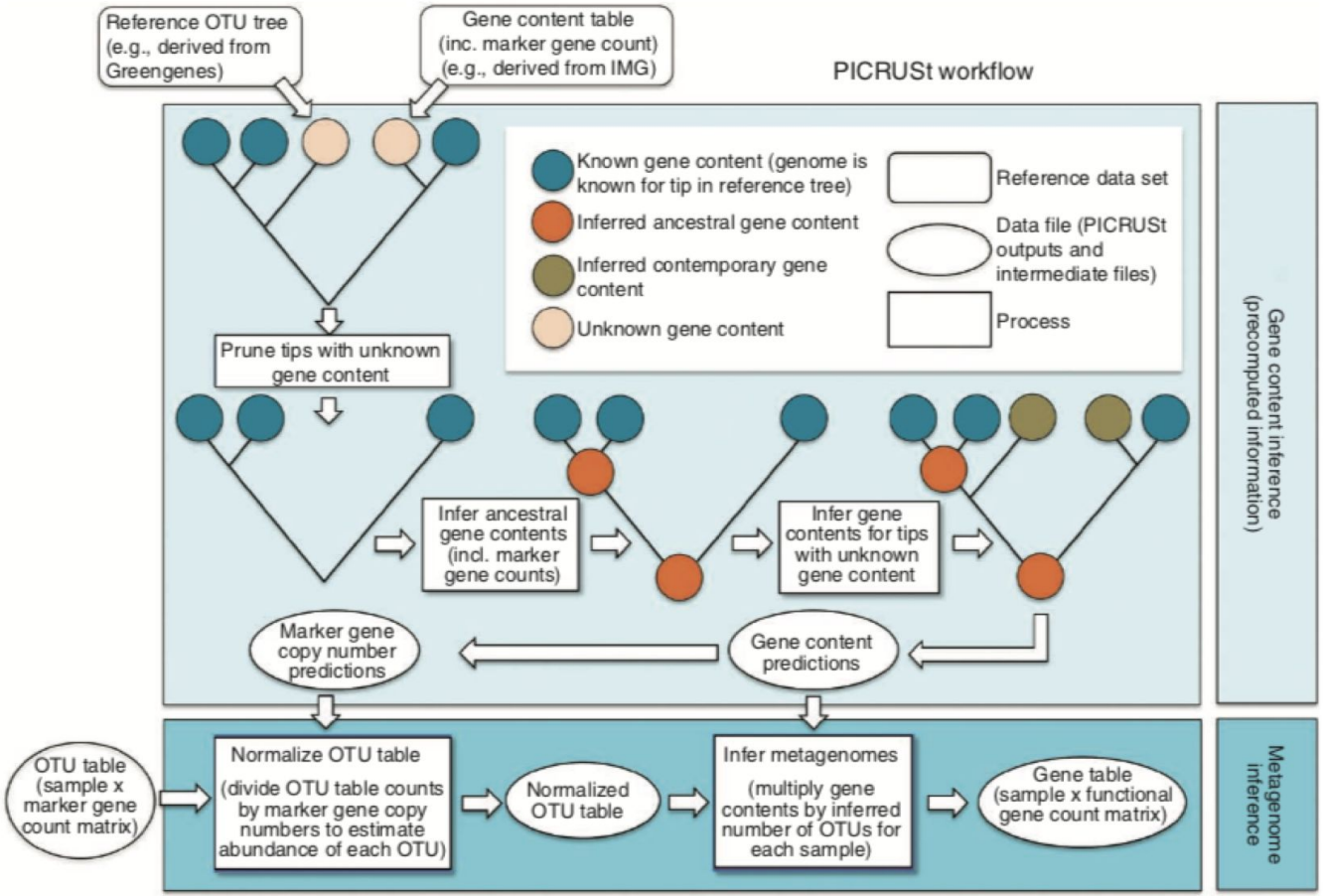
Ranjan et al. (2016) Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem. Biophys. Res. Commun. 469(4): 967–977



Konstantinidis and Tiedje (2007) Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. Curr. Opin. Microbiol. 10(5): 504–509

Predict Metagenome Function: PICRUSt

Phylogenetic Investigation of Communities by Reconstruction of Unobserved States



Ranjan et al. (2016) Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem. Biophys. Res. Commun. 469(4): 967–977

Douglas et al. 2020. “PICRUSt2 for Prediction of Metagenome Functions.” *Nature Biotechnology*, June. https://doi.org/10.1038/s41587-020-0548-6.

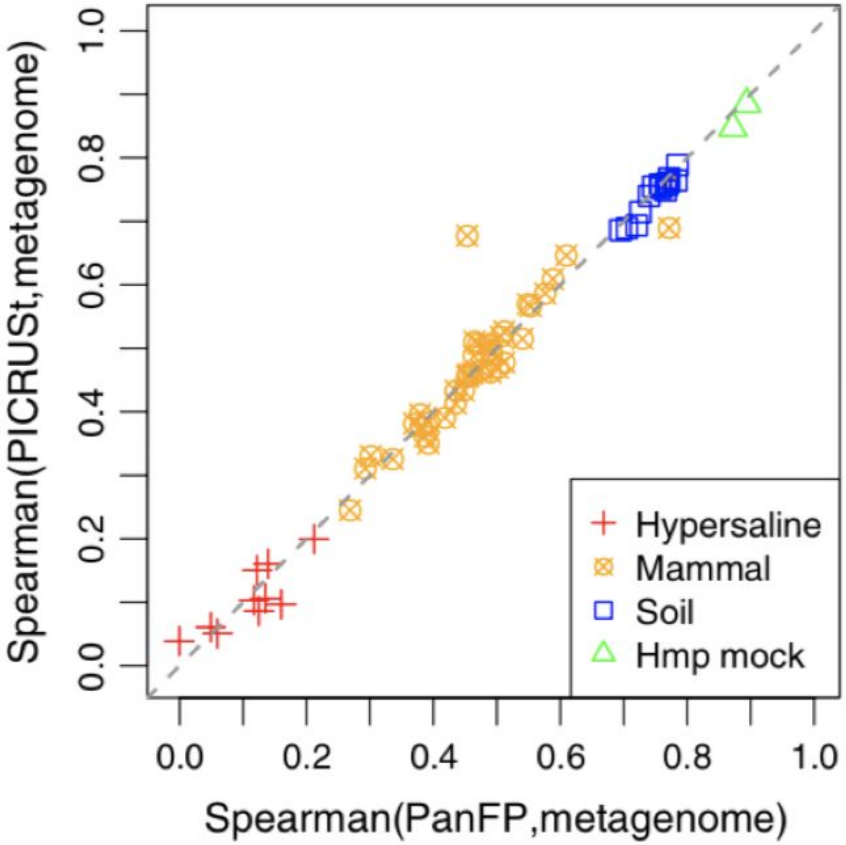


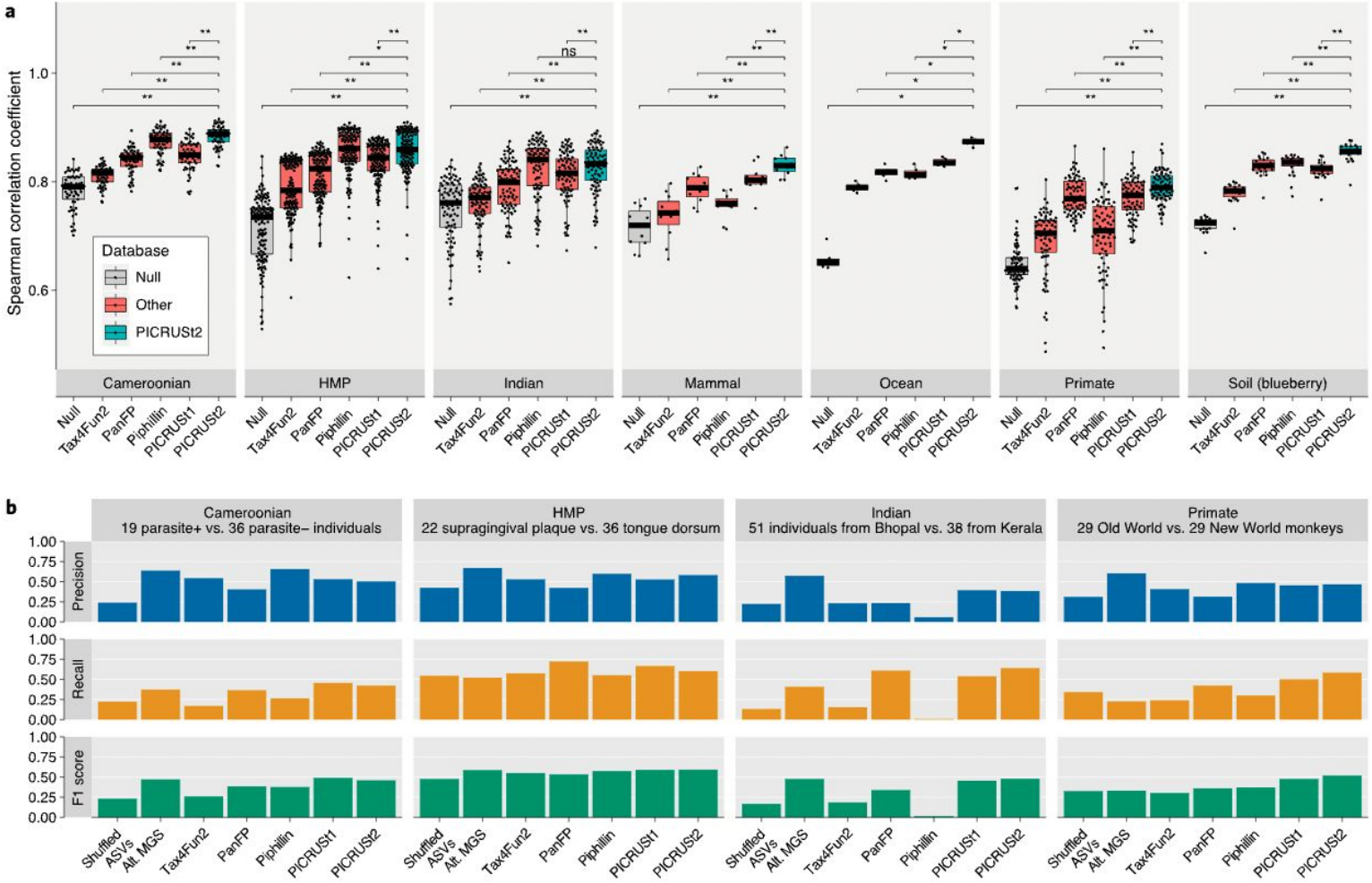
Fig. 2 A *scatter plot* of the two Spearman correlations for 65 samples: X-axis and Y-axis represent correlations between functional profiles inferred by PanFP and PICRUSt versus sequenced metagenomic functional profiles, respectively

Pan-Genome Functional Profiles

Based on Taxonomy

Jun, Se-Ran, Michael S. Robeson, Loren J. Hauser, Christopher W. Schadt, and Andrey A. Gorin. 2015. "PanFP: Pangenome-Based Functional Profiles for Microbial Communities." *BMC Research Notes* 8 (1): 479.

PICRUSt2 for Prediction of Metagenome Functions



Douglas *et al*. 2020. "PICRUSt2 for Prediction of Metagenome Functions." *Nature Biotechnology*, 38:6 685-688 https://doi.org/10.1038/s41587-020-0548-6.

Potential Pitfalls of metagenome prediction based on single genes

- Functional redundancy among taxa.
- Should be confirmed experimentally and/or with other metagenomic & metabolomic approaches.

Burke et al. 2011. Bacterial community assembly based on functional genes rather than species. Proc Natl Acad Sci U S A. doi:10.1073/pnas.1101591108.
Tian et al. 2017. Deciphering Functional Redundancy in the Human Microbiome. bioRxiv: 176313. Cold Spring Harbor Laboratory. doi:10.1101/176313.

Consistent and correctable bias in metagenomic sequencing experiments

Michael R McLaren[1], Amy D Willis[2], Benjamin J Callahan[1,3]*

[1]Department of Population Health and Pathobiology, North Carolina State University, Raleigh, United States; [2]Department of Biostatistics, University of Washington, Seattle, United States; [3]Bioinformatics Research Center, North Carolina State University, Raleigh, United States

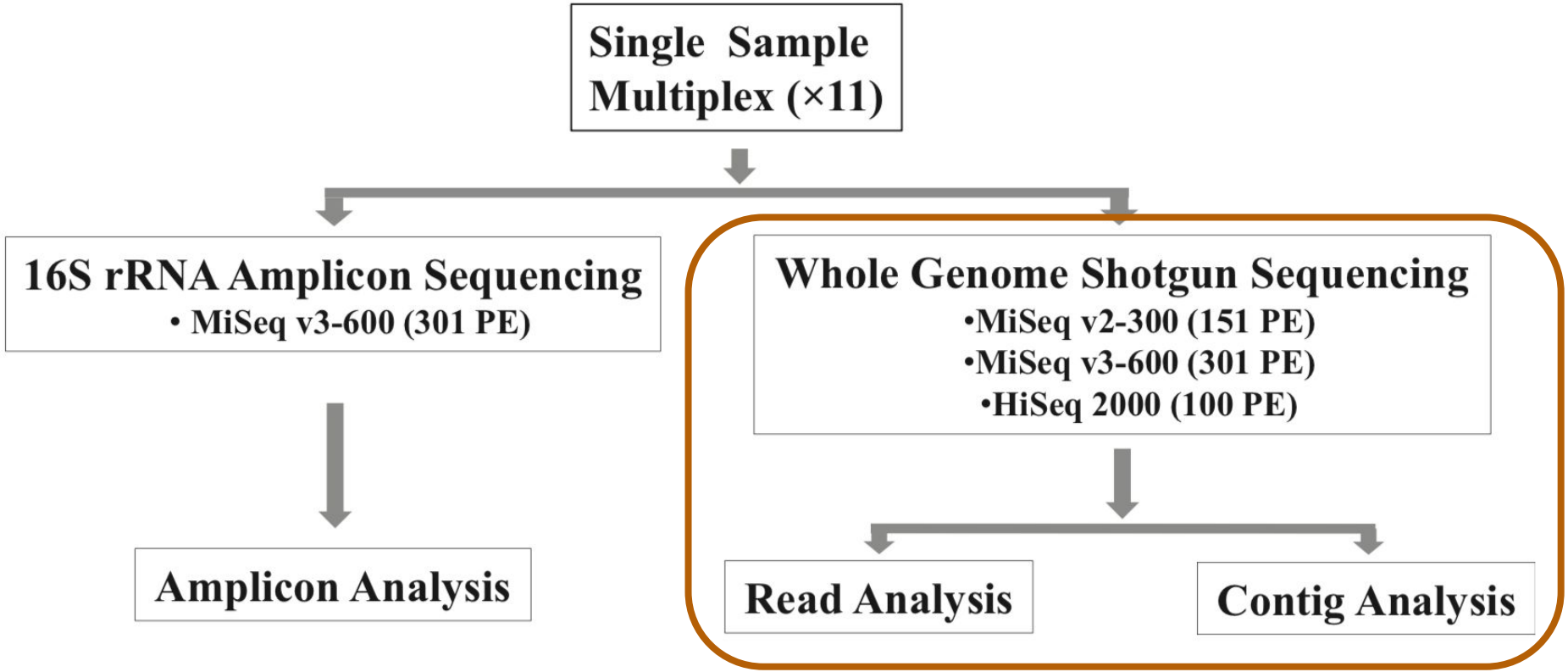
[Marker-gene and metagenomic sequencing (jointly MGS)] methods are now being adopted in fields ranging from food safety to wastewater remediation to forensics along with biology and medicine. Unfortunately, however, the community compositions measured by MGS are wrong.

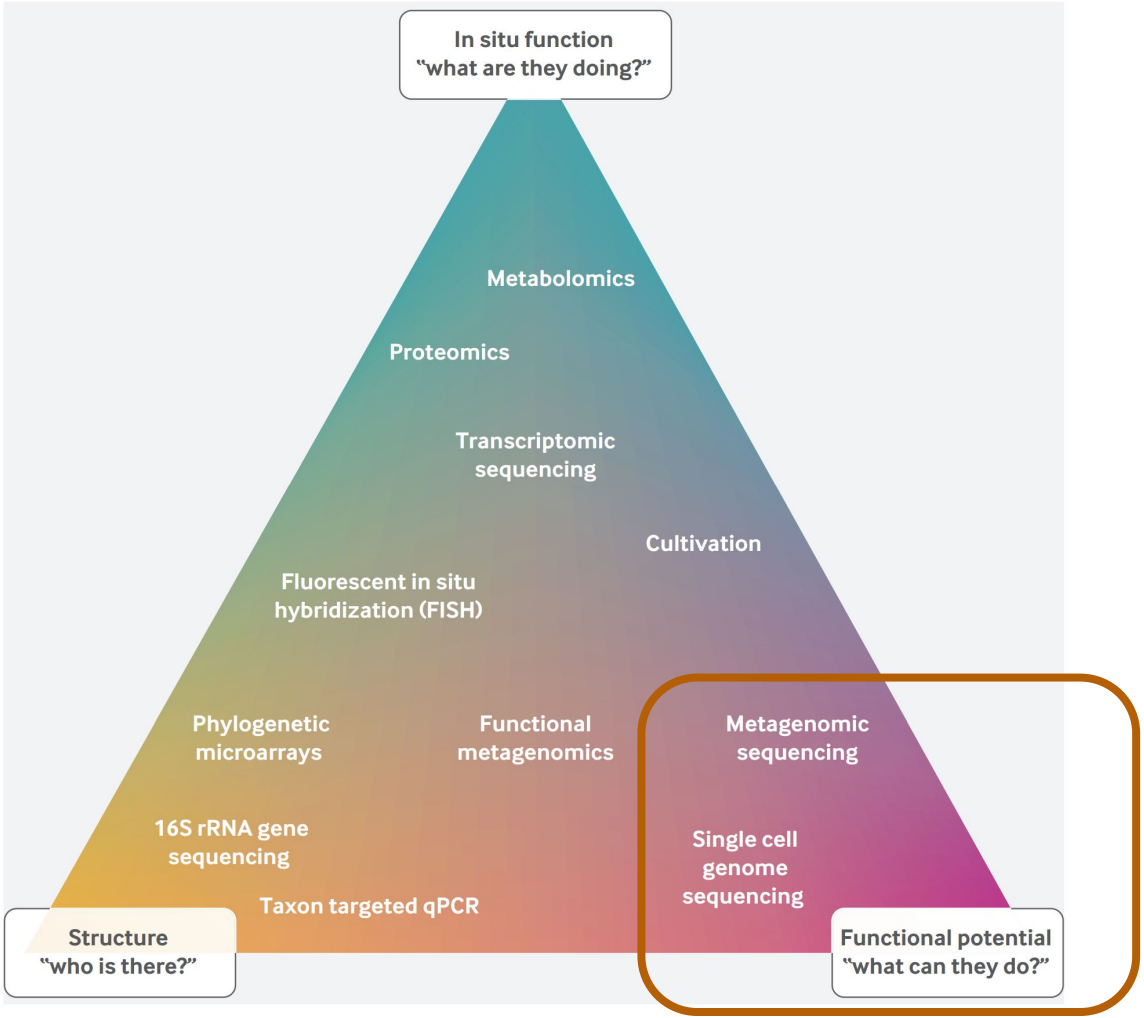
Biases of marker gene and metagenomics

That is, metagenomics is NOT necessarily >> amplicon sequencing!

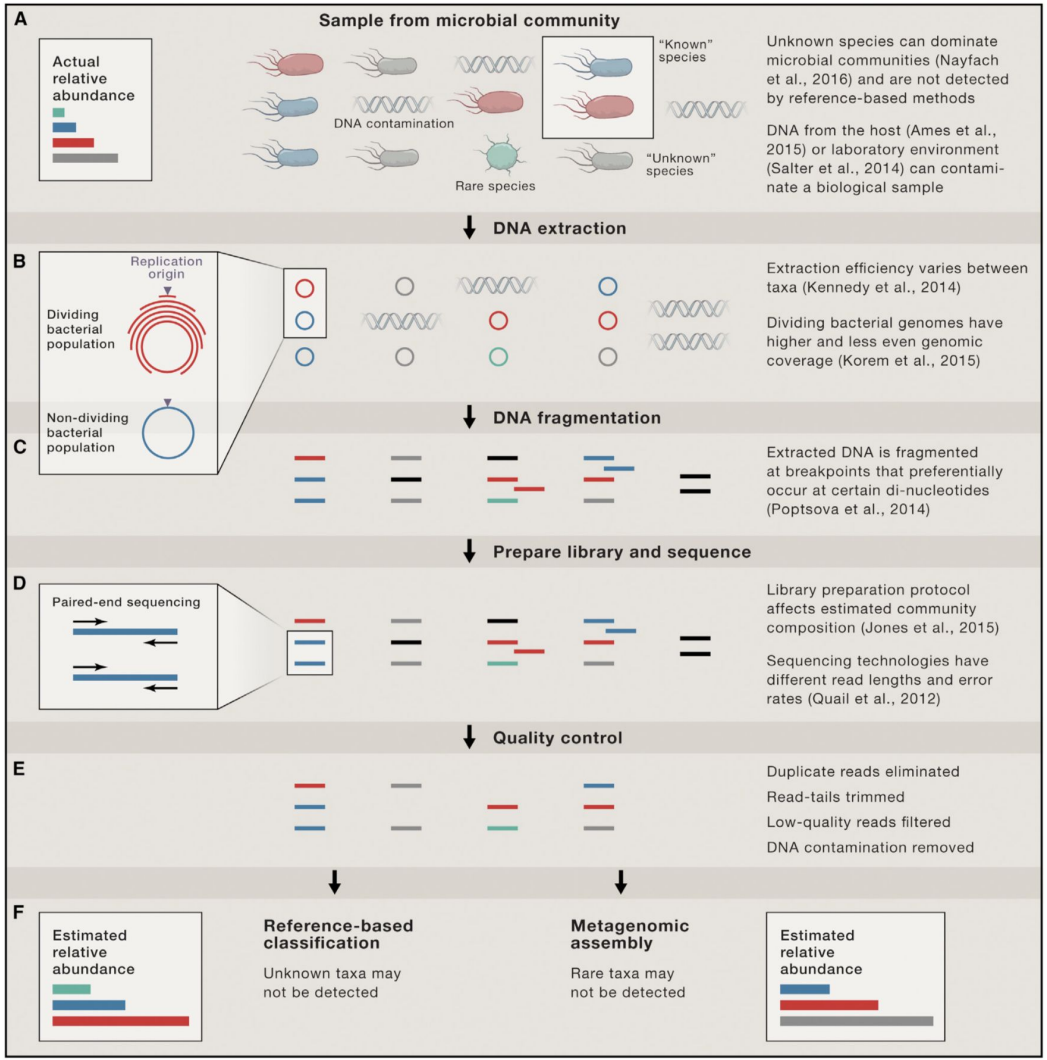
- bacterial species differ in how easily they are lysed
  - and therefore how much DNA they yield during DNA extraction
- bacteria differ in their number of 16S rRNA gene copies
  - how much PCR product we expect to obtain per cell
- Most sources of bias are protocol-dependent
  - PCR primers preferentially amplify different sets of taxa
  - different extraction protocols can produce 10-fold or greater differences in the measured proportion of a taxon from the same sample
- almost every choice in an MGS experiment has been implicated as contributing to bias.

Experimental Strategy



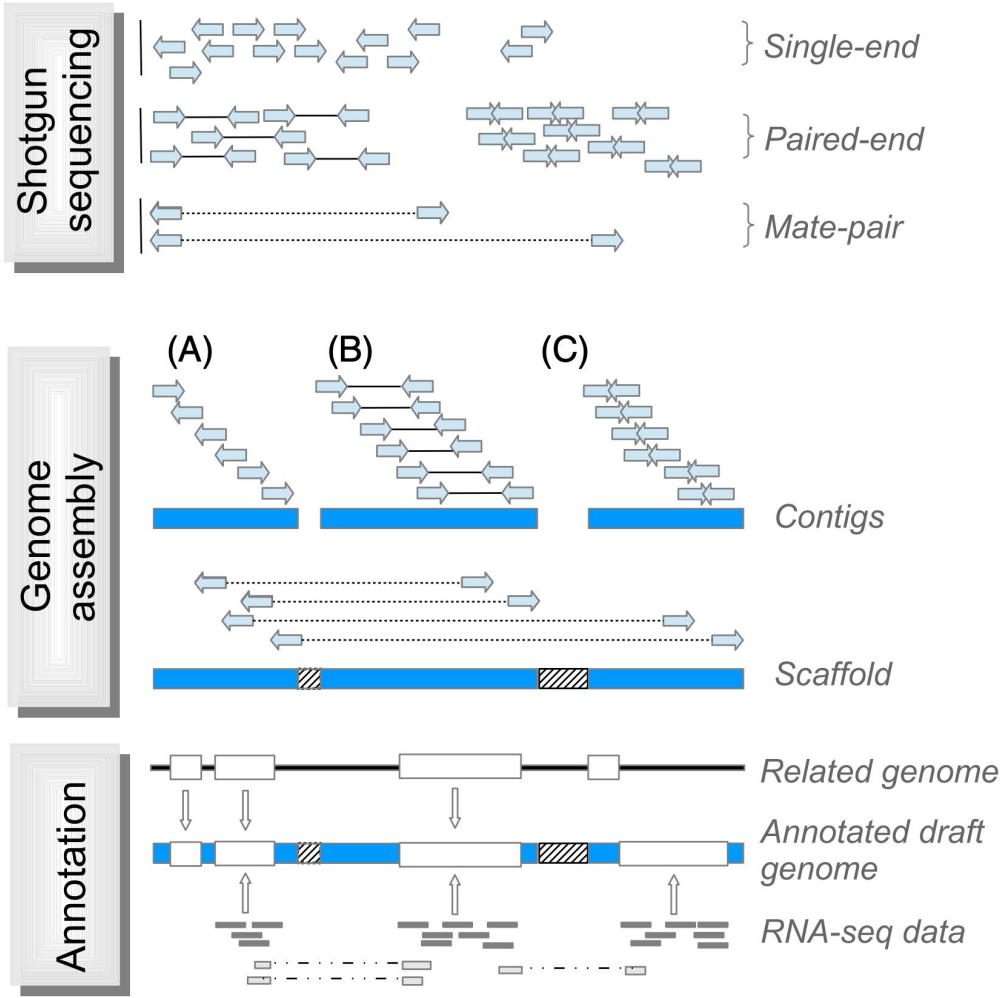


Young, Vincent B. 2017. "The Role of the Microbiome in Human Health and Disease: An Introduction for Clinicians." BMJ 356 (March): j831.

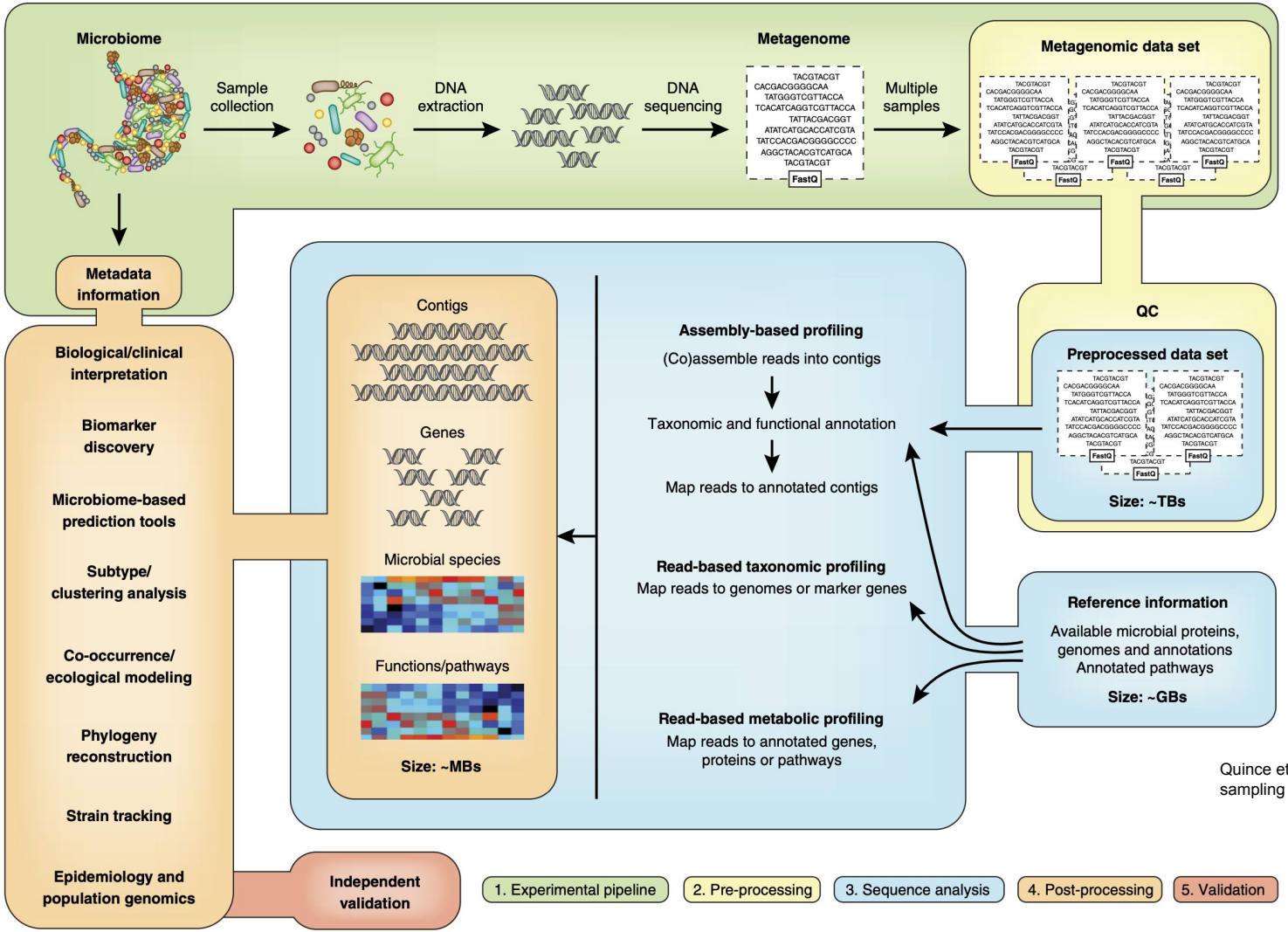


Challenges of metagenome data and analysis

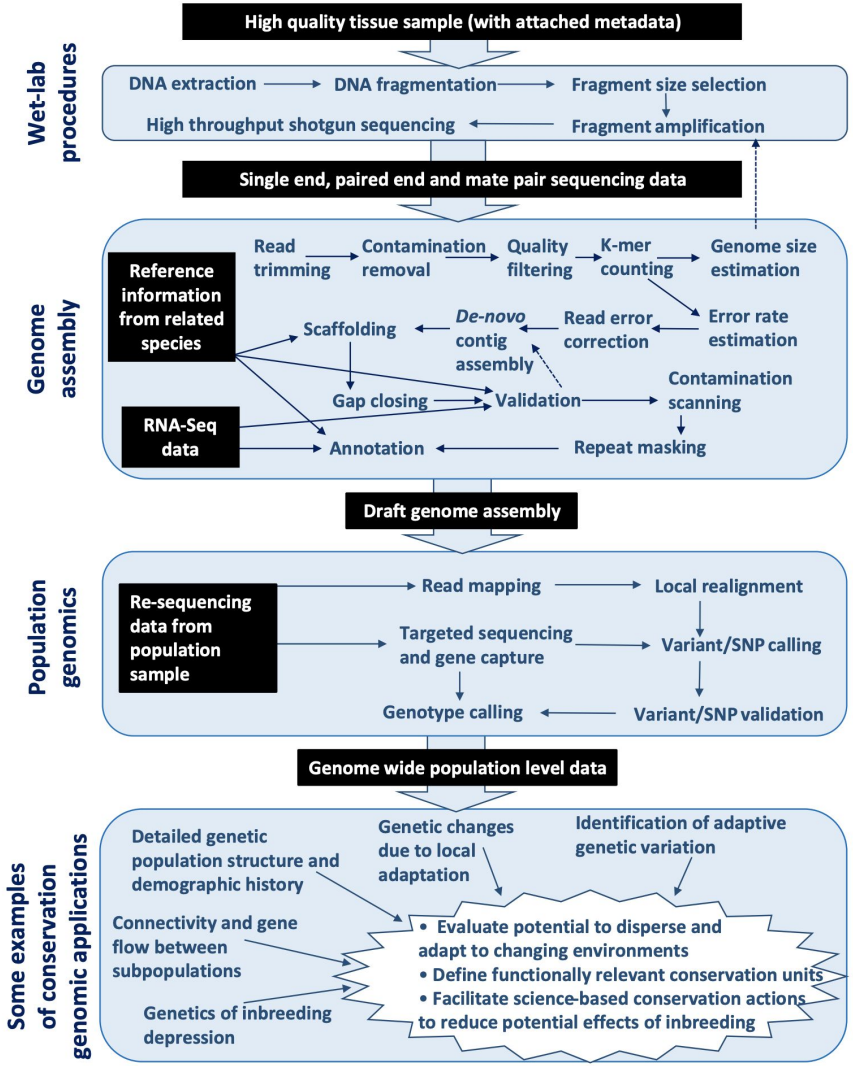
Nayfach, S., and Pollard, K.S. 2016. Toward Accurate and Quantitative Comparative Metagenomics. Cell 166(5): 1103–1116. doi:10.1016/j.cell.2016.08.007.



Ekblom et al. (2014) A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7(9): 1026–1042

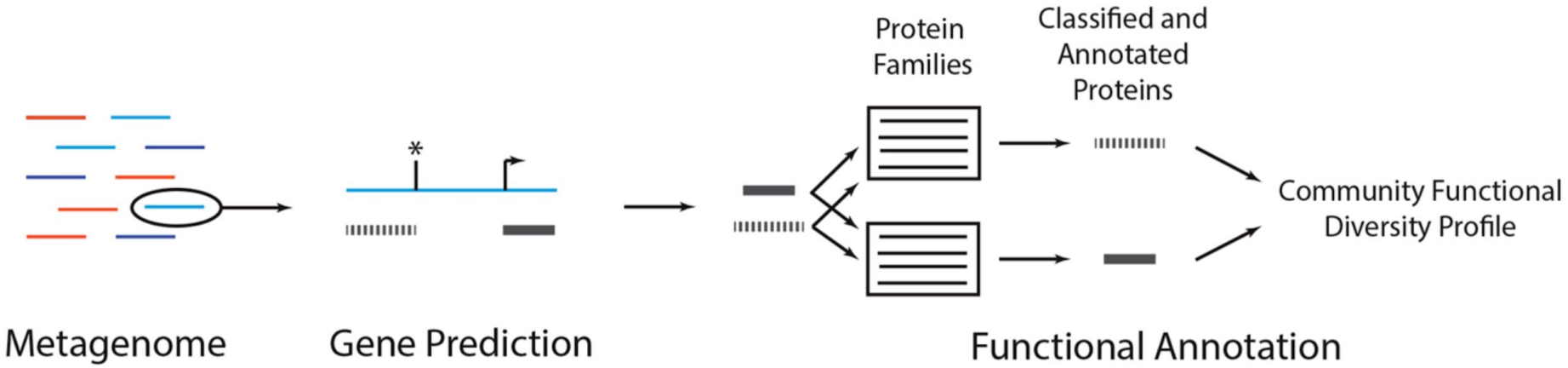


Quince et al. (2017) Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. 35(9): 833–844



Ekblom et al. (2014) A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7(9): 1026–1042

Direct assessment of functional capability.
e.g. prodigal, prokka, etc...



Sharpton (2014) An introduction to the analysis of shotgun metagenomic data. Frontiers in Plant Science 5: 209.

Types of metagenomes:

Single Amplified Genomes (SAGs): *"...are produced by isolating individual cells, amplifying the genome of each cell using whole genome amplification (WGA), and then sequencing the amplified DNA."*

This approach requires specialized instrumentation: *e.g.* flow cytometry, microfluidics, or micromanipulators for single-cell isolation, cleanrooms for downstream handling. Extremely low yields of genomic DNA from a single microbial cell (~1–6 fg). DNA from lysed cells must be amplified to generate enough material for current sequencing technologies

Metagenome-Assembled Genomes (MAGs): *"...are produced using computational binning tools that group assembled contigs into genomes from Gbp-level metagenomic data sets."*

Works best in low diversity datasets. Can produce chimeric genome fragments. Some cases limited to "PanGenome"-level information.

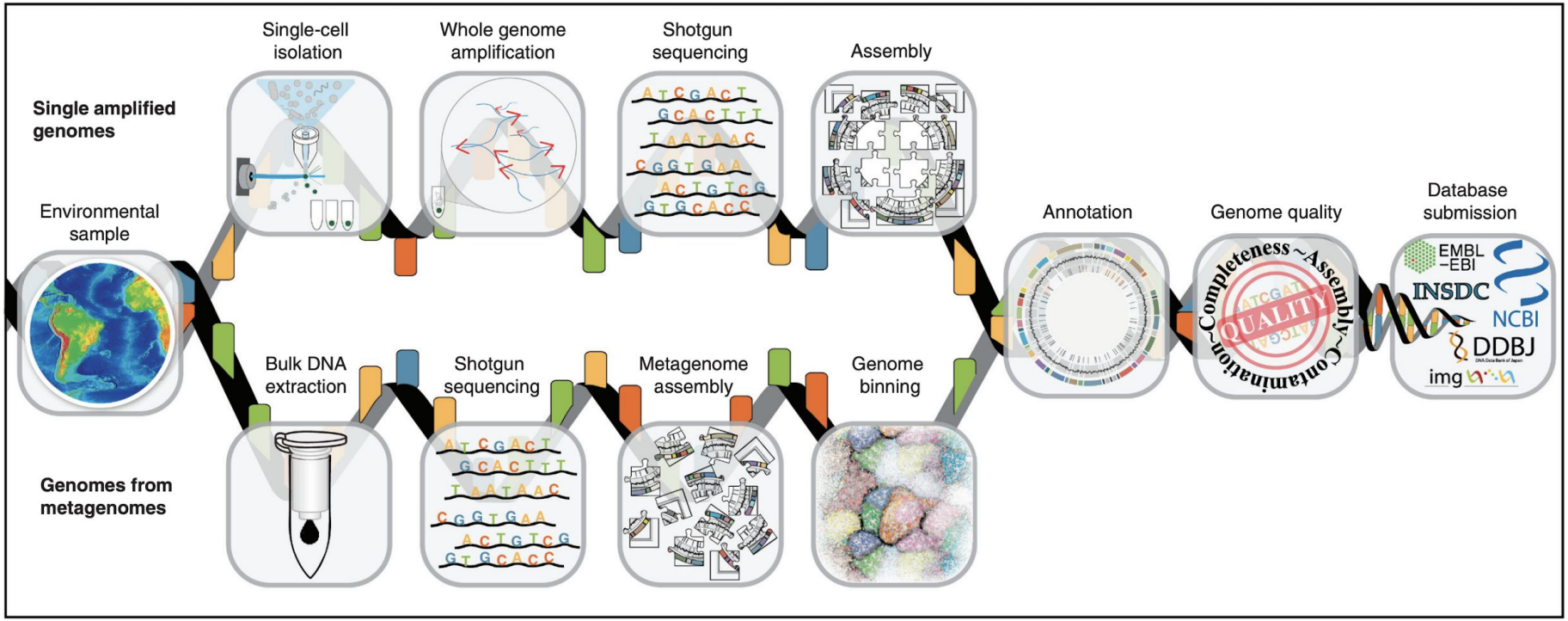


Figure 2 Generation of SAGs and MAGs. Flow diagram outlining the typical pipeline for the production of both SAGs and MAGs.

Bowers et al. 2017. "Minimum Information about a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea." *Nature Biotechnology* 35 (8): 725–31.

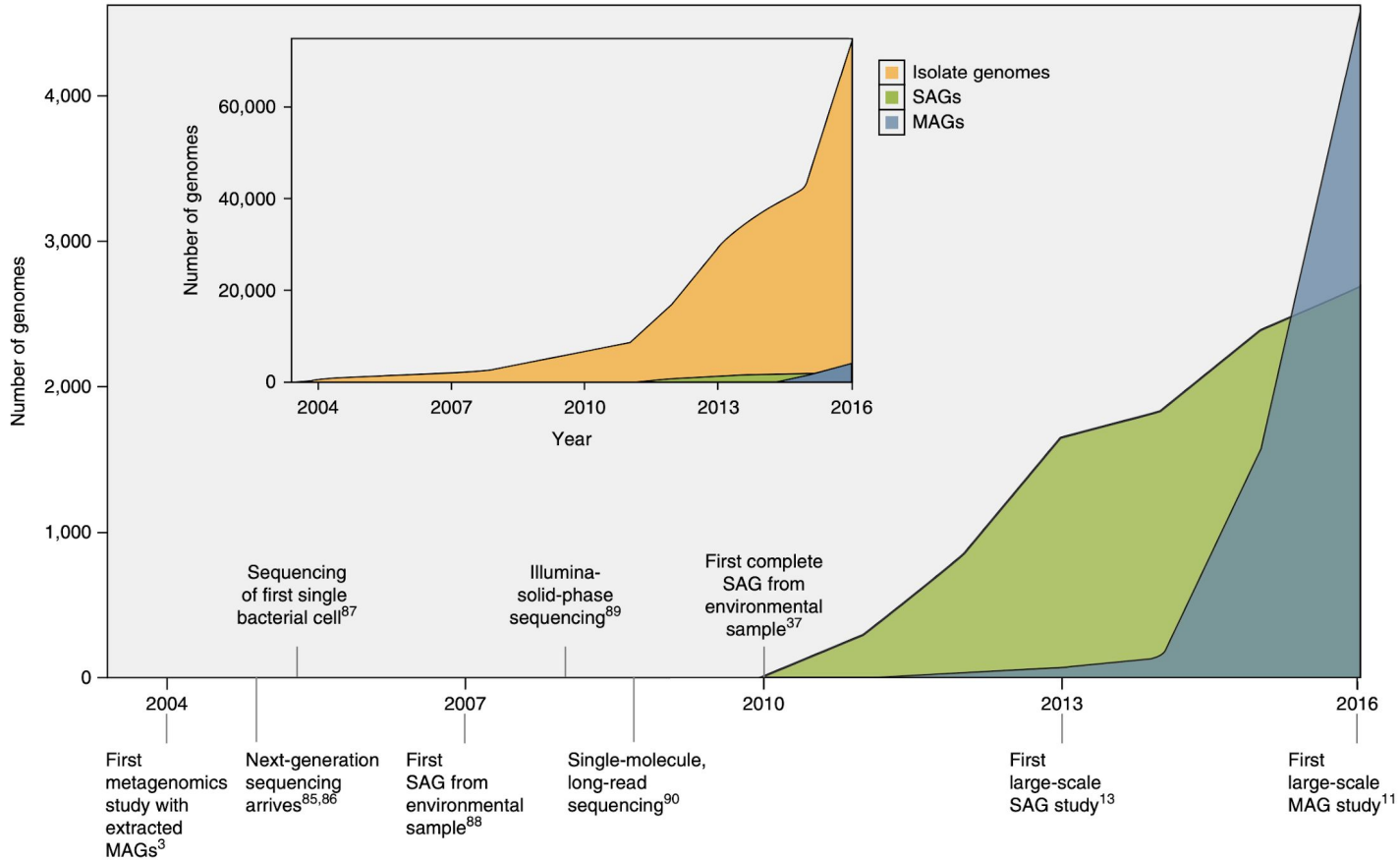


Figure 1 Sequencing of bacterial and archaeal genomes[3,11,13,37,85–90]. Increase in the number of SAGs and MAGs over time. Inset displays the number of isolate genomes over time for comparison. Data for figure were taken from IMG/GOLD[14] in January 2017.

Bowers et al. 2017. "Minimum Information about a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea." *Nature Biotechnology* 35 (8): 725–31.

Genome terminology: Pan / Meta

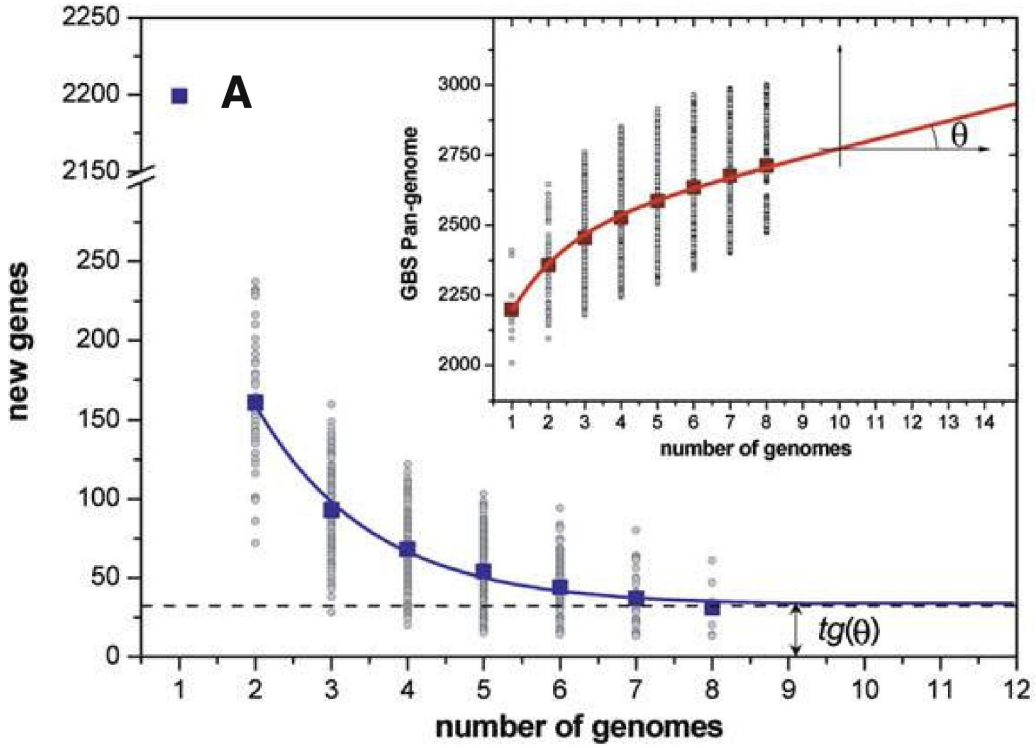
PanGenome: *"... the realization that the genetic repertoire of a [single] biological species, i.e. the pool of genetic material present across the organisms of the species, always exceeds each of the individual genomes and can be, in several cases, "unbounded": an open pangenome"*

Meta-PanGenome: extends the pangenome concept, by incorporating metagenome derived genes and genomes. *"...a representation of the totality of genes belonging to a species identified in multiple metagenomic samplings of a particular habitat"*. That is, the metapangenome is entire sequence space of a species in a given environment.

Pan-MetaGenome: *"...entire collection of all species' meta-pangenomes that exist in a specific environment"*. This is also referred to as the "habitome": the entirety of the genetic landscape of a given habitat.

These data are commonly generated through metagenomic sequencing to produce MAGs.

Tettelin, Hervé, and Duccio Medini, eds. 2020. *The Pangenome: Diversity, Dynamics and Evolution of Genomes*. Cham (CH): Springer.



GBS == Group B *Streptococcus*

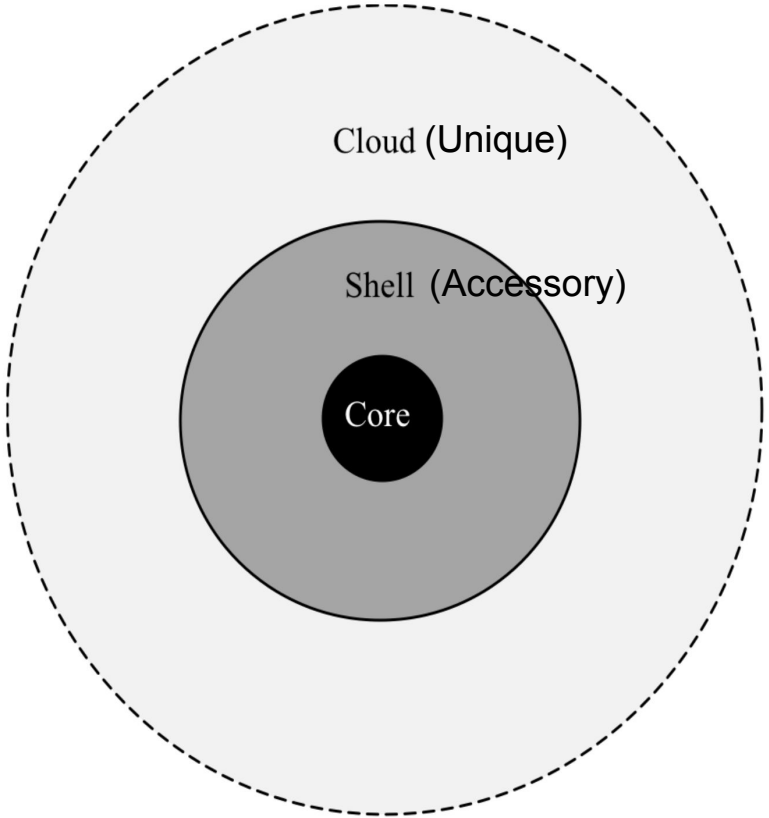
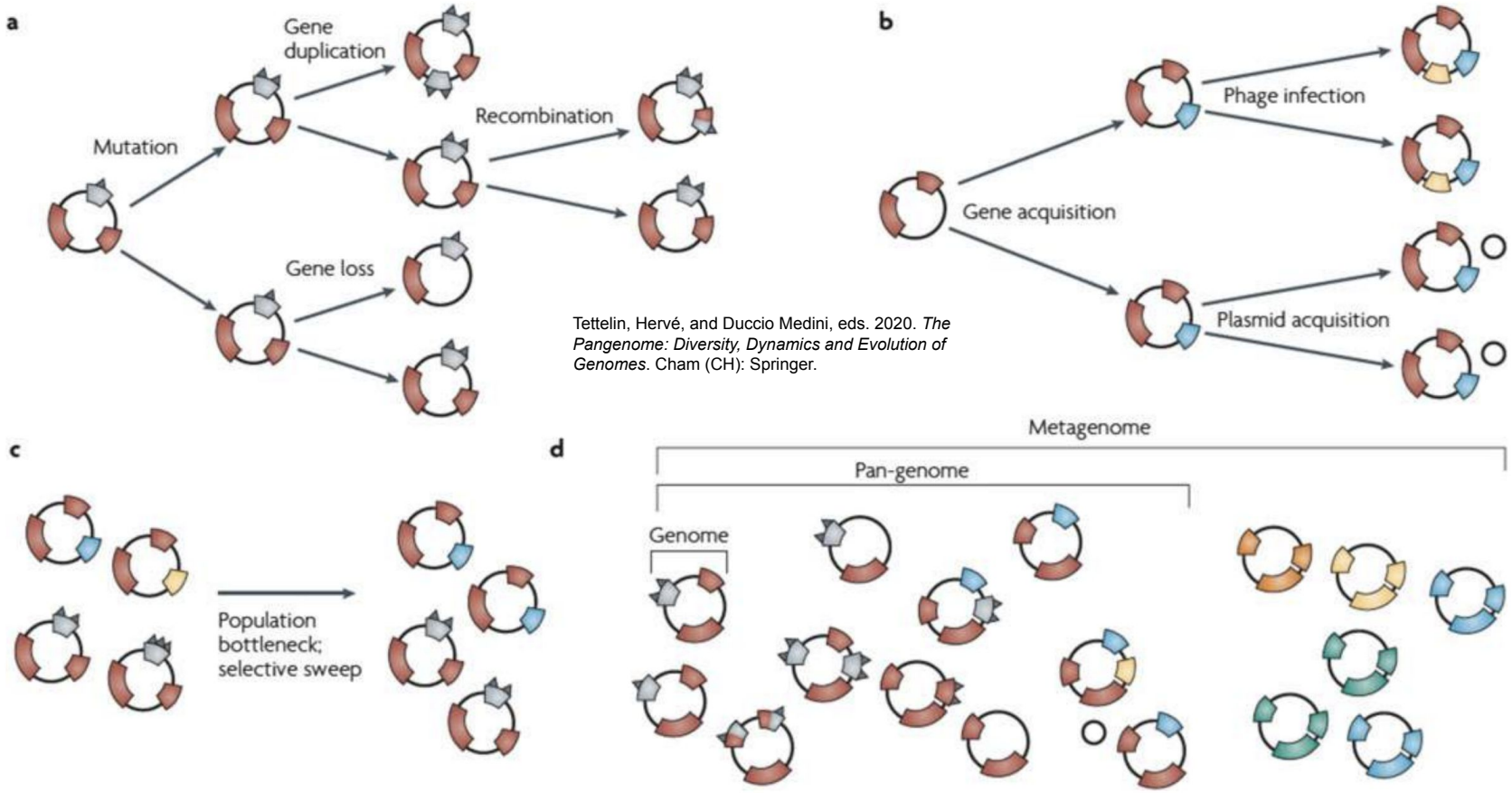


Figure 1. The bacterial pan-genome can be divided into the core (genes always occurring in any genome inside the pan-genome) the shell (genes frequently occurring) and cloud (rarely occurring genes).

Tettelin, Hervé, and Duccio Medini, eds. 2020. *The Pangenome: Diversity, Dynamics and Evolution of Genomes*. Cham (CH): Springer.

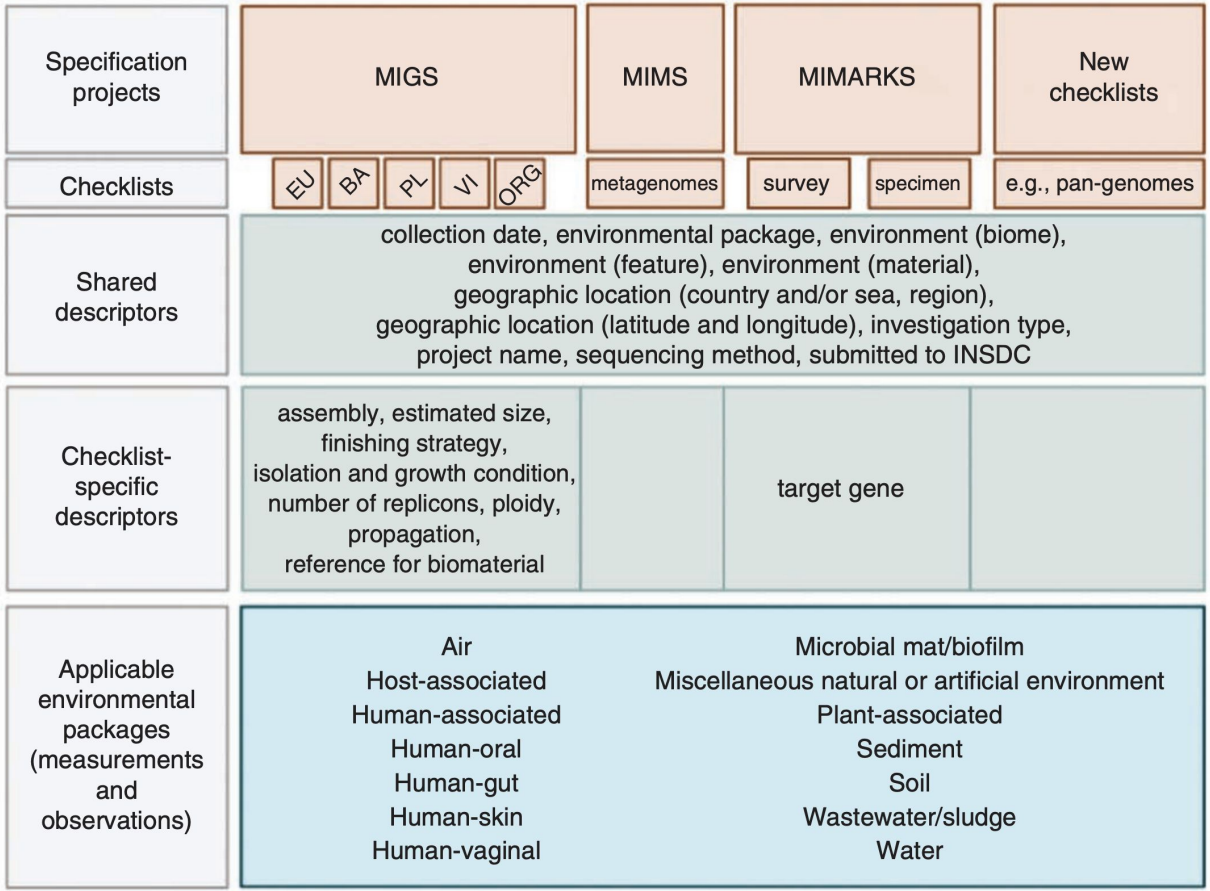
Snipen and Ussery (2010) Standard operating procedure for computing pangenome trees. Standards in Genomic Sciences 2(1): 135–141



Tettelin, Hervé, and Duccio Medini, eds. 2020. *The Pangenome: Diversity, Dynamics and Evolution of Genomes*. Cham (CH): Springer.

Fig. 4 Molecular evolutionary mechanisms that shape bacterial species diversity: one genome, pangenome, and metagenome (Medini et al. 2008). Intra-species (**a**), inter-species (**b**), and population dynamic (**c**) mechanisms manipulate the genomic diversity of bacterial species. For this reason, one genome sequence is inadequate for describing the complexity of species, genera and their interrelationships. Multiple genome sequences are needed to describe the pangenome, which represents, with the best approximation, the genetic information of a bacterial species. Metagenomics embraces the community as the unit of study and, in a specific environmental niche, defines the metagenome of the whole microbial population (**d**)

Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS)



Yilmaz et al. 2011. "Minimum Information about a Marker Gene Sequence (MIMARKS) and Minimum Information about Any (x) Sequence (MIxS) Specifications." *Nature Biotechnology* 29 (5): 415–20.

Minimum Information about a Single Amplified Genome (MISAG) and the Minimum Information about a Metagenome Assembled Genome (MIMAG)

Table 1 Genome reporting standards for SAGs and MAGs

Criterion	Description
Finished (SAG/MAG)	
Assembly quality[a]	Single contiguous sequence without gaps or ambiguities with a consensus error rate equivalent to Q50 or better
High-quality draft (SAG/MAG)	
Assembly quality[a]	Multiple fragments where gaps span repetitive regions. Presence of the 23S, 16S, and 5S rRNA genes and at least 18 tRNAs.
Completion[b]	>90%
Contamination[c]	<5%
Medium-quality draft (SAG/MAG)	
Assembly quality[a]	Many fragments with little to no review of assembly other than reporting of standard assembly statistics.
Completion[b]	≥50%
Contamination[c]	<10%
Low-quality draft (SAG/MAG)	
Assembly quality[a]	Many fragments with little to no review of assembly other than reporting of standard assembly statistics.
Completion[b]	<50%
Contamination[c]	<10%

This is a compressed set of genome reporting standards for SAGs and MAGs. For a complete list of mandatory and optional standards, see **Supplementary Table 1**.

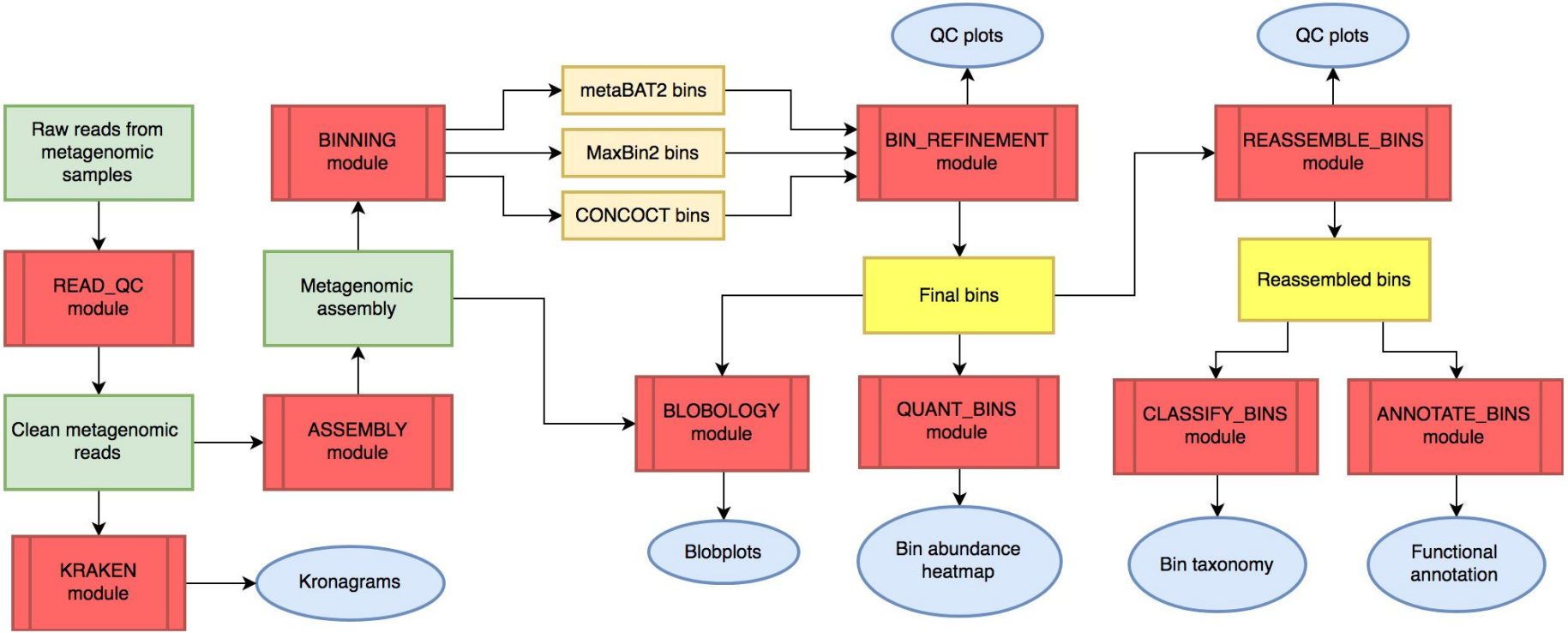
[a]Assembly statistics include but are not limited to: N50, L50, largest contig, number of contigs, assembly size, percentage of reads that map back to the assembly, and number of predicted genes per genome. [b]Completion: ratio of observed single-copy marker genes to total single-copy marker genes in chosen marker gene set. [c]Contamination: ratio of observed single-copy marker genes in ≥2 copies to total single-copy marker genes in chosen marker gene set.

Bowers et al. 2017. "Minimum Information about a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea." *Nature Biotechnology* 35 (8): 725–31.

Tools of the trade. *The tools are legion.*

- Anvi'o
- metaWRAP
- biobakery suite
  - HUMAnN
  - MEtaPhlAn
  - PhyloPhlAn
  - GraPhlAn
- MEGAN
  - DIAMOND
  - MALT
- MetAMOS
- Kraken
- Centrifuge
- BBtools
- DAS tool
- CONCOCT
- MetaBAT
- metaSPAdes
- PhyloSift
- PhyloPythiaS+
- MG-RAST
- PAUDA
- ShortBRED
- CIRCOS
- MG-RAST
- Integrated Microbial Genomes and Microbiomes (IMG/M)
- MicrobiomeAnalyst
- iRep
- ANI
- Krona
- ...

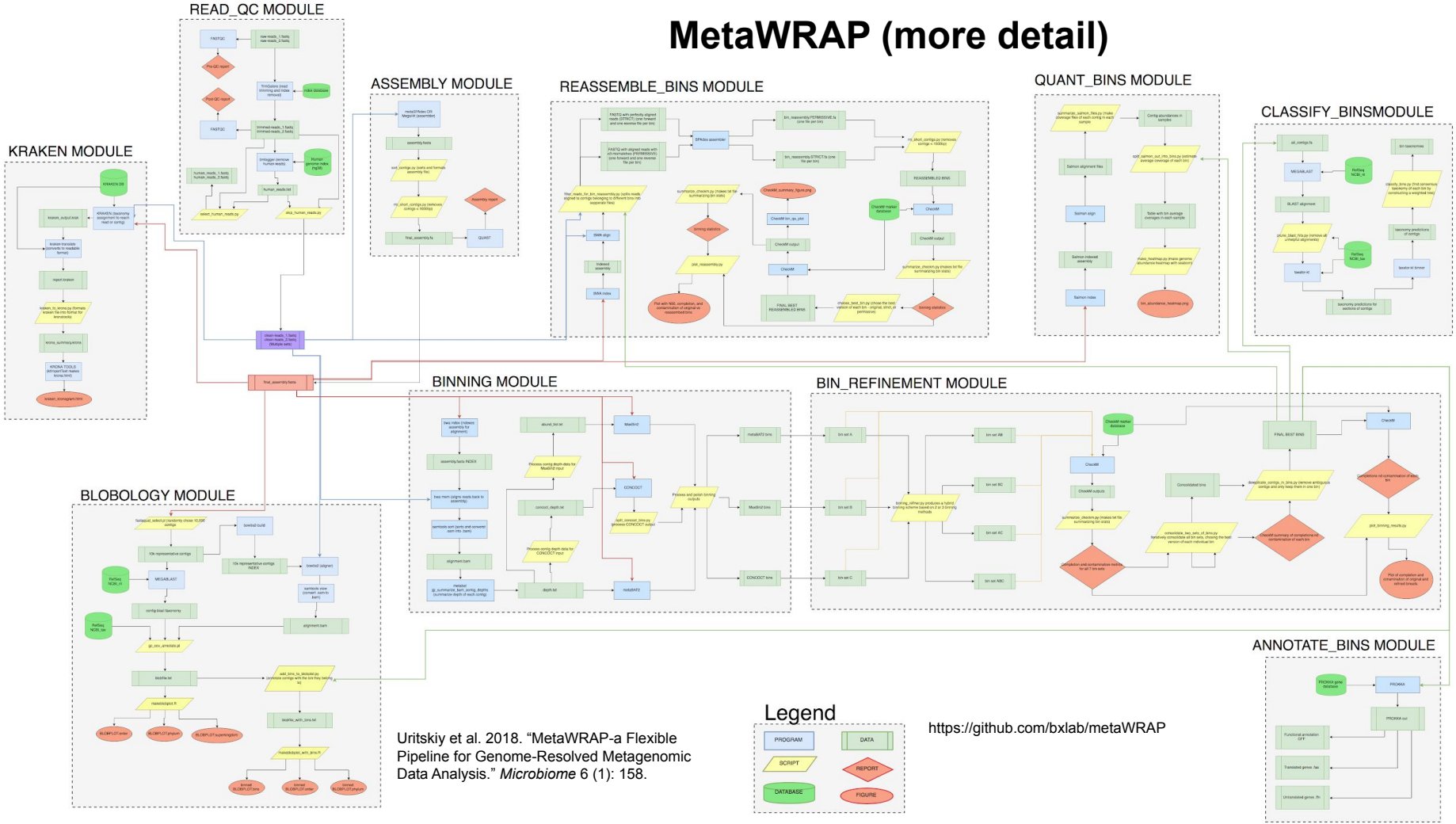
MetaWRAP (basic overview)



https://github.com/bxlab/metaWRAP

Uritskiy et al. 2018. “MetaWRAP-a Flexible Pipeline for Genome-Resolved Metagenomic Data Analysis.” *Microbiome* 6 (1): 158.

MetaWRAP (more detail)



Uritskiy et al. 2018. “MetaWRAP-a Flexible Pipeline for Genome-Resolved Metagenomic Data Analysis.” *Microbiome* 6 (1): 158.

https://github.com/bxlab/metaWRAP

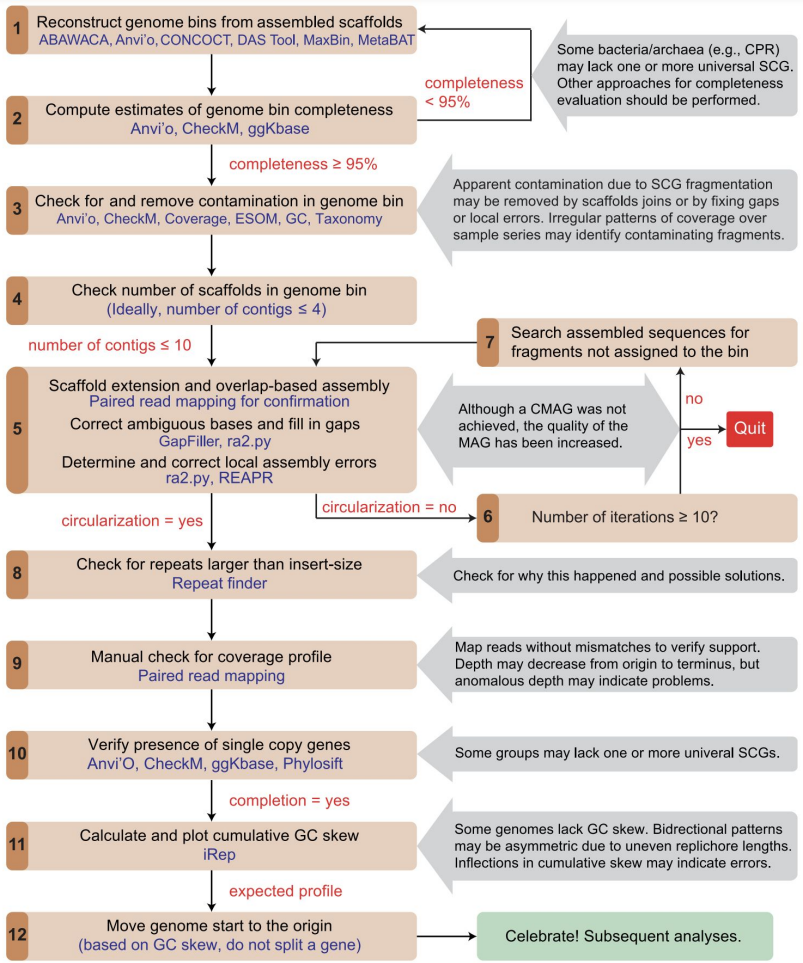


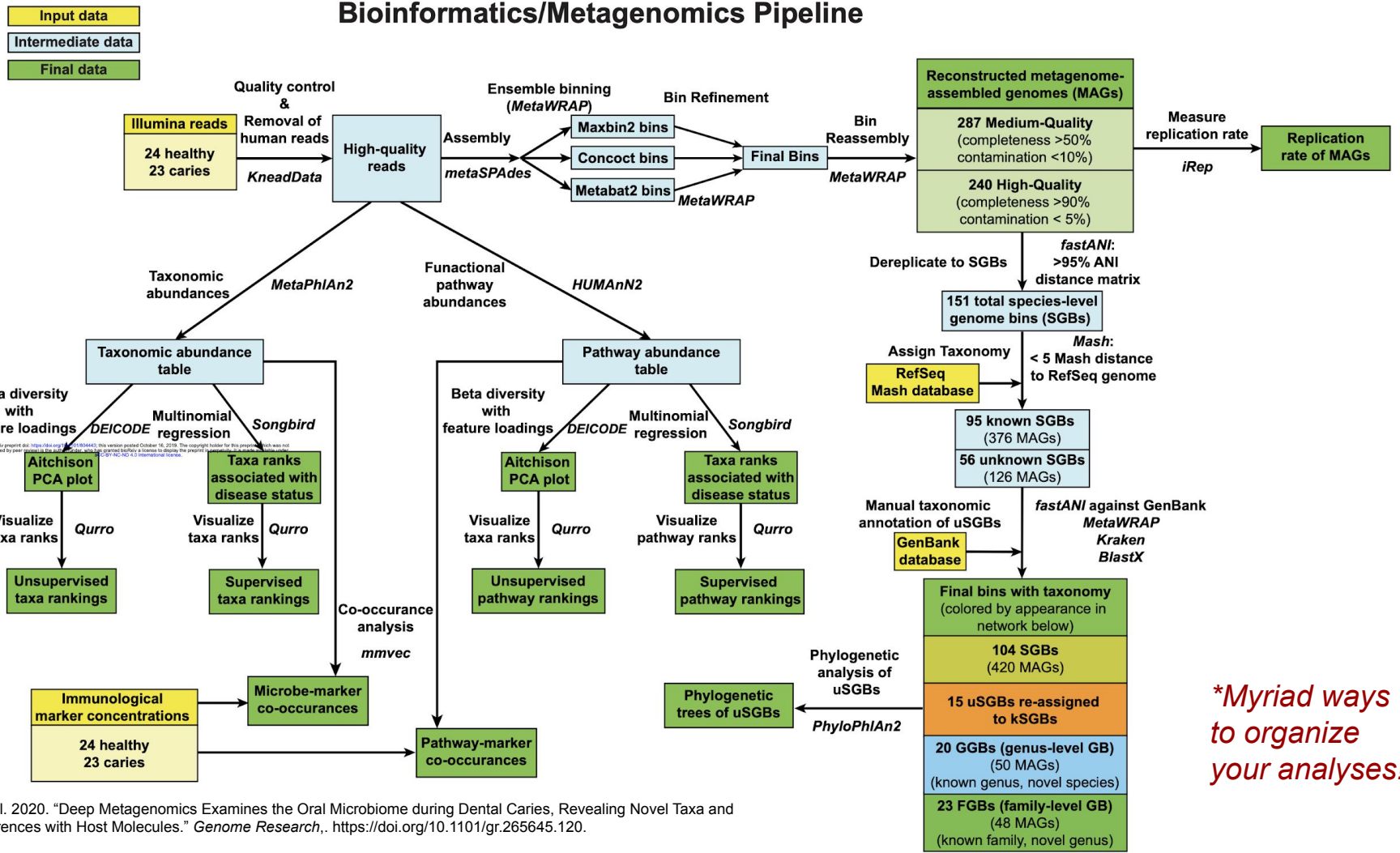
Figure 5. The workflow for generating curated and complete genomes from metagenomes. Steps are shown in black, and the tools or information used in blue. Notes for procedures are shown in gray boxes. The detailed procedures for scaffold extension and gap closing are available in the Supplemental Methods and also online (https://ggkbase-help.berkeley.edu/genome_curation/scaffold-extension-and-gap-closing/).

Chen et al.. 2020. "Accurate and Complete Genomes from Metagenomes." *Genome Research* 30 (3): 315–33.

Shaiber et al. 2019. "Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories." *mBio* 10 (3): e00725–19.

Bowers et al. 2017. "Minimum Information about a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea." *Nature Biotechnology* 35 (8): 725–31.

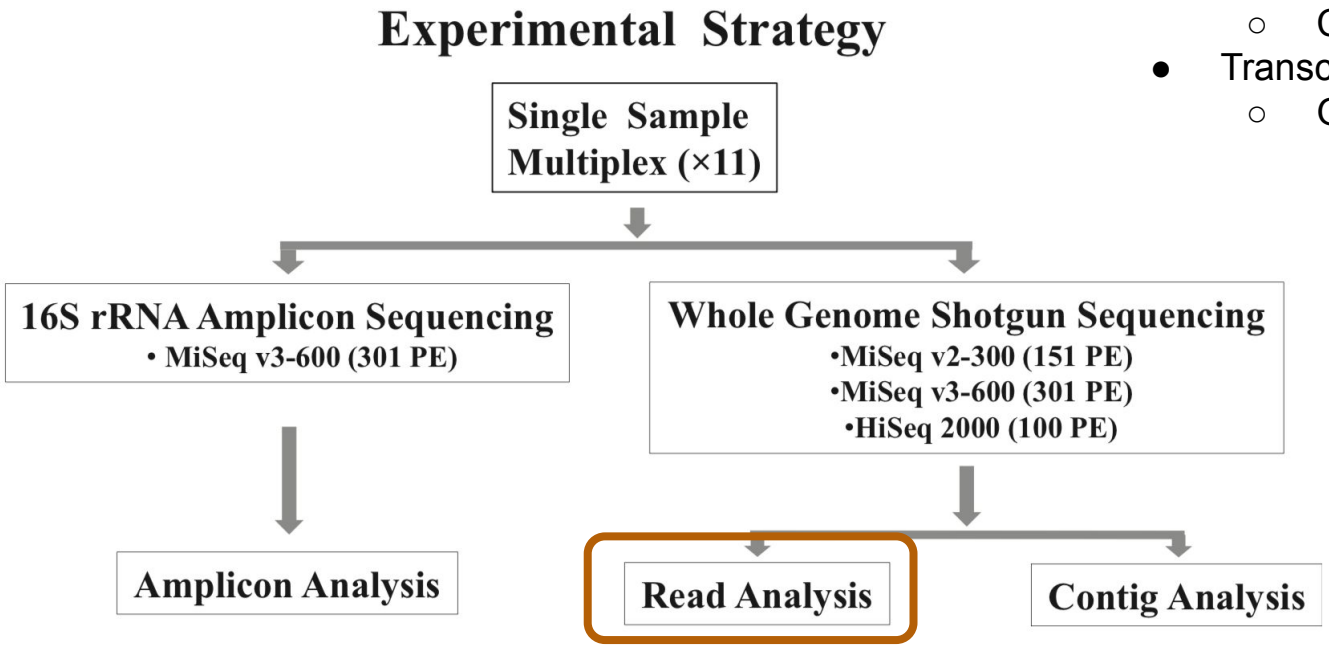
Bioinformatics/Metagenomics Pipeline



Myriad ways to organize your analyses!

Baker et al. 2020. “Deep Metagenomics Examines the Oral Microbiome during Dental Caries, Revealing Novel Taxa and Co-Occurrences with Host Molecules.” *Genome Research*,. https://doi.org/10.1101/gr.265645.120.

Read mapping approaches



`Omics use cases:

- Classify many short reads.
  - Coding sequence / pathways
- Transcriptome (RNASeq)
  - Gene / pathway expression

Classifying / Analyzing shotgun reads

Pros:

- Overcome primer / amplicon biases
- Can use if unable make assemblies
- Increased taxonomic specificity
- Best for viromics (via recruitment plots)

Cons:

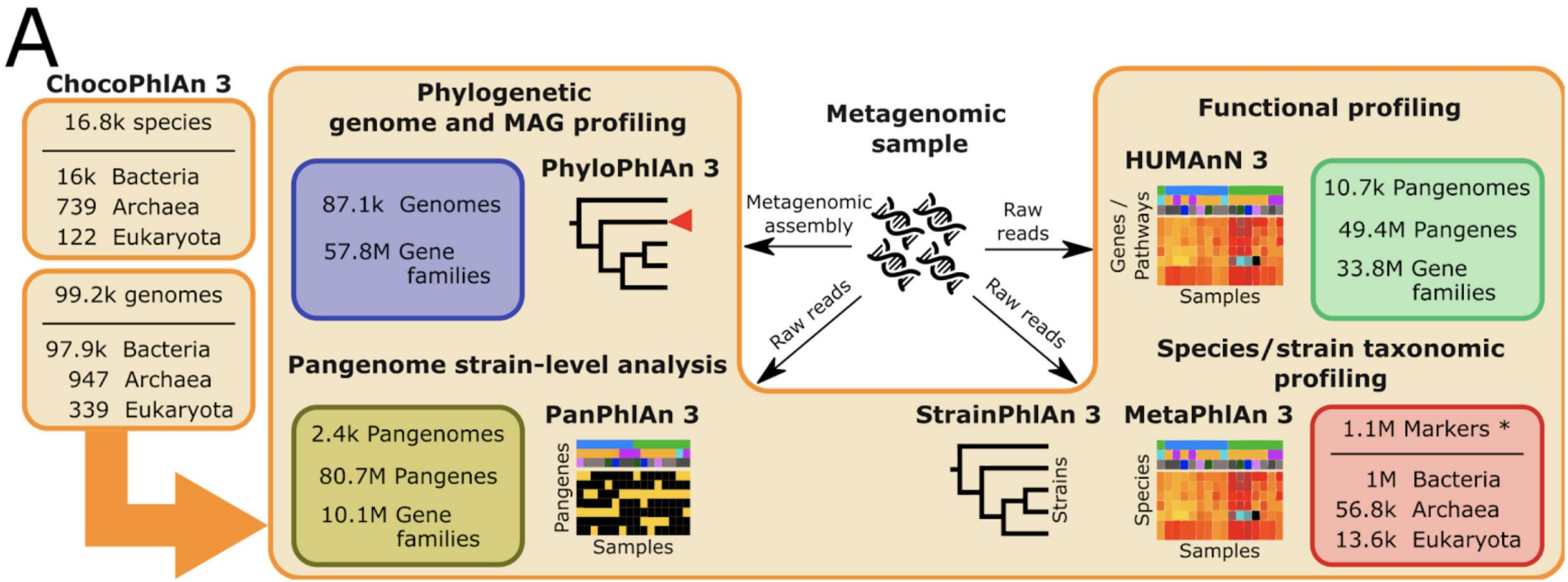
- Unclear how to normalize data for community analysis
- Cost prohibitive when tasked with many samples
- Harder to limit sequencing of host as you can with amplicon surveys (i.e. blocking primer, PNA)

Beghini, *et al*. 2021. "Integrating Taxonomic, Functional, and Strain-Level Profiling of Diverse Microbial Communities with bioBakery 3." *eLife* 10. https://doi.org/10.7554/eLife.65088.

Huttenhower's bioBakery:
https://github.com/biobakery/biobakery/wiki

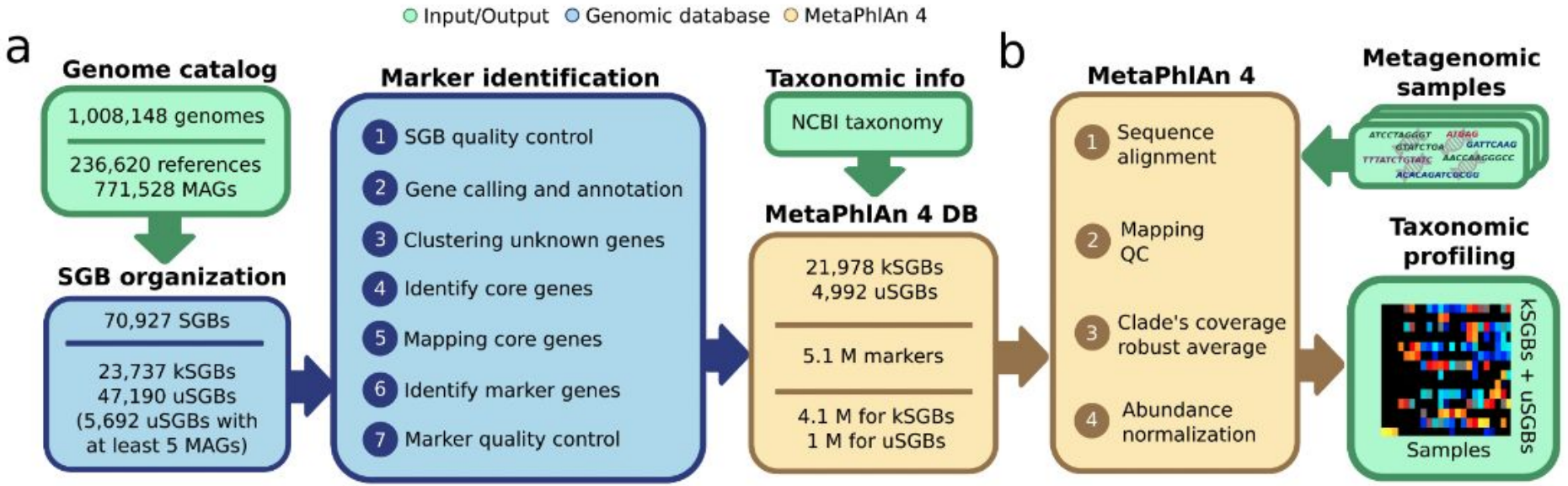
Forum:
https://forum.biobakery.org/

Biobakery



https://github.com/biobakery/biobakery/wiki

Extending and Improving Metagenomic Taxonomic Profiling with Uncharacterized Species with MetaPhlAn 4



SGB: species-level genome bins
kSGB: known SGB
uSGB: unknown SGB

Blanco-Miguez *et al*. 2022. "Extending and Improving Metagenomic Taxonomic Profiling with Uncharacterized Species with MetaPhlAn 4." *bioRxiv*. https://doi.org/10.1101/2022.08.22.504593.

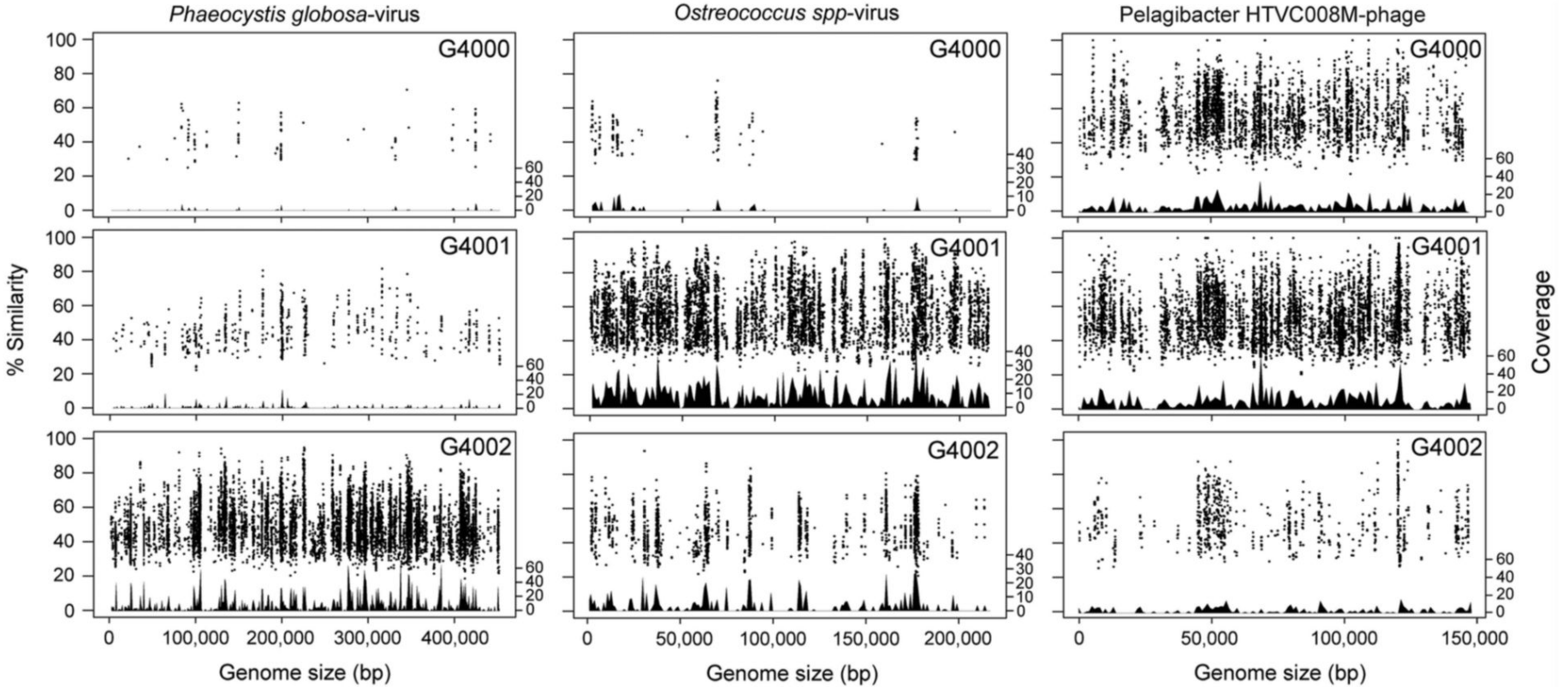
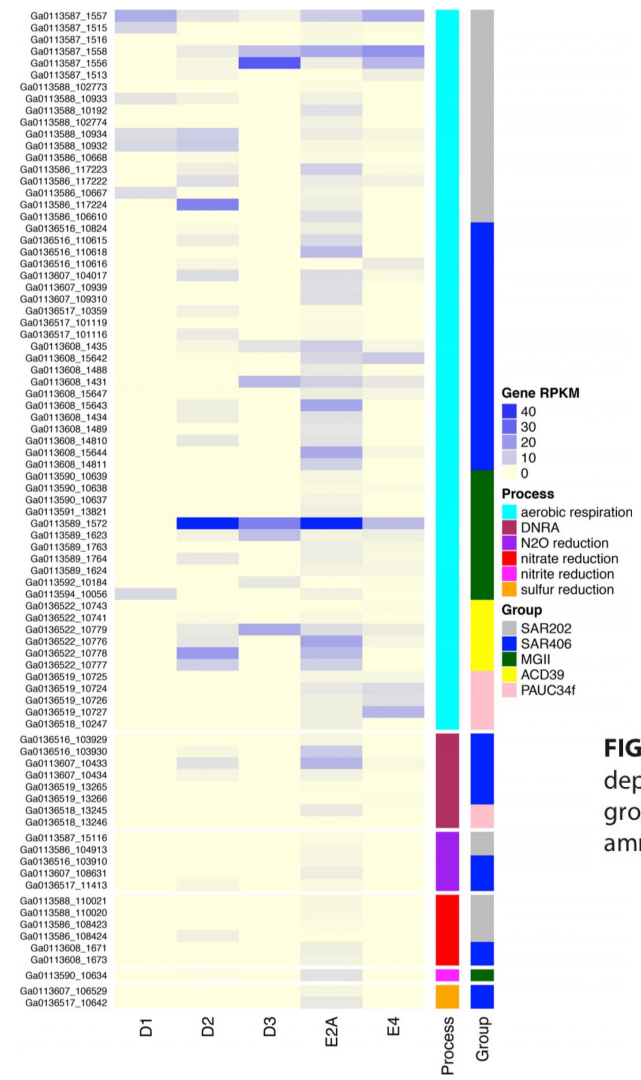


Figure 4 Fragment recruitment of sorted metagenomes to selected viral genomes. Similarity values are based on tBLASTx alignments, and coverage across each genome was plotted using a sliding window of 4000 bp.

RNASeq Recruitment Plot

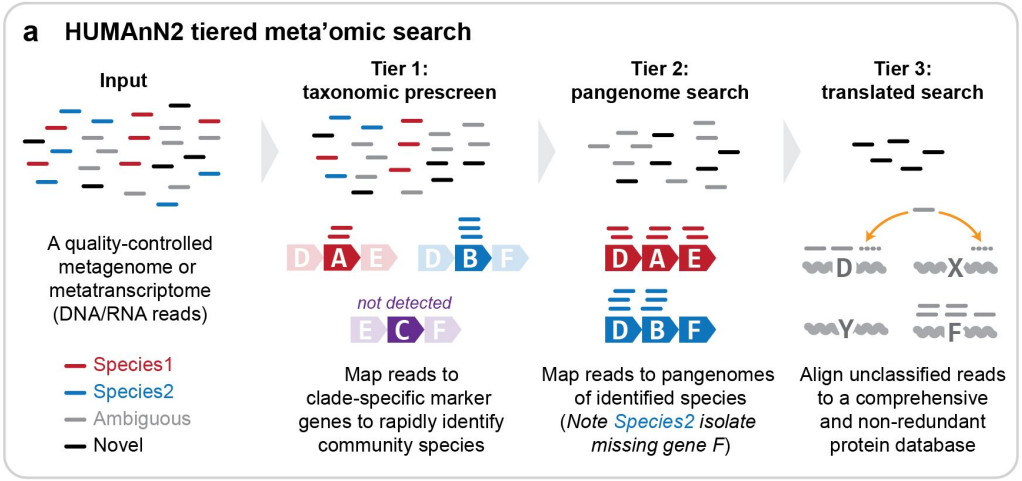


Easier / faster approach if you have explicit questions about microbes of interest.

FIG 3 Expression of predicted respiratory genes. RPKM values of RNA recruitment for each gene, by sample, are depicted with colors according to the Gene RPKM key (pale yellow to blue shows increasing intensity). Genes are grouped by bin, taxonomic affiliation, and specific respiratory process. DNRA, dissimilatory nitrite reduction to ammonia.

Thrash (2017) Metabolic Roles of Uncultivated Bacterioplankton Lineages in the Northern Gulf of Mexico "Dead Zone". MBio 8(5)

HUMAnN example



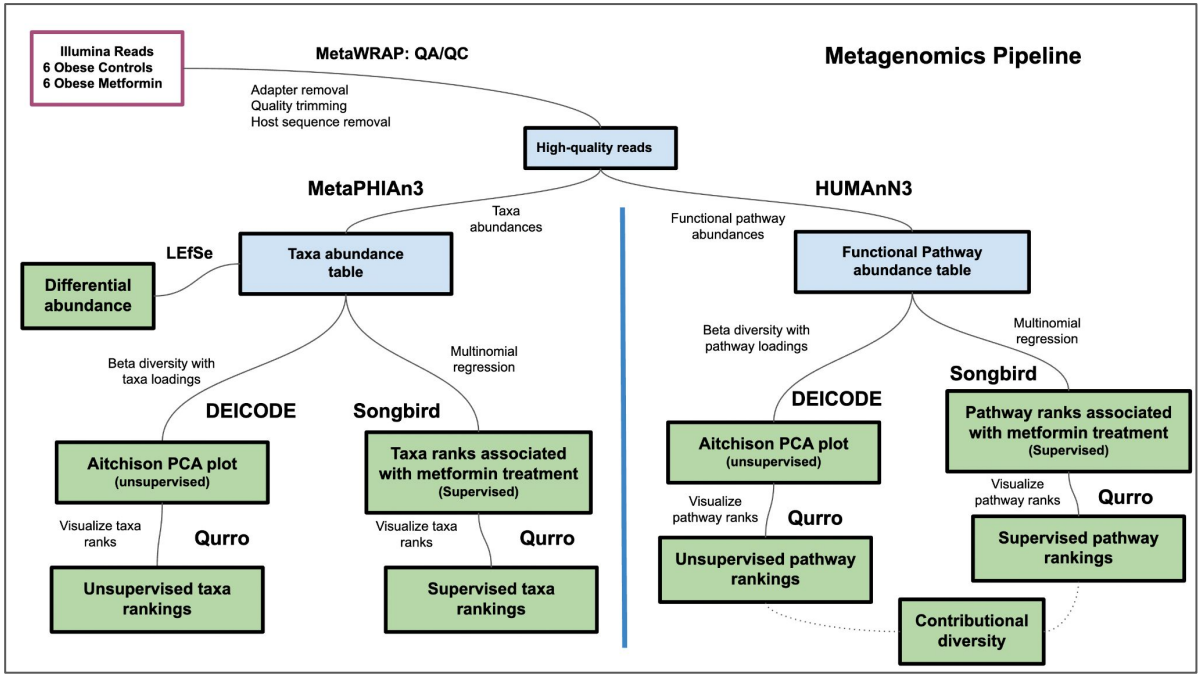
https://github.com/biobakery/biobakery/wiki/humann3

Beghini, *et al*. 2021. "Integrating Taxonomic, Functional, and Strain-Level Profiling of Diverse Microbial Communities with bioBakery 3." *eLife* 10. https://doi.org/10.7554/eLife.65088.

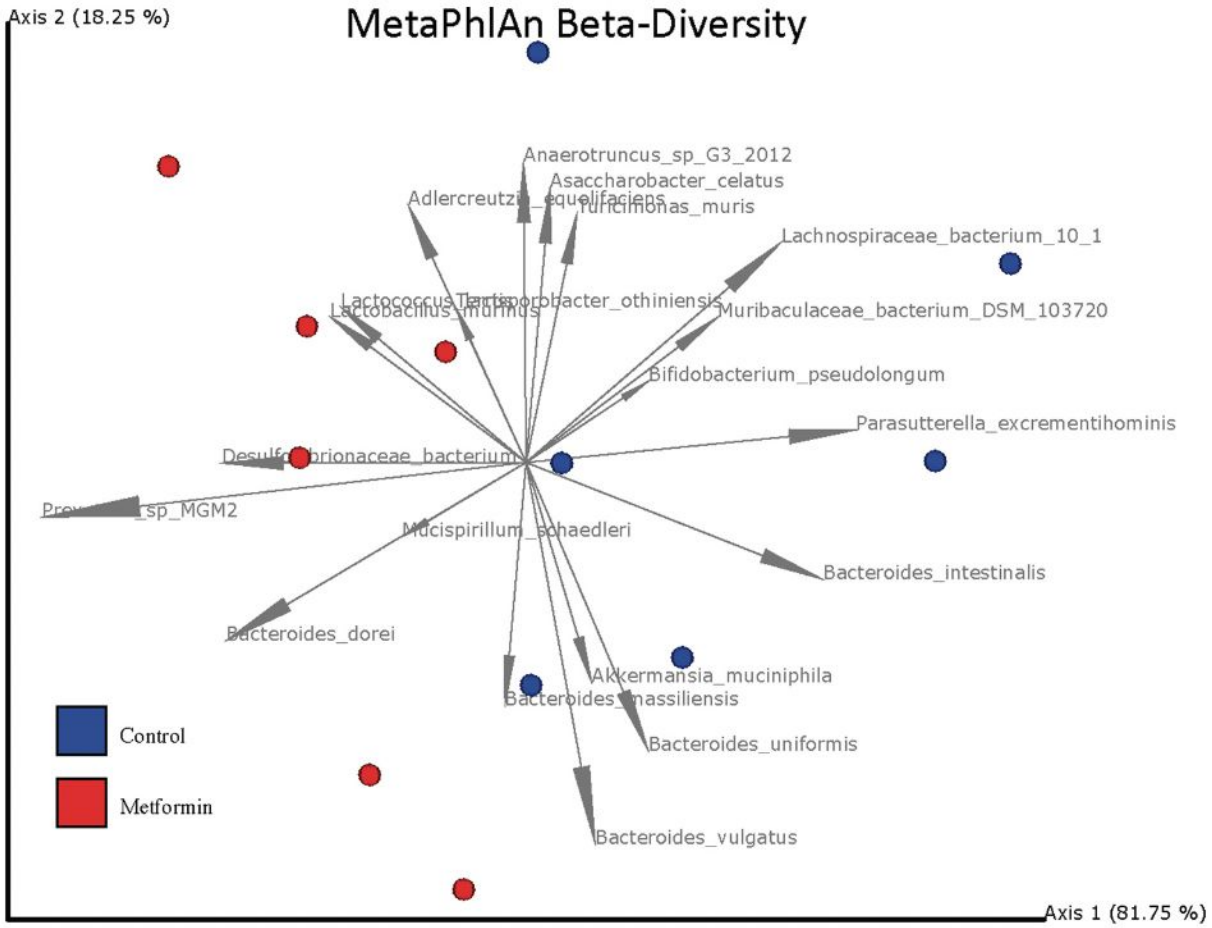
Mixing tools.

MataWRAP + bioBakery A meta'omic analysis environment + qiime2

=

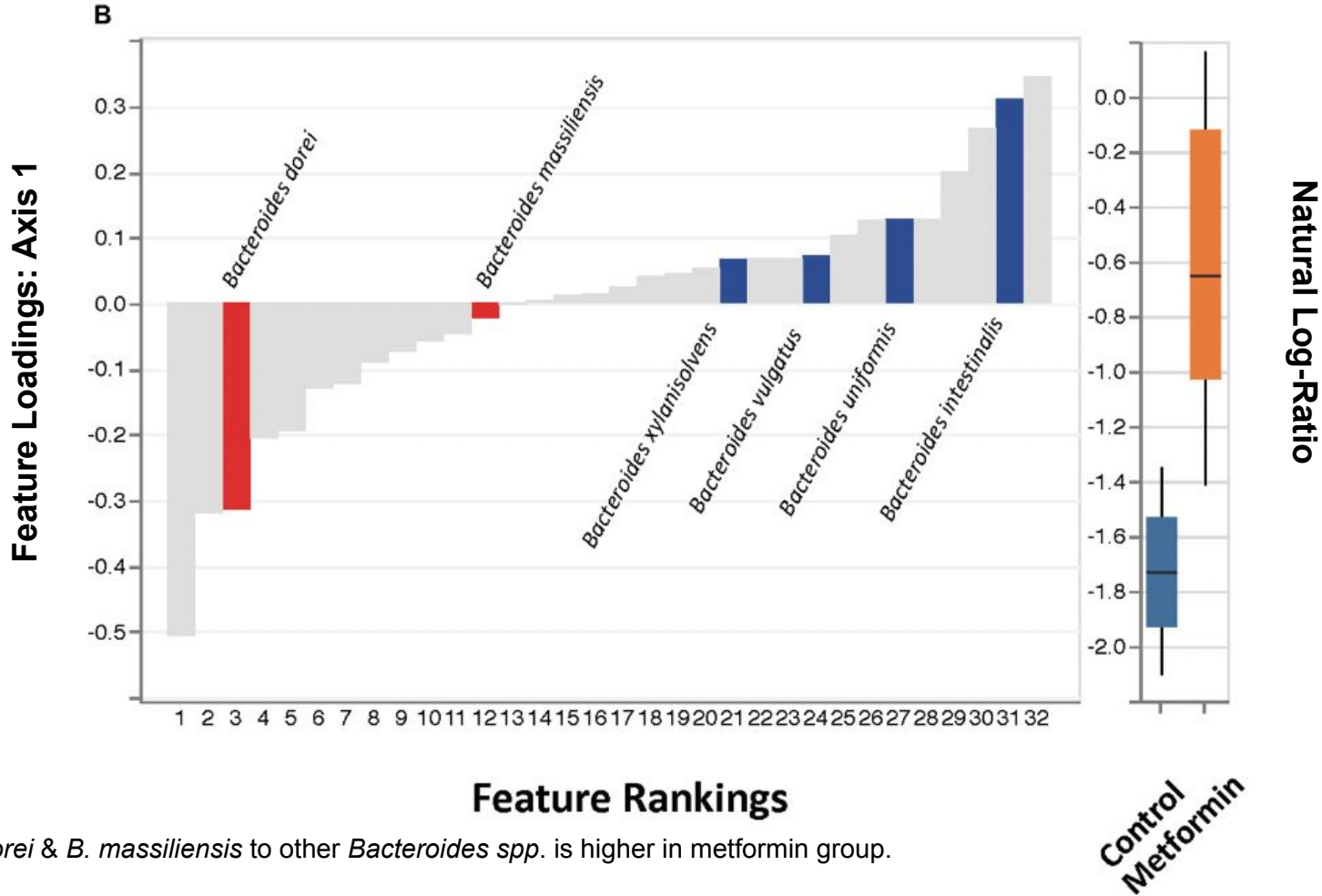


Short-Term Metformin Treatment Enriches *Bacteroides dorei* in an Obese Liver Steatosis Zucker Rat Model



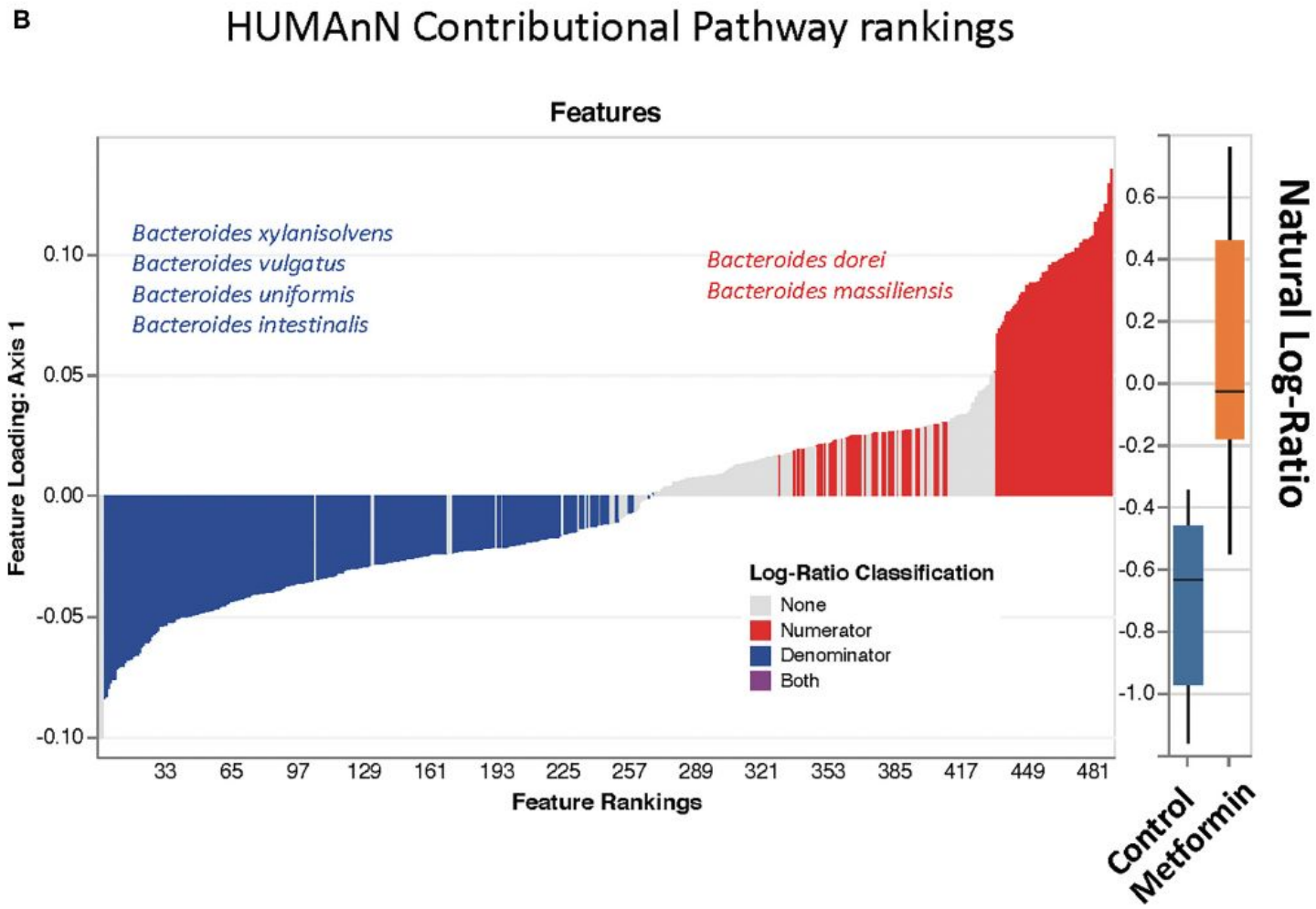
Robeson *et al*. 2022. “Short-Term Metformin Treatment Enriches Bacteroides Dorei in an Obese Liver Steatosis Zucker Rat Model.” *Frontiers in Microbiology* 13 (March): 834776. http://dx.doi.org/10.3389/fmicb.2022.834776

Bacteroides dorei continued

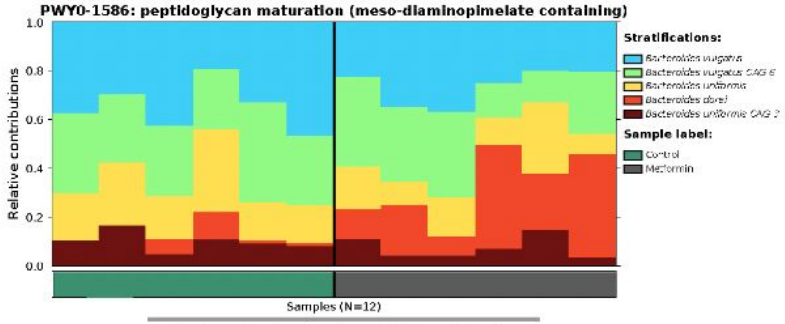


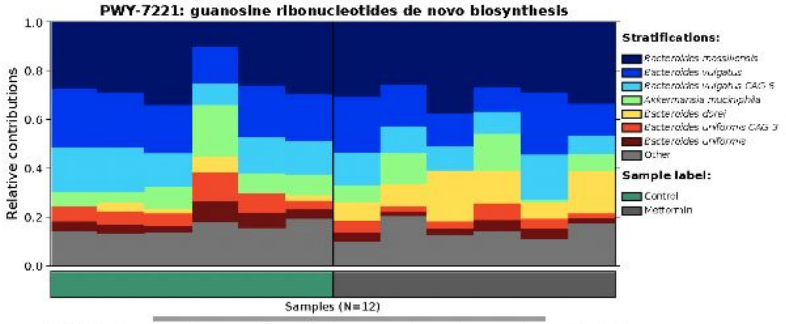
Ratio of *B. dorei* & *B. massiliensis* to other *Bacteroides spp*. is higher in metformin group.

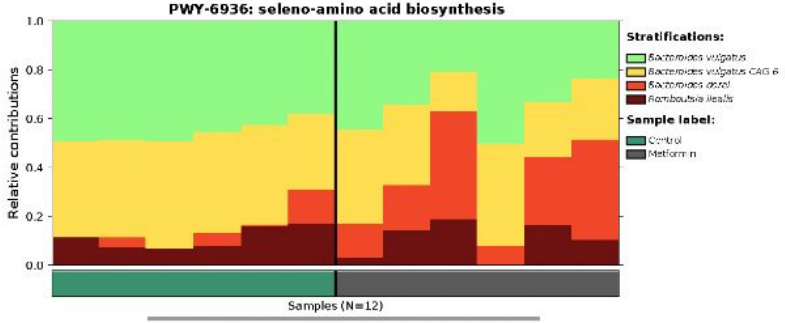
Bacteroides dorei continued

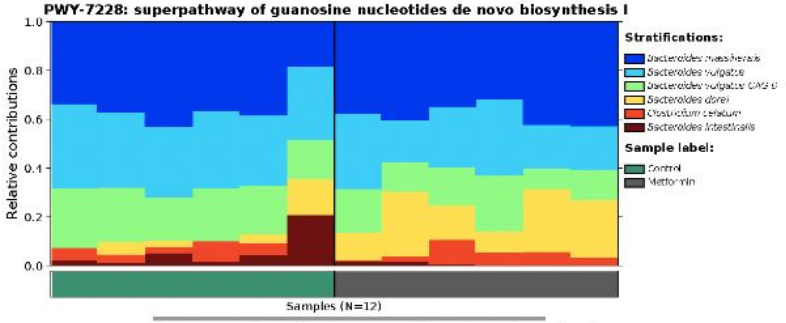


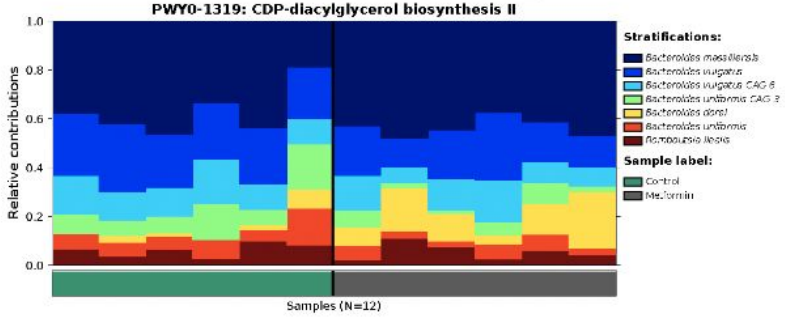
Bacteroides dorei continued

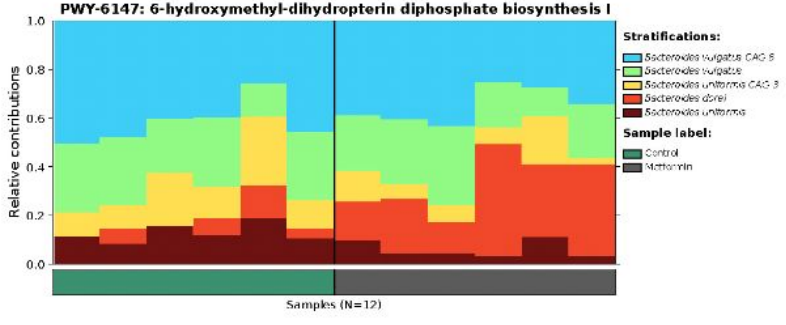




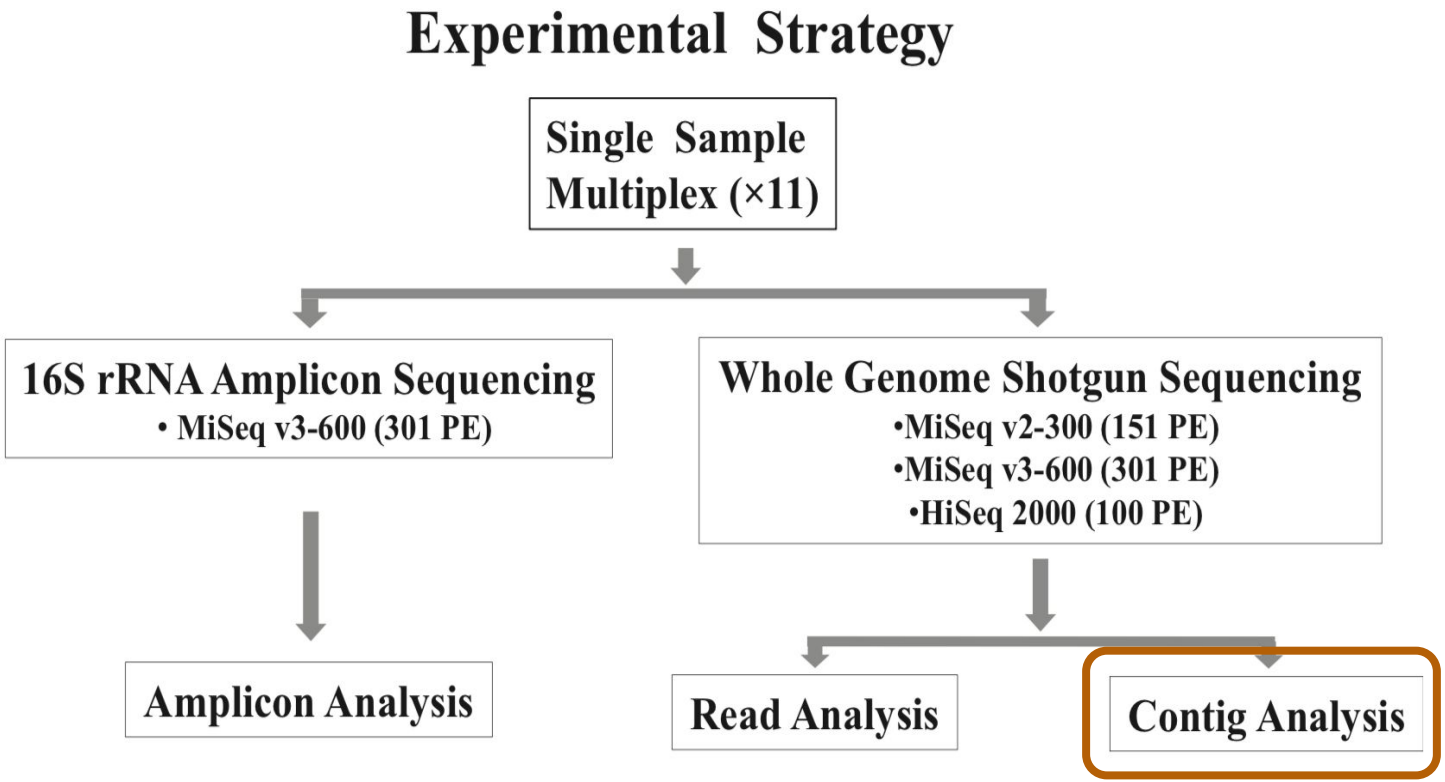


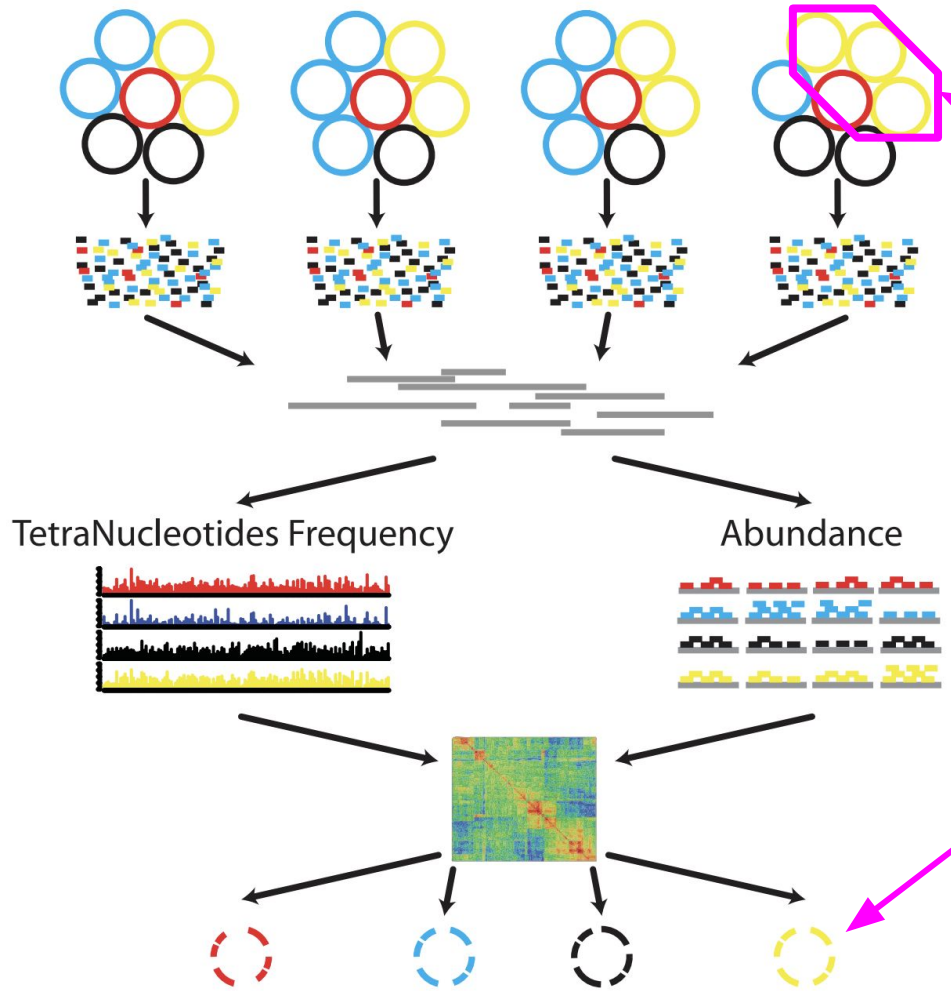






Genome Assembly





Preprocessing

1. Samples from multiple sites or times
2. Metagenome libraries
3. Initial de-novo assembly using the combined library

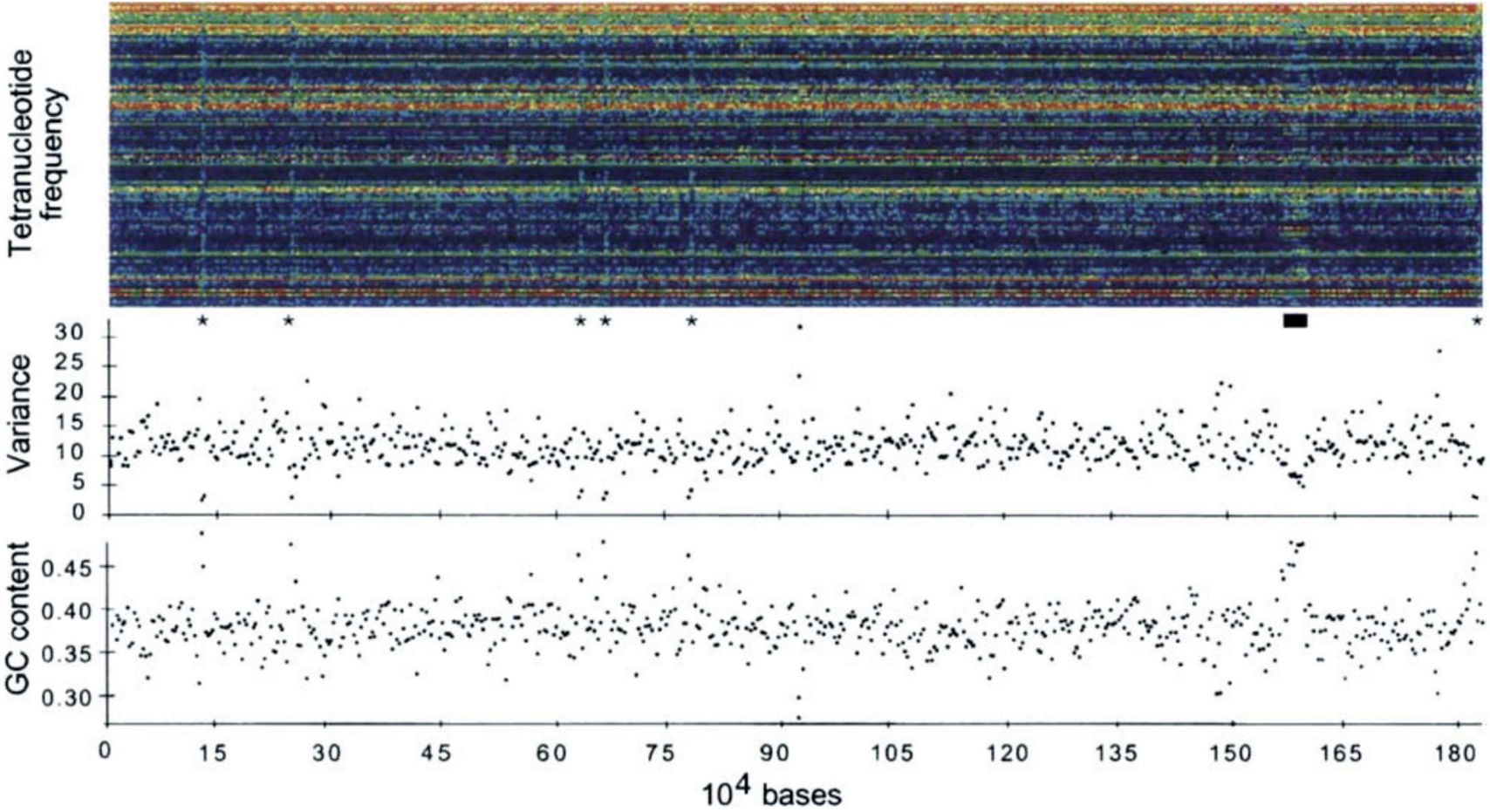
MetaBAT

4. Calculate TNF for each contig
5. Calculate Abundance per library for each contig
6. Calculate the pairwise distance matrix using pre-trained probabilistic models
7. Forming genome bins iteratively

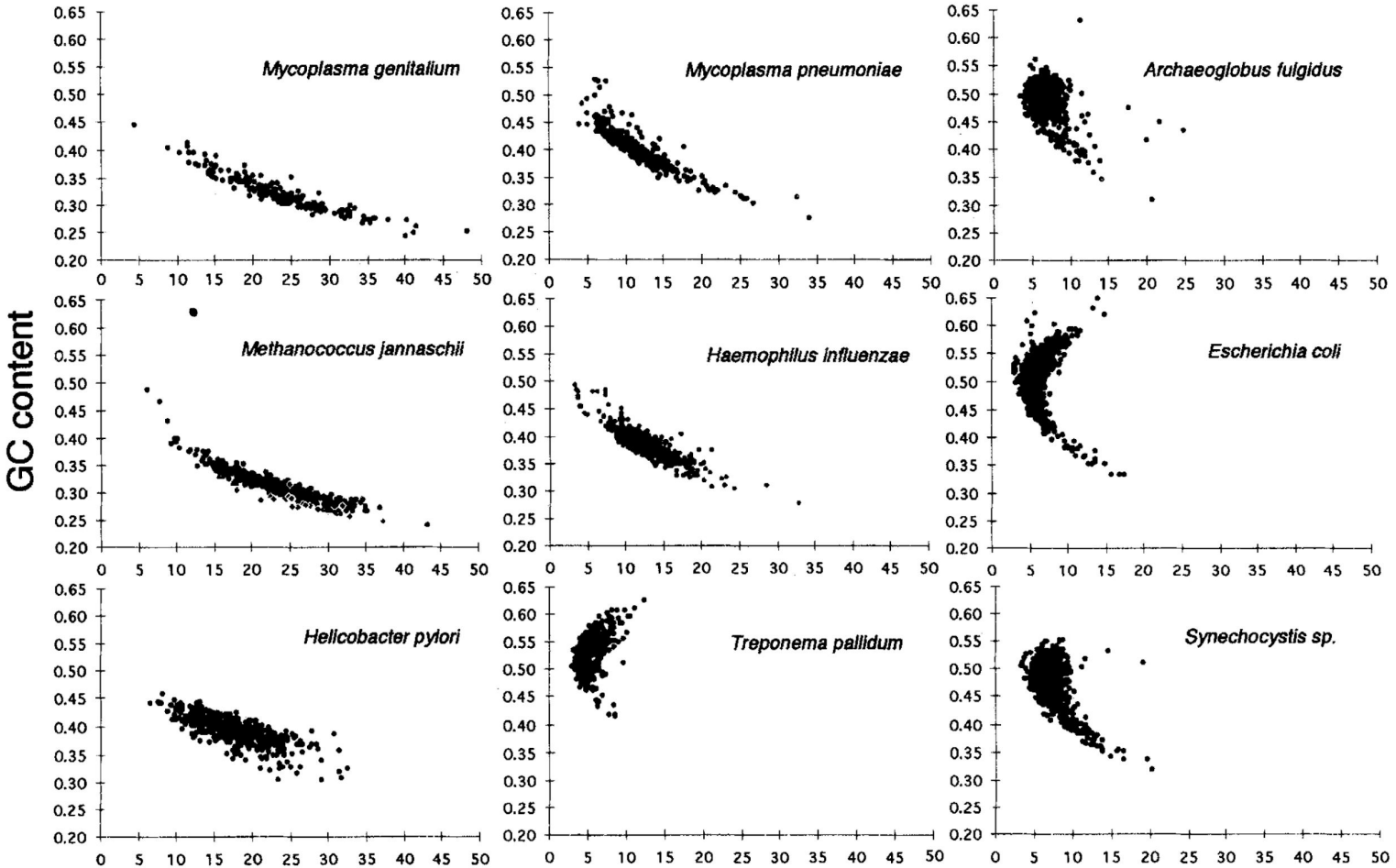
You can see why shot-gun metagenomics is often referred to as a "pangenome approach".

Kang et al. 2015. "MetaBAT, an Efficient Tool for Accurately Reconstructing Single Genomes from Complex Microbial Communities." *PeerJ* 3 (8): e1165.

Binning

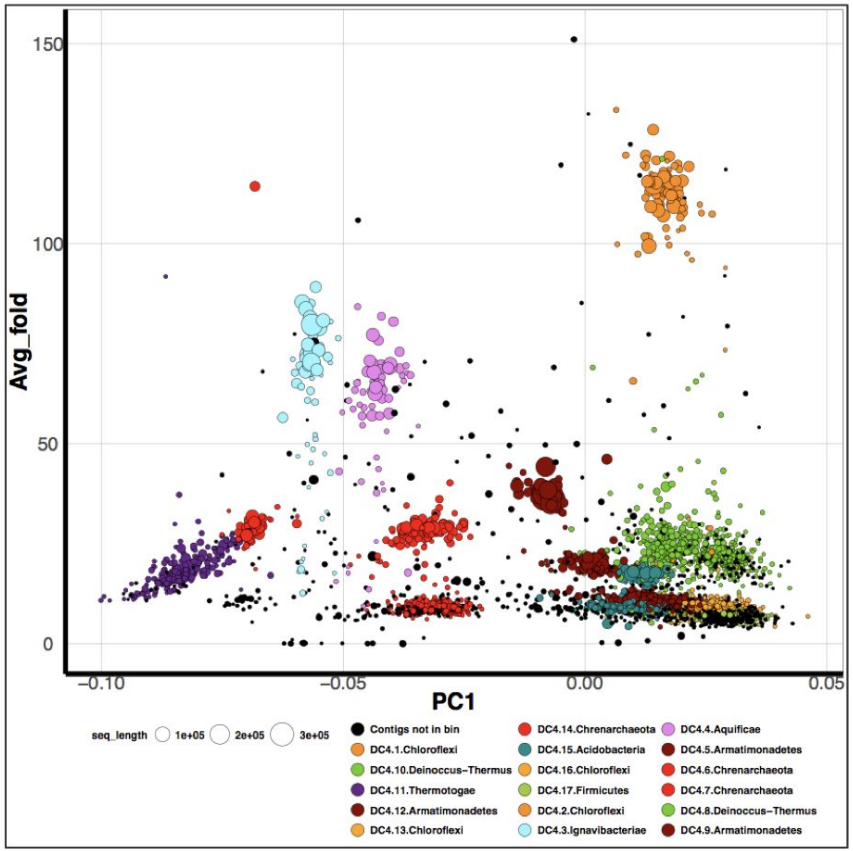


Noble, P. A., R. W. Citek, and O. A. Ogunseitan. 1998. “Tetranucleotide Frequencies in Microbial Genomes.” *Electrophoresis* 19 (4): 528–35.



Noble, P. A., R. W. Citek, and O. A. Ogunseitan. 1998. “Tetranucleotide Frequencies in Microbial Genomes.” *Electrophoresis* 19 (4): 528–35.

Blob plots



Anvi'o example

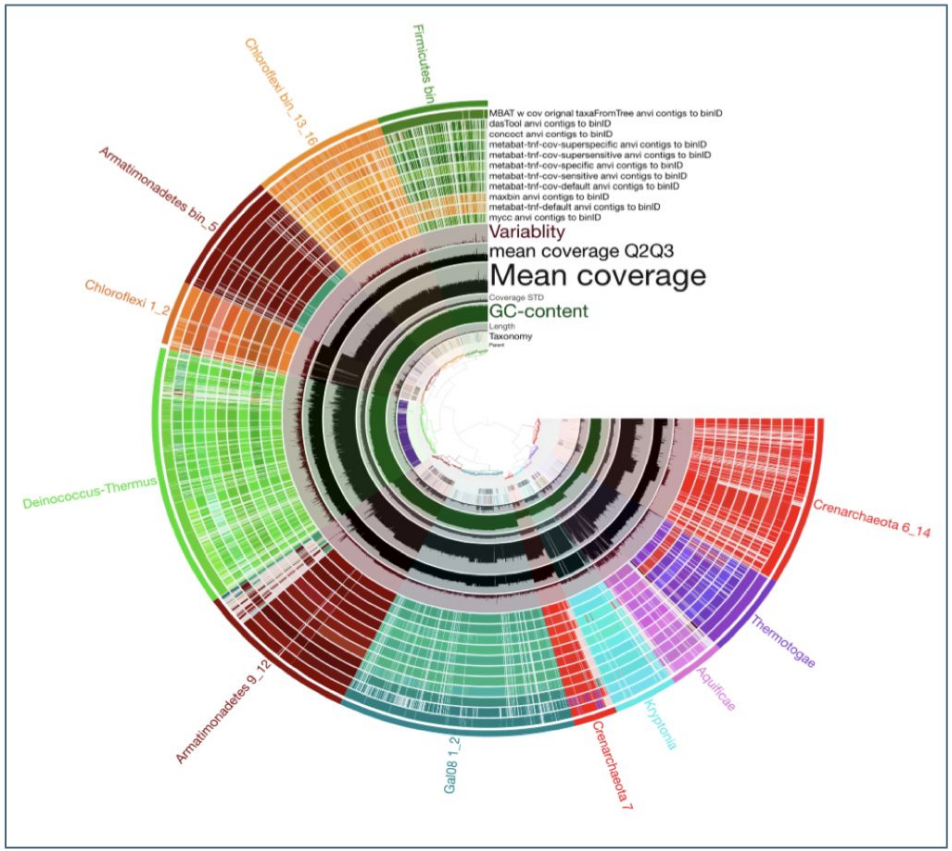
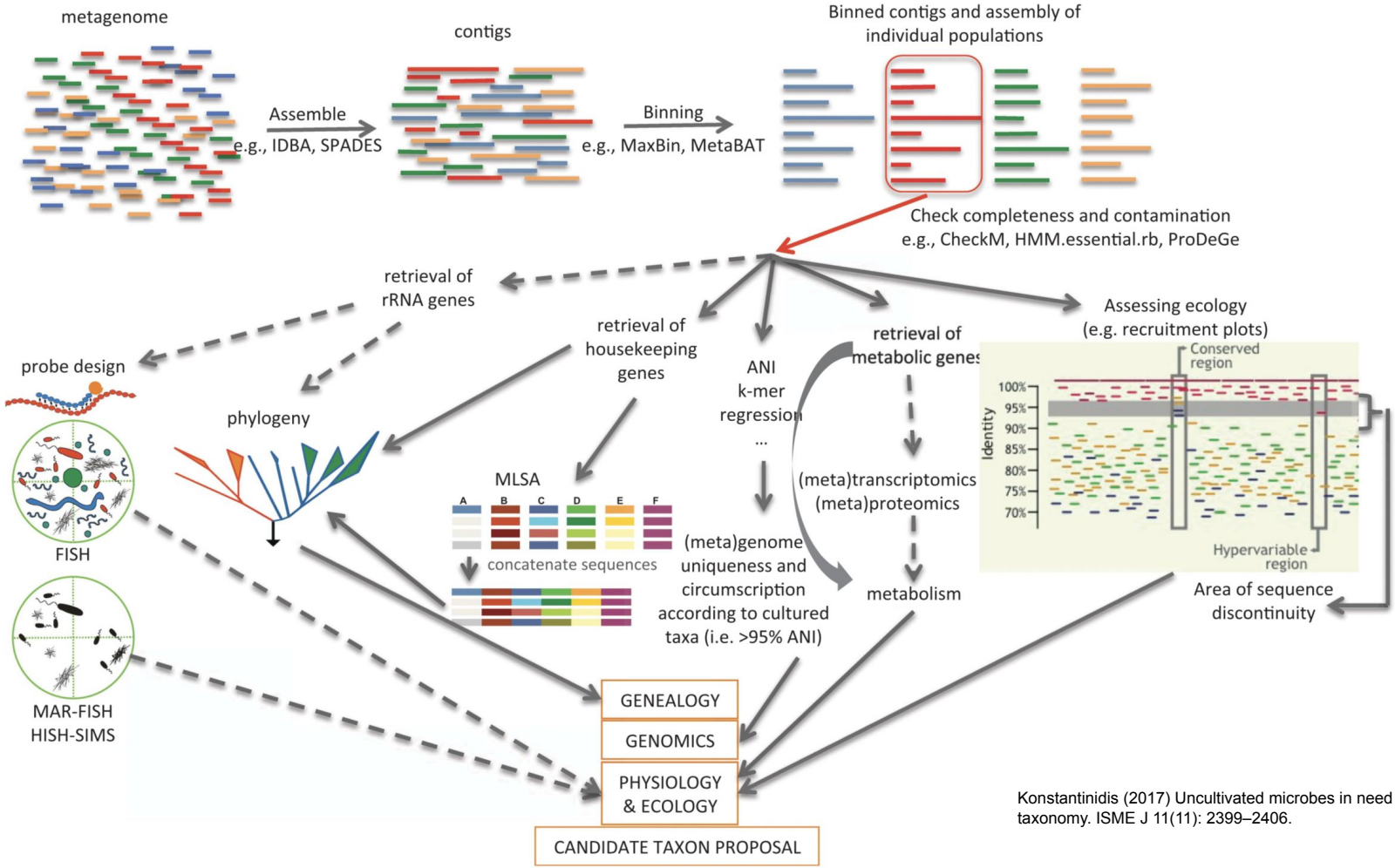


Image Credit: Bob Bowers
Tool: Anvi'o (Eren et al. (2015) PeerJ 3(358): e1319–29. PeerJ doi:10.7717/peerj.1319.)



Konstantinidis (2017) Uncultivated microbes in need of their own taxonomy. ISME J 11(11): 2399–2406.

Table 4 Strengths and weaknesses of assembly-based and read-based analyses for primary analysis of metagenomics data

	Assembly-based analysis	Read-based analysis ('mapping')
Comprehensiveness	Can construct multiple whole genomes, but only for organisms with enough coverage to be assembled and binned.	Can provide an aggregate picture of community function or structure, but is based only on the fraction of reads that map effectively to reference databases.
Community complexity	In complex communities, only a fraction of the genomes can be resolved by assembly.	Can deal with communities of arbitrary complexity given sufficient sequencing depth and satisfactory reference database coverage
Novelty	Can resolve genomes of entirely novel organisms with no sequenced relatives.	Cannot resolve organisms for which genomes of close relatives are unknown.
Computational burden	Requires computationally costly assembly, mapping and binning.	Can be performed efficiently, enabling large meta-analyses.
Genome-resolved metabolism	Can link metabolism to phylogeny through completely assembled genomes, even for novel diversity.	Can typically resolve only the aggregate metabolism of the community, and links with phylogeny are only possible in the context of known reference genomes.
Expert manual supervision	Manual curation required for accurate binning and scaffolding and for misassembly detection.	Usually does not require manual curation, but selection of reference genomes to use could involve human supervision.
Integration with microbial genomics	Assemblies can be fed into microbial genomic pipelines designed for analysis of genomes from pure cultured isolates.	Obtained profiles cannot be directly put into the context of genomes derived from pure cultured isolates.

Quince et al. (2017) Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. 35(9): 833–844