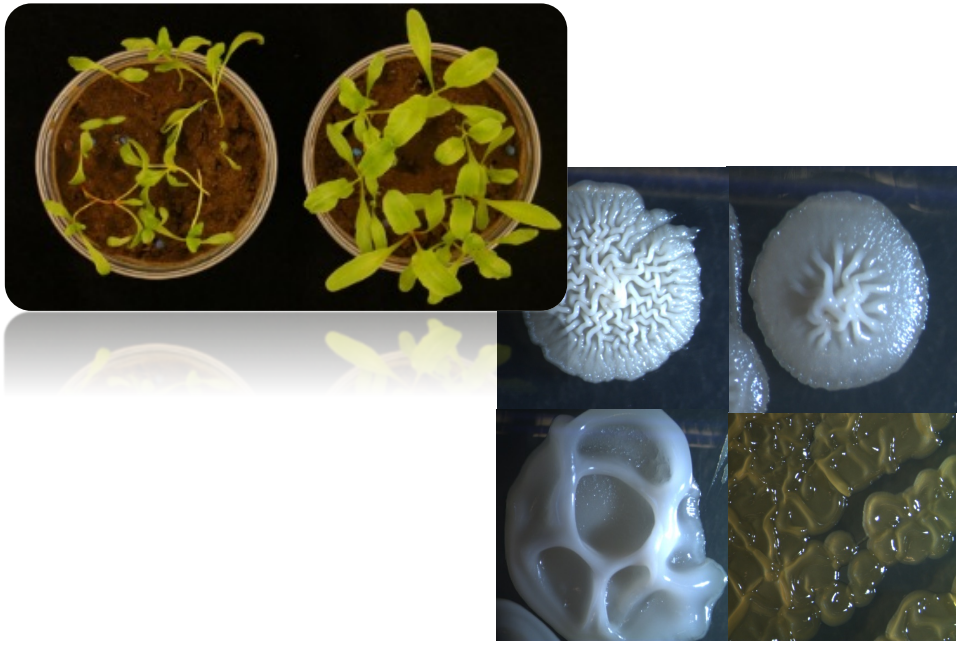
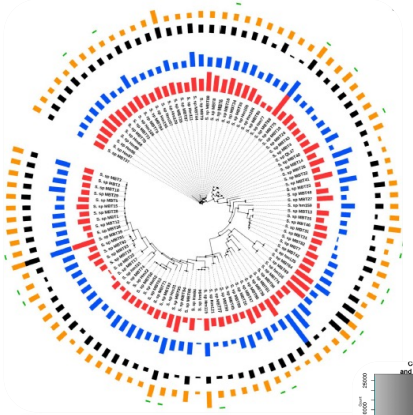
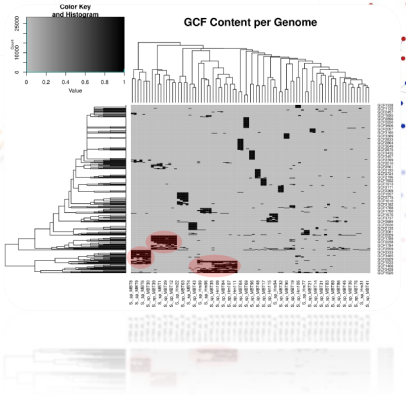
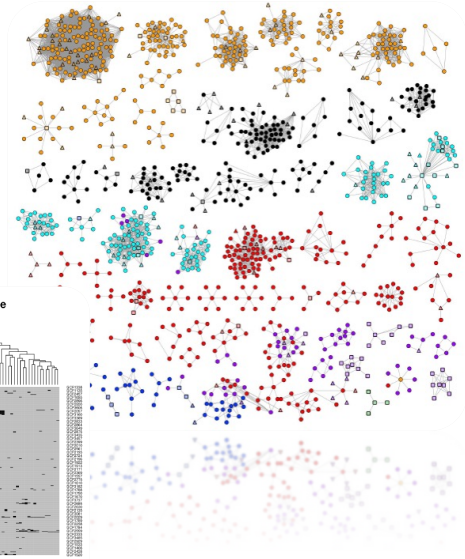
Meta(genomics)









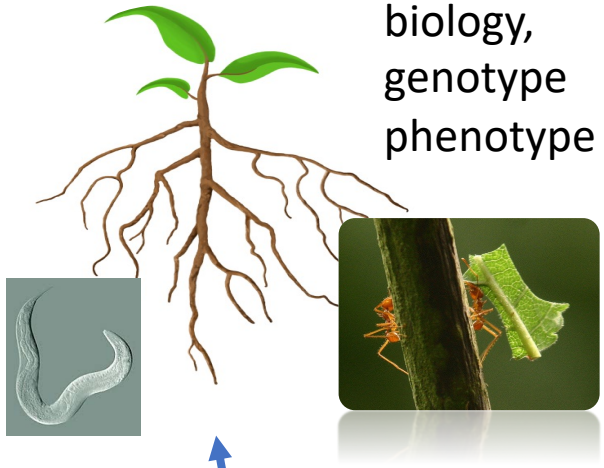
Victor J **Carrión**
v.j.carrion.bravo@biology.leidenuniv.nl
@VCarryOn1

THREE MAIN QUESTIONS FOR MICROBIOME RESEARCH

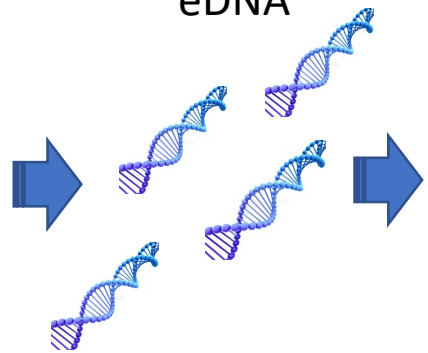
What are they doing?
How are they doing it?

Metagenomics, Metatranscriptomics …. all omics

biology,
genotype
phenotype

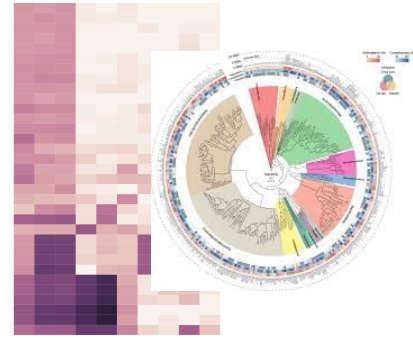


eDNA

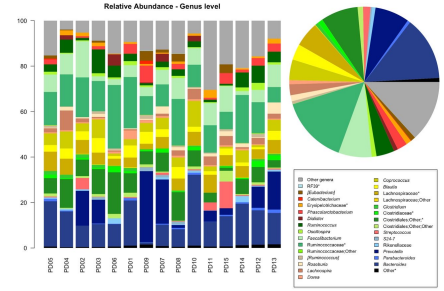


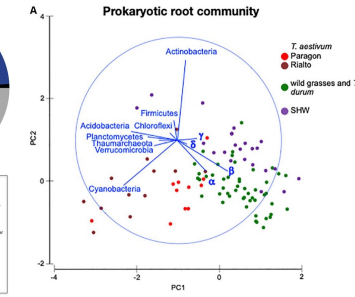
NGS sequencing

functions, meta'omics



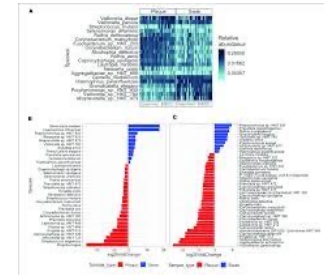
taxonomy, phylogeny

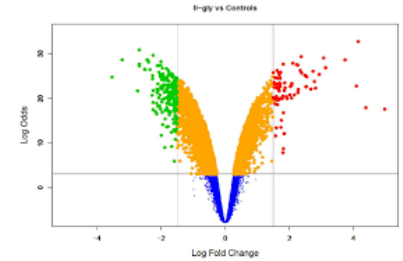




culturing
phenotyping

differential abundance (function & taxonomy)



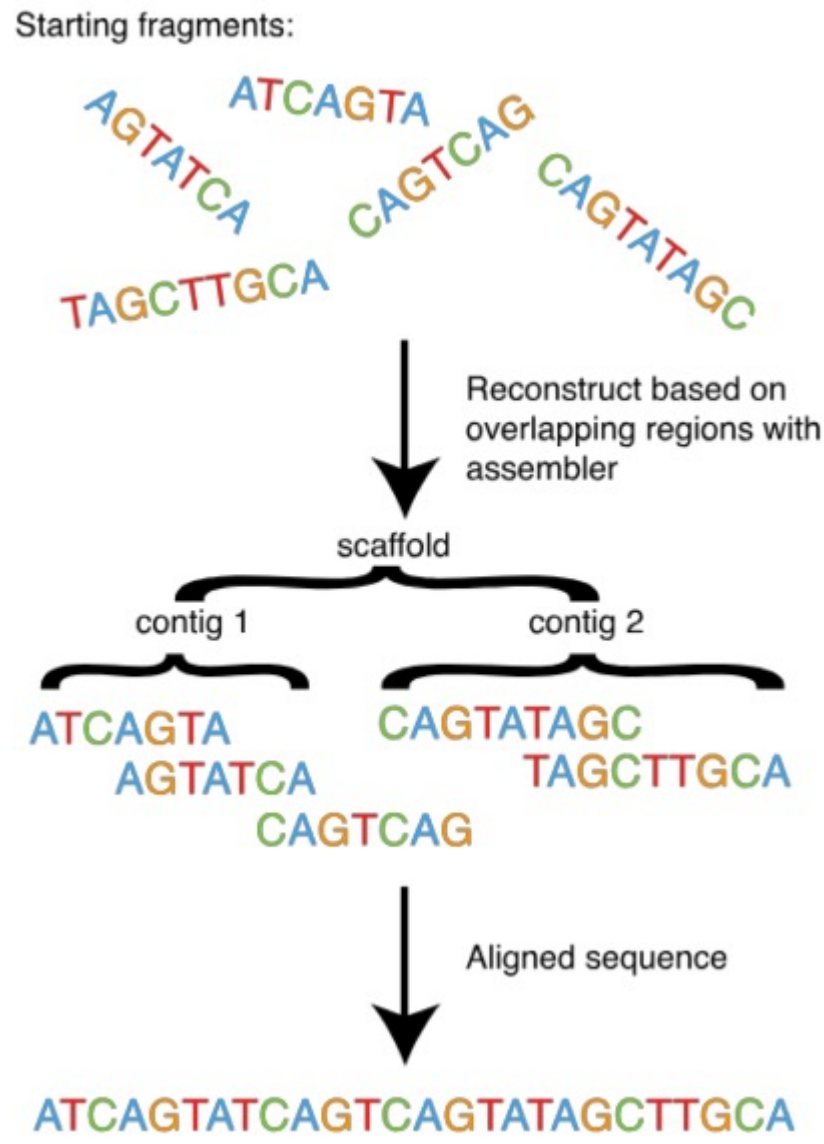


genome assembly
comparative genomics

cross-link

SHOTGUN SEQUENCING (METAGENOMICS)

Randomly break up the DNA, sequence all of the fragments to study potential gene function and assemble genomes/partial genomes in addition to taxonomy



DEFINITIONS

Metagenomics is the study of microbial communities in their original living places. Metagenome sequencing refers to sequencing the entire genomes of all microbes present in a sample in order to explore taxonomic, functional, and evolutionary aspects

Metatranscriptomics informs us of the genes that are expressed by the community as a whole. With the use of functional annotations of expressed genes, it is possible to infer the functional profile of a community under specific conditions, which are usually dependent on the status of the host.

Metabolomics is the comprehensive analysis by which all metabolites of a sample (small molecules released by the organism into the immediate environment) are identified and quantified.

Metaproteomics is the large-scale identification and quantification of proteins from microbial communities and thus provides direct insight into the phenotypes of microorganisms on the molecular level.

Unculturable
Ecology
Natural state
Health

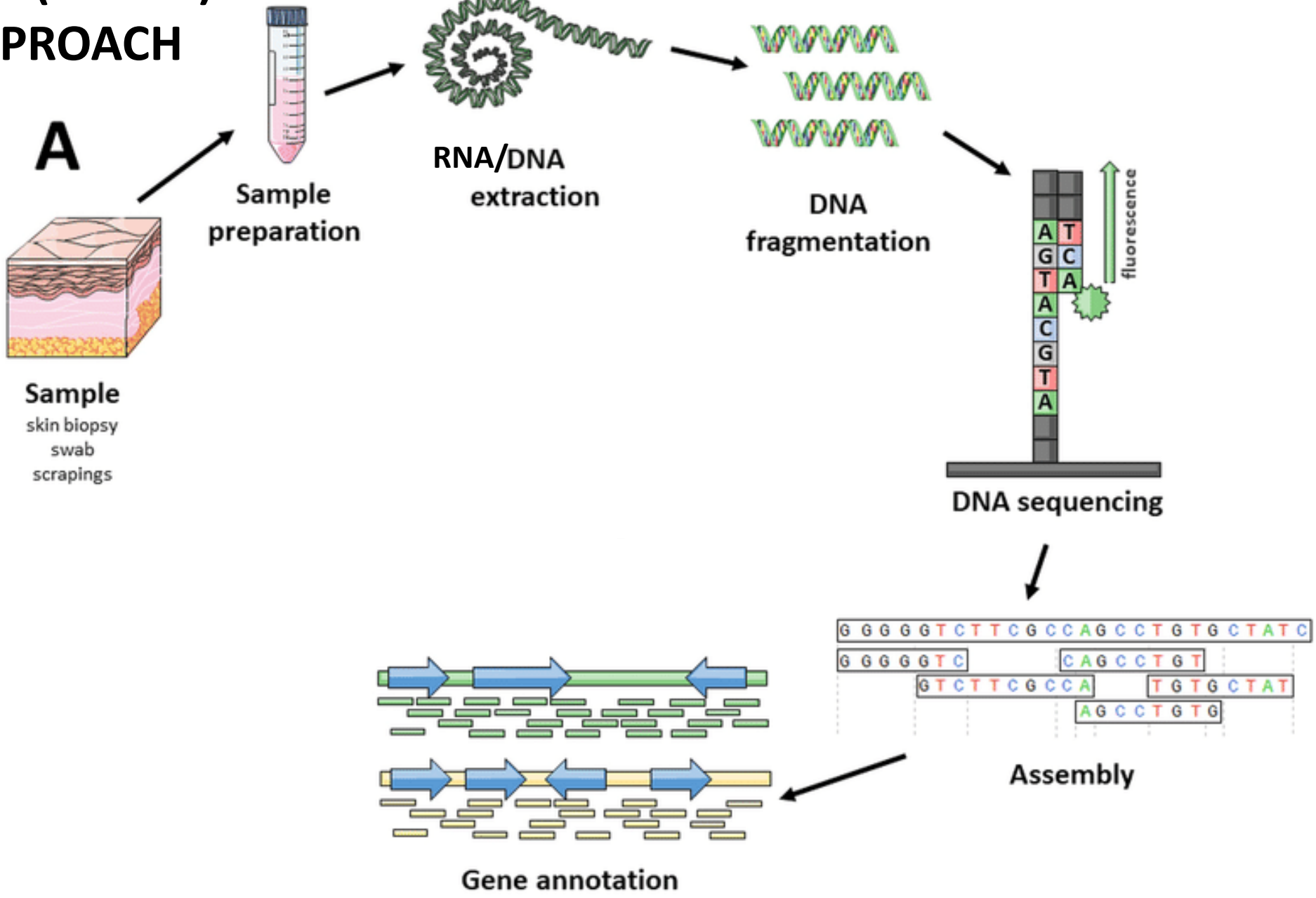
PITFALLS AND TROUBLESHOOTING

What can affect results of your communities?

Sampling method
Experimental design
Sequencing strategy
Workflows
Analysis

Tip: pilot experiments, consult bioinformatics experts

META(OMICS) APPROACH



Godlewska et al., 2020

WORKFLOWS …

metagenome-atlas/ **genecatalog_atlas**

1 Quality Control
- PCR duplicates removal
- Quality trimming
- Host removal
- Common contaminant removal

➢ **QC reads**

2 Assembly
- Error correction
- Paired-end merging
- Assembly (metaSpades/megahit)
- Post-filtering

➢ **High-quality Scaffolds**

3 Genomic Binning
- Binning (metabat, maxbin2)
- Quality Assessment (checkM)
- Bin refining (DAS Tool)
- Dereplication (dRep)
- Quantification
- Robust taxonomic classification (CAT)

➢ **Genomes**
➢ **Abundances**

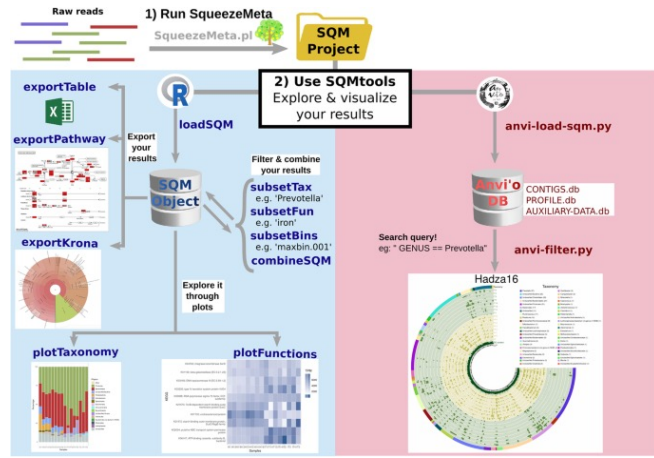
Clostridiaceae nov.
A. muciniphila

4 Annotation
- Gene prediction (prodigal)
- Cluster redundant genes (linclust/ cd-hit)
- Annotation (eggNOG)

➢ **Comparable gene catalog**

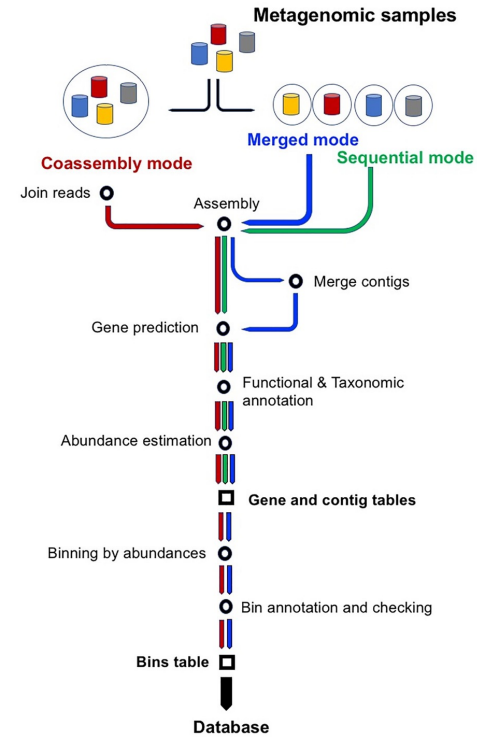
https://github.com/metagenome-atlas/atlas

anvi'o



https://merenlab.org/software/anvio/

SqueezeMeta

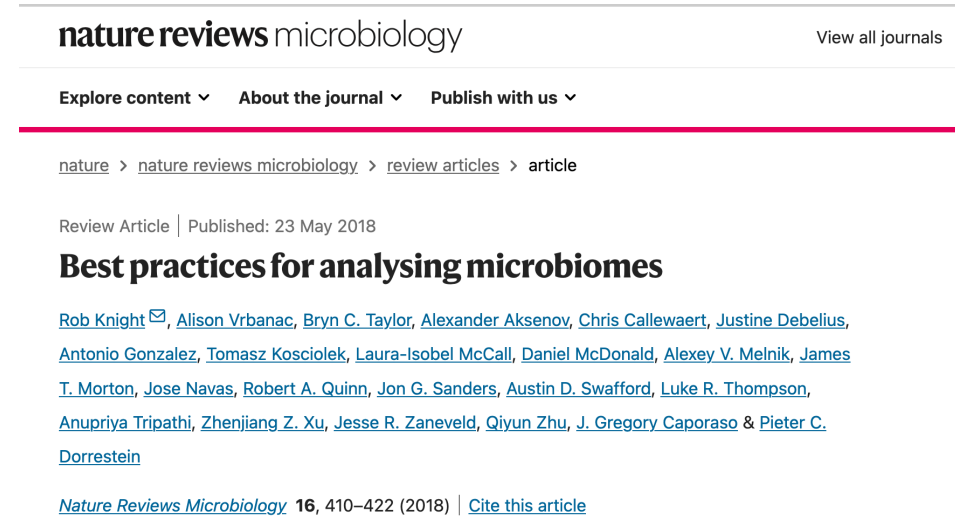


https://github.com/jtamames/SqueezeMeta

Many more!!!

Types of normalization methods

1. Rarefaction
2. Total Sum Scaling (TSS)
3. Relative Abundance
4. **Counts per Million (CPM)**
5. **RPKM/FPKM/TPM**
6. Cumulative Sum Scaling (CSS)
7. DESeq2 / EdgeR Normalization
8. Variance Stabilizing Transformation (VST)
9. Log and CLR Transformations



nature reviews microbiology

View all journals

Explore content · About the journal · Publish with us

nature › nature reviews microbiology › review articles › article

Review Article | Published: 23 May 2018

Best practices for analysing microbiomes

Rob Knight, Alison Vrbanac, Bryn C. Taylor, Alexander Aksenov, Chris Callewaert, Justine Debelius, Antonio Gonzalez, Tomasz Kosciolek, Laura-Isobel McCall, Daniel McDonald, Alexey V. Melnik, James T. Morton, Jose Navas, Robert A. Quinn, Jon G. Sanders, Austin D. Swafford, Luke R. Thompson, Anupriya Tripathi, Zhenjiang Z. Xu, Jesse R. Zaneveld, Qiyun Zhu, J. Gregory Caporaso & Pieter C. Dorrestein

Nature Reviews Microbiology **16**, 410–422 (2018) | Cite this article

CPM, RPKM, FPKM & TPM

- CPM: Counts per million, adjusts for sequencing depth.
- RPKM/FPKM: Normalizes gene length and sequencing depth.
- TPM: Similar to RPKM but corrects for transcript length biases.

TAXONOMIC AND FUNCTIONAL ANNOTATION

General

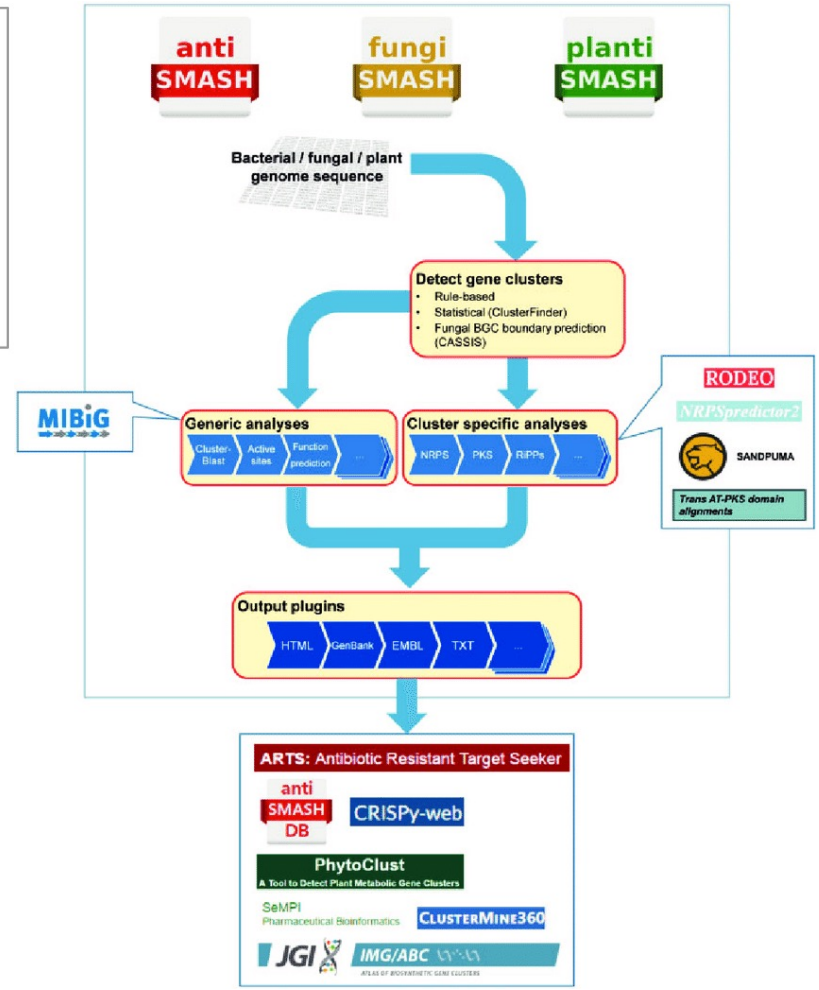
Cantalapiedra et al., 2021

Specialized

CAZypedia

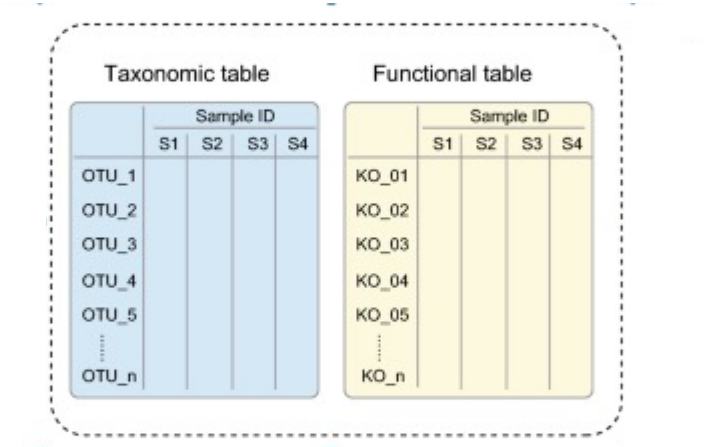
carbohydrate-active
ENZYMES

http://www.cazy.org/



Medema group

α AND β DIVERSITY IN META(OMICS)



Liu et al., 2020

CAN WE RECONSTRUCT GENOMES FROM METAGENOMES?

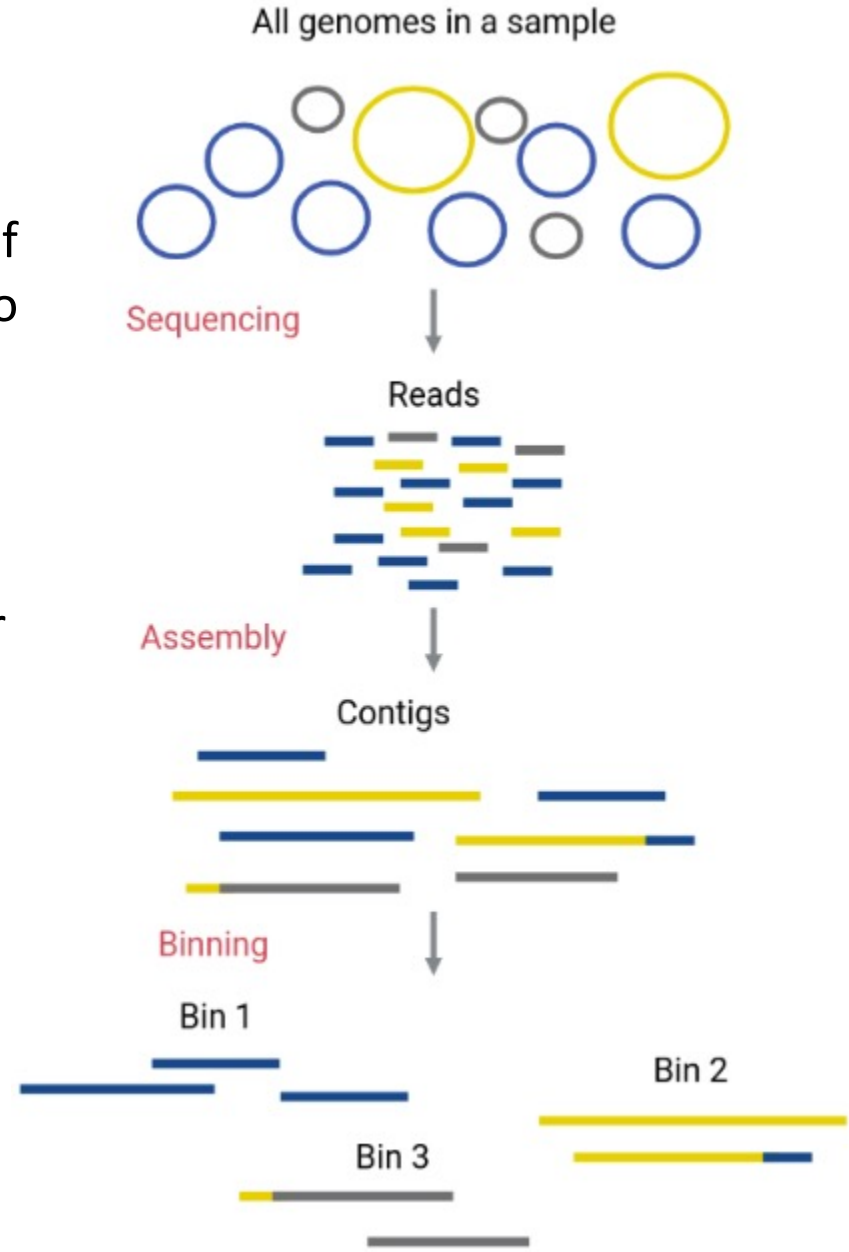
Binning. in metagenomics, binning is the process of grouping reads or contigs and assigning them to individual genome.

Supervised approach

- relies on known reference genomes
- uses homology or sequence composition similarity for binning

Unsupervised approach

- does not need a reference genome
- relies on sequence composition and/or species abundance for binning

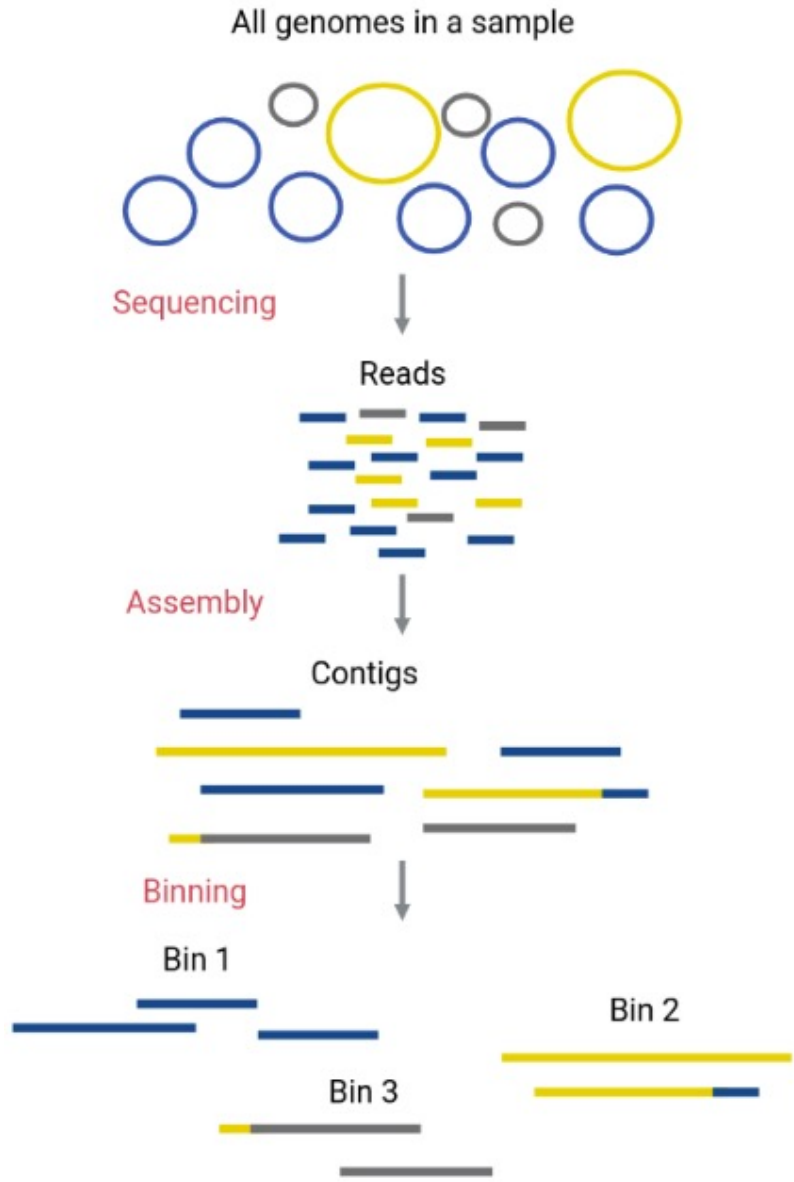


https://academic.oup.com/bib/article/13/6/669/193900?login=false

CAN WE RECONSTRUCT GENOMES FROM METAGENOMES?

Binning. in metagenomics, binning is the process of grouping reads or contigs and assigning them to individual genome.

- **Mapping statistics**
  - ✓ Read coverage
- **Composition**
  - ✓ Kmer frequency (tetranucleotides)
  - ✓ %GC content
  - ✓ Codon usage
- **Other**
  - ✓ Taxonomic assignment



What is K-mer frequency?

- A K-mer is a sequence of length k in a genome.
- Example: 4-mers (tetranucleotides) → AAAA, AAAC, AAGT, etc.
- The frequency of K-mers provides a genomic signature useful for classification.
- Taxonomic classification, genome binning, and HGT detection

https://www.sciencedirect.com/science/article/pii/S2001037024001703

Genome Assembly	In *de novo* genome assembly, sequencing reads undergo fragmentation into k-mers, and their overlaps are employed to assemble longer contiguous sequences. This process commonly utilizes k-mers to build De Bruijn graphs or implements overlap-layout-consensus methods.	Allpaths-LG[239], Bifrost [240], Canu[241], Cortex [242], ELBA[243], KAT[223], MEGAHIT[244], Merqury [46], QUAST-LG[245], SKESA [246], SPAdes[75], TandemQUAST[247], TandemMapper[248]
Sequence Comparison	Alignment-free methods are increasingly used for DNA and protein sequence comparison since they are much faster than traditional alignment-based approaches. Most alignment-free algorithms are based on the word or k-mer composition of the sequences under study. [249]	BBMap[250], Bowtie2[251, 252], BWA[253–255], iMOKA [256], MiniMap2[71]
Protein Sequence Searching and Alignment	Sequence match is determined by aligning translated DNA sequences to a reference protein database.	BLAT[65], BLAST[68], BLASTX[64,282], DIAMOND [283], MMSeqs2[284], PAUDA[285], RAPSearch2 [286], USEARCH & UBLAST [287]
Taxonomic Classification	In sequence composition-based methods, the frequency and distribution of k-mers in metagenomic data are analyzed to assess genome similarity across various taxonomic ranks. [257,258]	ARK[259], BinDash[122], Bracken[260], CDKAM[261], CLARK[262], Dashing[124], fmh-funprofiler[128], Genometa[263], Kaiju[138], KMCP[264], KmerFinder [265], Kraken2[136], KrakenUniq[19], LMAT[266], Mash[72], Mash Screen[34], Matchtigs[267], MetaCache [268], MetaPalette[269], MetaProFi[50], NIQKI[126], SEK[270], StrainSeeker[271], SuperSampler[127], TACOA

https://www.sciencedirect.com/science/article/pii/S2001037024001703

Codon usage in bacteria

- Codon usage refers to the preference of certain codons over others for the same amino acid.

- It varies among bacterial species due to factors like tRNA availability, GC content, gene expression levels, and evolutionary adaptations.

- Example: Leucine has six codons (UUA, UUG, CUU, CUC, CUA, CUG).

- Different bacteria prefer specific codons.

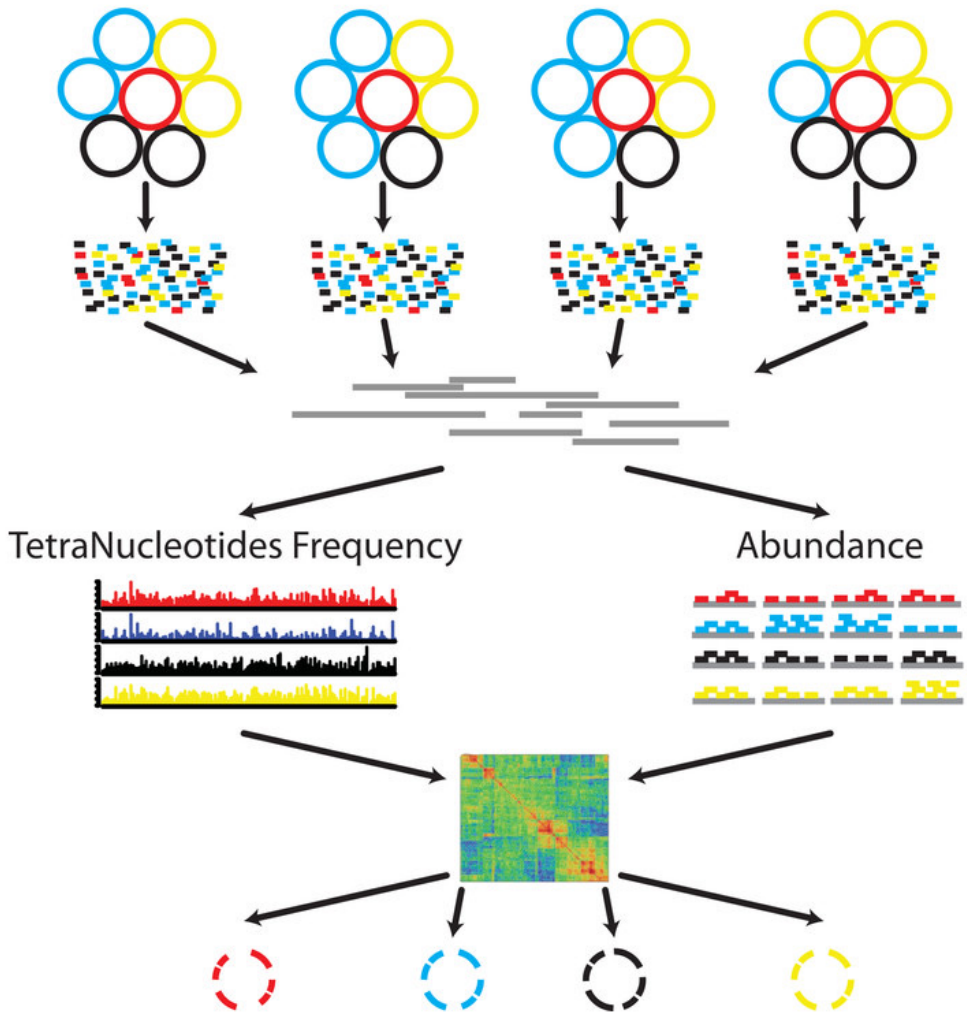
Examples of codon usage in bacteria

- *E. coli*: moderate GC content codons (e.g., CGU for arginine).

- *Bacillus subtilis*: Low GC content preference (e.g., AAA for lysine).

- *Pseudomonas aeruginosa*: High GC content preference (e.g., CUG for leucine).

BINNING METHODS: EXAMPLE METABAT



Preprocessing

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

MetaBAT

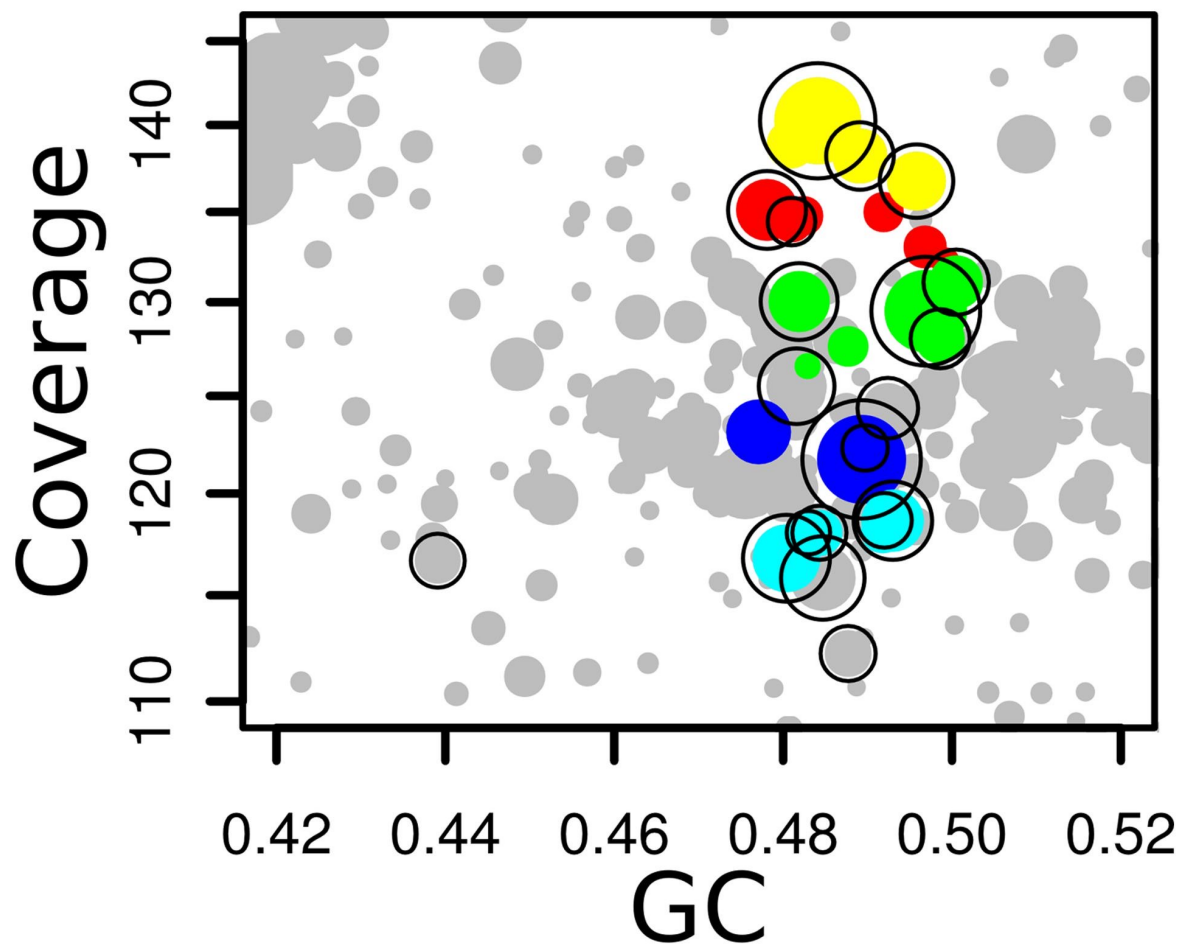
4. Calculate TNF for each contig
5. Calculate Abundance per library for each contig
6. Calculate the pairwise distance matrix using pre-trained probabilistic models
7. Forming genome bins iteratively

Kang DD *et al.*, *PeerJ* (2015)

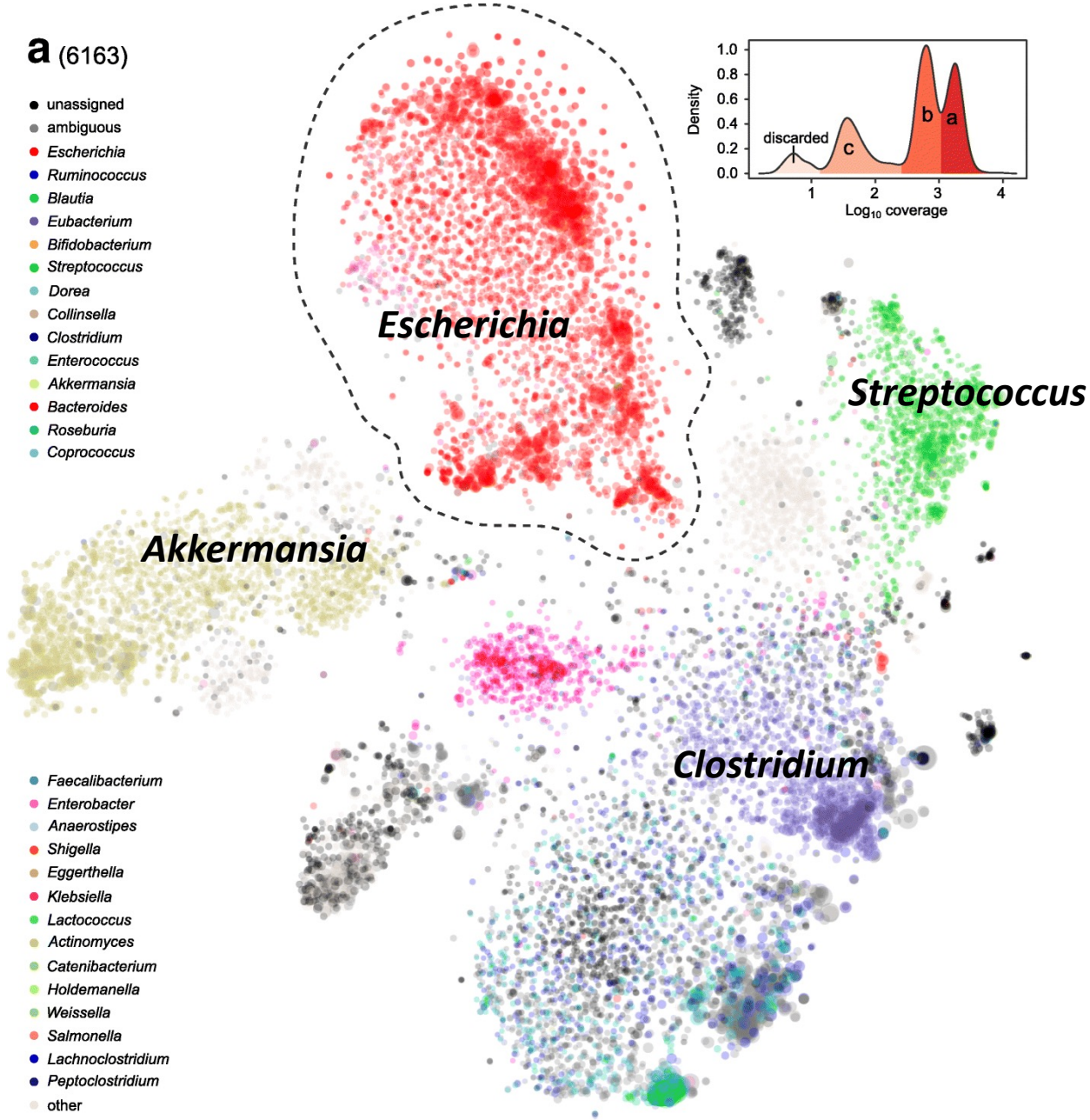
Table 1 Summary of twelve original genome binner and three refinning genome binner

Genome binner	Parameters	Model	Version to validate	Publication	Last update	Resources
MaxBin	k-mer frequencies, coverage, single-copy genes	Expectation-maximization, bin number estimated from single-copy marker gene analysis	2.2.6	2014	2019	https://sourceforge.net/projects/maxbin
MetaBat	4-mer frequencies, coverage	Modified K-medoids algorithm	1&2.13	2015	2020	https://bitbucket.org/berkeleylab/metabat/src/master
Groopm	coverage, contig's length, tetranucleotide frequency	Two way clustering, Hough partitioning, self-organizing map	2	2014	2017	https://github.com/timbalam/GroopM
CONCOCT	k-mer frequencies, coverage	Gaussian mixture models, bin number determined by variable Bayesian	1.0.0	2014	2019	https://github.com/BinPro/CONCOCT
MyCC	k-mer frequencies, coverage (optional), universal single-copy genes	Affinity propagation	1	2016	2017	https://sourceforge.net/projects/sb2nhri
MetaWatt	tetranucleotide frequency, coverage	Firstly clustering by empirical relationship of the average standard deviation at tetranucleotide frequency mean, then employing interpolated Markov models	3.5.3	2012	2016	https://sourceforge.net/projects/metawatt
BMC3C	frequency variation of oligonucleotides, coverage, codon usage	Ensemble k-means, construct a weigh graph and partition it by Normalized cuts [49, 50]	\	2018	2018	http://mlda.swu.edu.cn/codes.php?name=BMC3C
Binsanity	coverage, tetranucleotide frequency, percent GC content	Affinity propagation	0.2.8	2017	2020	https://github.com/edgraham/BinSanity
Autometa	sequence homology, single-copy genes, 5-mer frequency, coverage, single-copy genes	Lowest common ancestor analysis, DBSCAN algorithm, supervised decision tree classifier recruite unclustered contigs	\	2019	2020	https://bitbucket.org/jason_c_kwan/autometa/src/master
COCACOLA	k-mer frequency, coverage, co-alignment, paired-end read linkage	K-means based on L1 distance, non-negative matrix factorization with sparse regularization, hierarchical clustering	\	2017	2017	https://github.com/younglululu/COCACOLA
SolidBin-naive	single-copy mark genes, tetranucleotide frequencies, coverage, pairwise constraints	Semi-supervised spectral Normalized cut	1.1	2019	2020	https://github.com/solforest/SolidBin
Vamb	tetranucleotide frequencies, coverage	Variational autoencoders, iterative medoid clustering algorithm	2.0.1	2018	2020	https://github.com/RasmussenLab/vamb
DAS Tool	original binner output bin sets	Refine bins according shared contigs between two original binner results	1.1.1	2018	2019	https://github.com/cmks/DAS_Tool
MetaWrap	original binner output bin sets	Separating every pair of contigs in different bins, selecting the best bin sets according completion and contamination	1.2.2	2018	2019	https://github.com/bxlab/metaWRAP
Binning_refiner	original binner output bin sets, single-copy genes	Scoring bins based on single-copy genes and picking up high-score bins iteratively	1.4.0	2017	2019	https://github.com/songweizhi/Binning_refiner

Yue et al., 2020

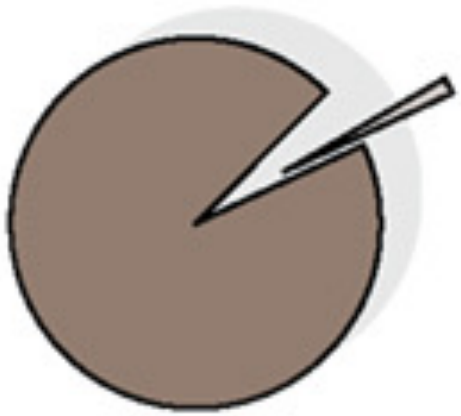


Kang et al., 2016



Qiyun Zhu et al., 2018

WHAT CAN GO WRONG?



Pseudomonas
Enterobacter

Pseudomonas fluorescens
Pseudomonas putida
Pseudomonas aeruginosa
Pseudomonas savastanoi
Pseudomonas syringae

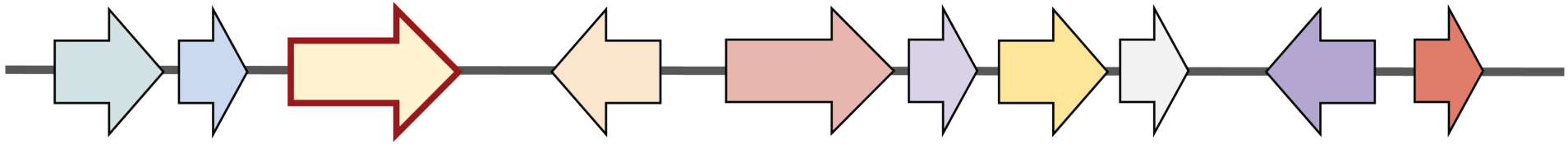
Modified from Strous et al., 2012

Bin Id	Marker lineage	# genomes	# markers	# marker sets	0	1	2	3	4	5+	Completeness	Contamination	Strain heterogeneity
res.005	N/A (0)	-1	43	1	0	43	0	0	0	0	100.00	0.00	80.00
res.004	N/A (0)	-1	43	1	0	43	0	0	0	0	100.00	0.00	50.00
res.002	N/A (0)	-1	43	1	0	43	0	0	0	0	100.00	0.00	0.00
res.001	N/A (0)	-1	43	1	0	43	0	0	0	0	100.00	0.00	0.00
res.003	N/A (0)	-1	43	1	1	42	0	0	0	0	97.67	0.00	0.00
res.014	N/A (0)	-1	43	1	7	19	9	8	0	0	83.72	58.14	4.69
res.012	N/A (0)	-1	43	1	8	19	16	0	0	0	81.40	37.21	4.17
res.007	N/A (0)	-1	43	1	8	35	0	0	0	0	81.40	0.00	33.33
res.009	N/A (0)	-1	43	1	10	21	12	0	0	0	76.74	27.91	0.00
res.008	N/A (0)	-1	43	1	11	32	0	0	0	0	74.42	0.00	0.00
res.011	N/A (0)	-1	43	1	19	24	0	0	0	0	55.81	0.00	25.00
res.013	N/A (0)	-1	43	1	23	18	2	0	0	0	46.51	4.65	0.00
res.010	N/A (0)	-1	43	1	37	6	0	0	0	0	13.95	0.00	0.00
res.006	N/A (0)	-1	43	1	43	0	0	0	0	0	0.00	0.00	0.00

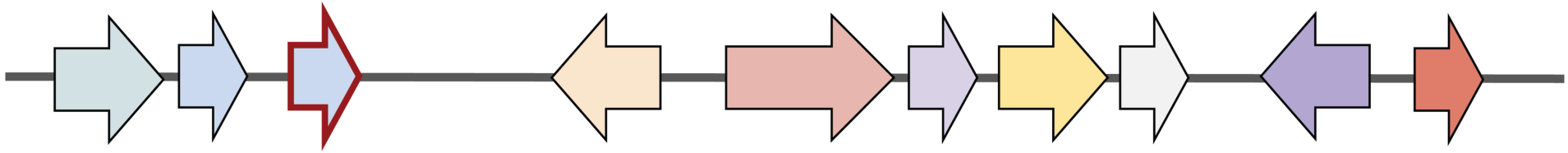
QUALITY ASSESSMENT: CheckM

Uses a set of lineage-specific single-copy marker genes (SCMG) - genes that are present in nearly every genome within a lineage and are single copy.

Reference SCMG set



New genome assembly to evaluate



Completeness: 90% (9 out of 10 genes are present)
Contamination: 10% (1 gene occurs twice)

Parks DH *et al.*, *Genome Res.* (2015)

QUALITY ASSESSMENT: CheckM

Strain heterogeneity: indicates the source of contamination (other strains of the same species vs more distant taxa)

Completeness: 85%
Contamination: 15%
Strain heterogeneity: 100%

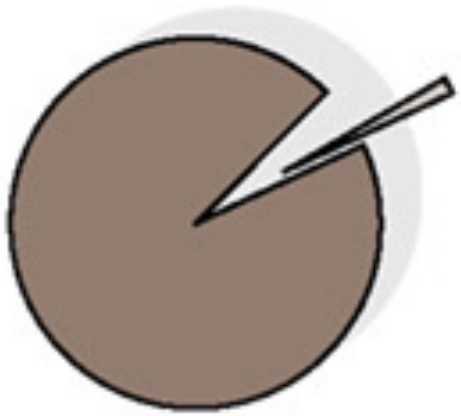
all contamination is likely to be from other strains of the same species

Completeness: 85%
Contamination: 15%
Strain heterogeneity: 0%

all contamination is likely to be from different species

Tools to remove contamination:
GUNC (https://grp-bork.embl-community.io/gunc/)
MAGpurify (https://github.com/snayfach/MAGpurify)

WHAT CAN GO WRONG?



Pseudomonas
Enterobacter

Pseudomonas fluorescens
Pseudomonas putida
Pseudomonas aeruginosa
Pseudomonas savastanoi
Pseudomonas syringae

Modified from Strous et al., 2012

Bin Id	Marker lineage	# genomes	# markers	# marker sets	0	1	2	3	4	5+	Completeness	Contamination	Strain heterogeneity
res.005	N/A (0)	-1	43	1	0	43	0	0	0	0	100.00	0.00	80.00
res.004	N/A (0)	-1	43	1	0	43	0	0	0	0	100.00	0.00	50.00
res.002	N/A (0)	-1	43	1	0	43	0	0	0	0	100.00	0.00	0.00
res.001	N/A (0)	-1	43	1	0	43	0	0	0	0	100.00	0.00	0.00
res.003	N/A (0)	-1	43	1	1	42	0	0	0	0	97.67	0.00	0.00
res.014	N/A (0)	-1	43	1	7	19	9	8	0	0	83.72	58.14	4.69
res.012	N/A (0)	-1	43	1	8	19	16	0	0	0	81.40	37.21	4.17
res.007	N/A (0)	-1	43	1	8	35	0	0	0	0	81.40	0.00	33.33
res.009	N/A (0)	-1	43	1	10	21	12	0	0	0	76.74	27.91	0.00
res.008	N/A (0)	-1	43	1	11	32	0	0	0	0	74.42	0.00	0.00
res.011	N/A (0)	-1	43	1	19	24	0	0	0	0	55.81	0.00	25.00
res.013	N/A (0)	-1	43	1	23	18	2	0	0	0	46.51	4.65	0.00
res.010	N/A (0)	-1	43	1	37	6	0	0	0	0	13.95	0.00	0.00
res.006	N/A (0)	-1	43	1	43	0	0	0	0	0	0.00	0.00	0.00

DAS TOOL FOR GENOME RESOLVED METAGENOMICS

https://merenlab.org/2020/01/02/visualizing-metagenomic-bins/#example-1

DAS Tool is an automated method that integrates the results of a flexible number of binning algorithms to calculate an optimized, non-redundant set of bins from a single assembly.

Sieber et al., 2018

QUALITY ASSESSMENT STANDARDS

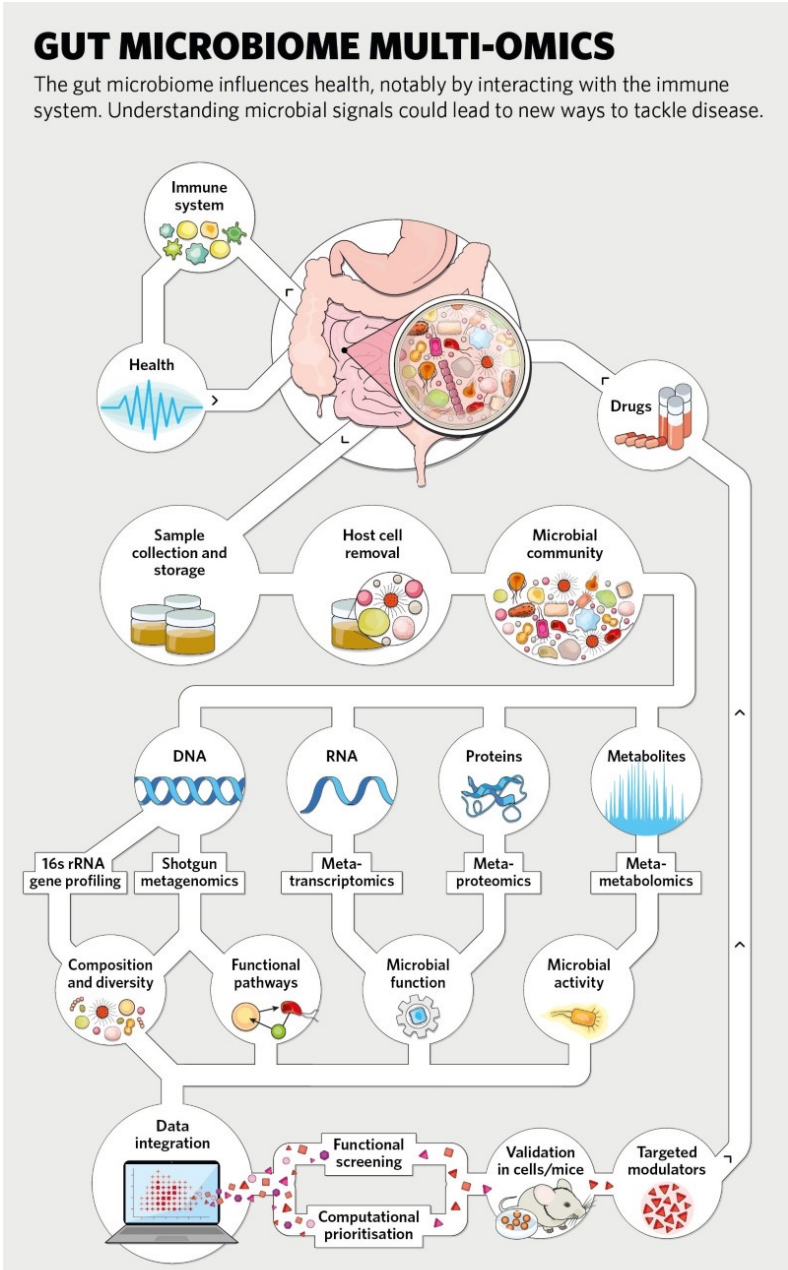
Minimum Information about a Metagenome-Assembled Genome (MIMAG)

Table 1 Genome reporting standards for SAGs and MAGs

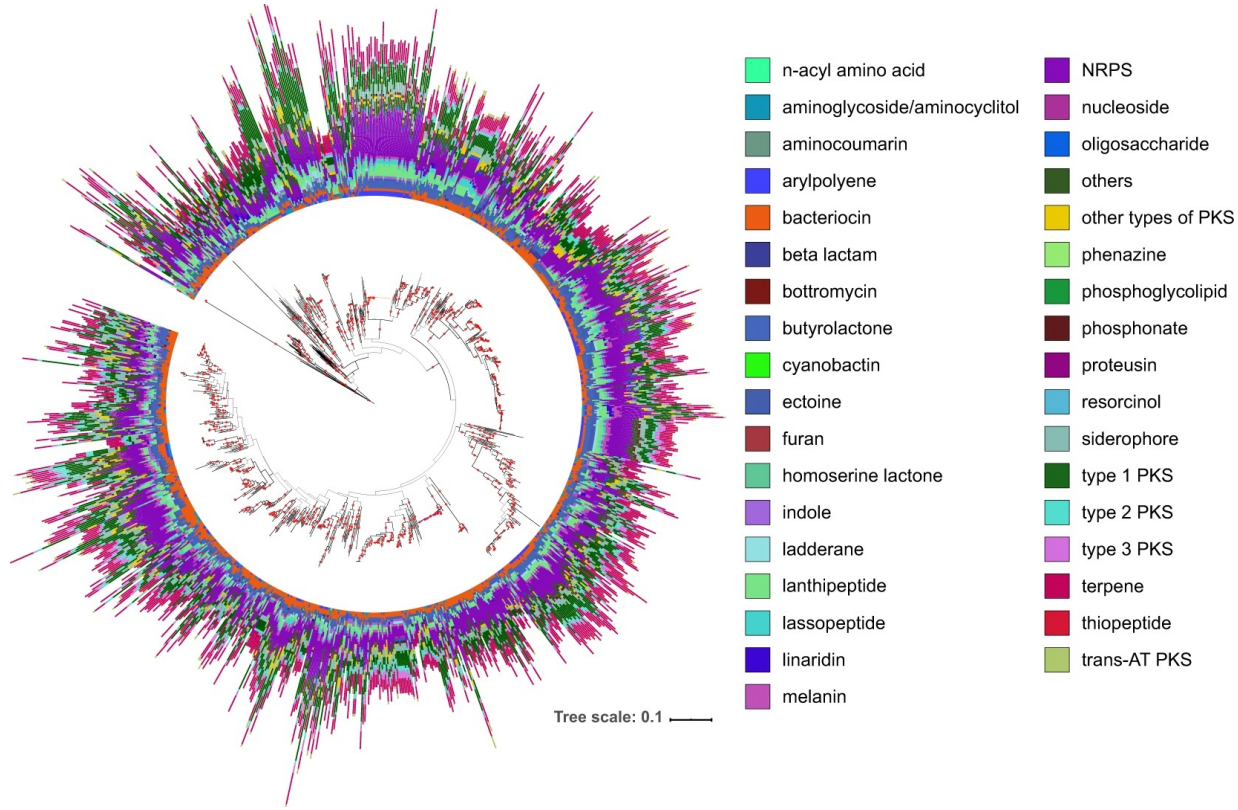
Criterion	Description
Finished (SAG/MAG)	
Assembly quality[a]	Single contiguous sequence without gaps or ambiguities with a consensus error rate equivalent to Q50 or better
High-quality draft (SAG/MAG)	
Assembly quality[a]	Multiple fragments where gaps span repetitive regions. Presence of the 23S, 16S, 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%

Bowers RM *et al.*, *Nat Biotechnol.* (2017)

DATA INTEGRATION



https://www.nature.com/articles/d42473-020-00214-9



Belknap et al., 2020

SUMMARY

Amplicon data - 16S rRNA, 18S rRNA, ITS markers

> provides a snapshot of the taxonomic diversity
> Inexpensive, can process a lot of samples cheaply
> Works well with low biomass samples and samples with high amounts of host DNA
> not good for strain level identification
> can be biased based on primer choice, sample preservation methods,
and other technical artifacts

Shotgun metagenomic data

> Can also generate taxonomic profiles (using multiple target genes)
> can provide potential functional capacity of genome
> can provide strain level taxonomy information
> expensive, requires a lot of DNA compared to amplicon methods
> Aren't great methods for samples with high host DNA content (like for example endosphere)

Which method do you think is best for your specific research question??