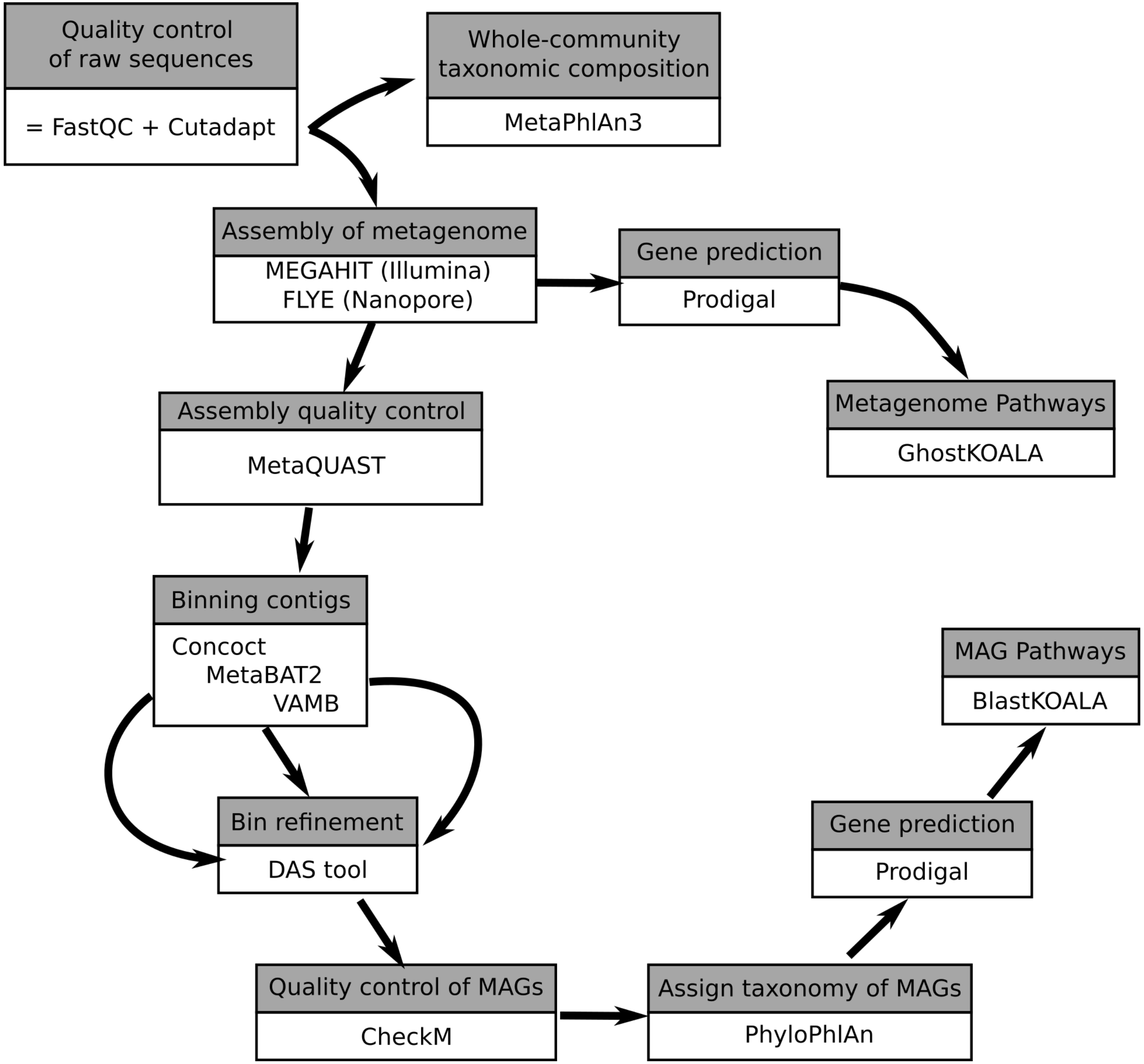
Functional microbiome research – bioinformatics section

Day 4 – binning and refining



Quality control of raw sequences

= FastQC + Cutadapt

Whole-community taxonomic composition

MetaPhlAn3

Assembly of metagenome

MEGAHIT (Illumina)
FLYE (Nanopore)

Gene prediction

Prodigal

Metagenome Pathways

GhostKOALA

Assembly quality control

MetaQUAST

Binning contigs

Concoct
MetaBAT2
VAMB

Bin refinement

DAS tool

Quality control of MAGs

CheckM

Assign taxonomy of MAGs

PhyloPhlAn

Gene prediction

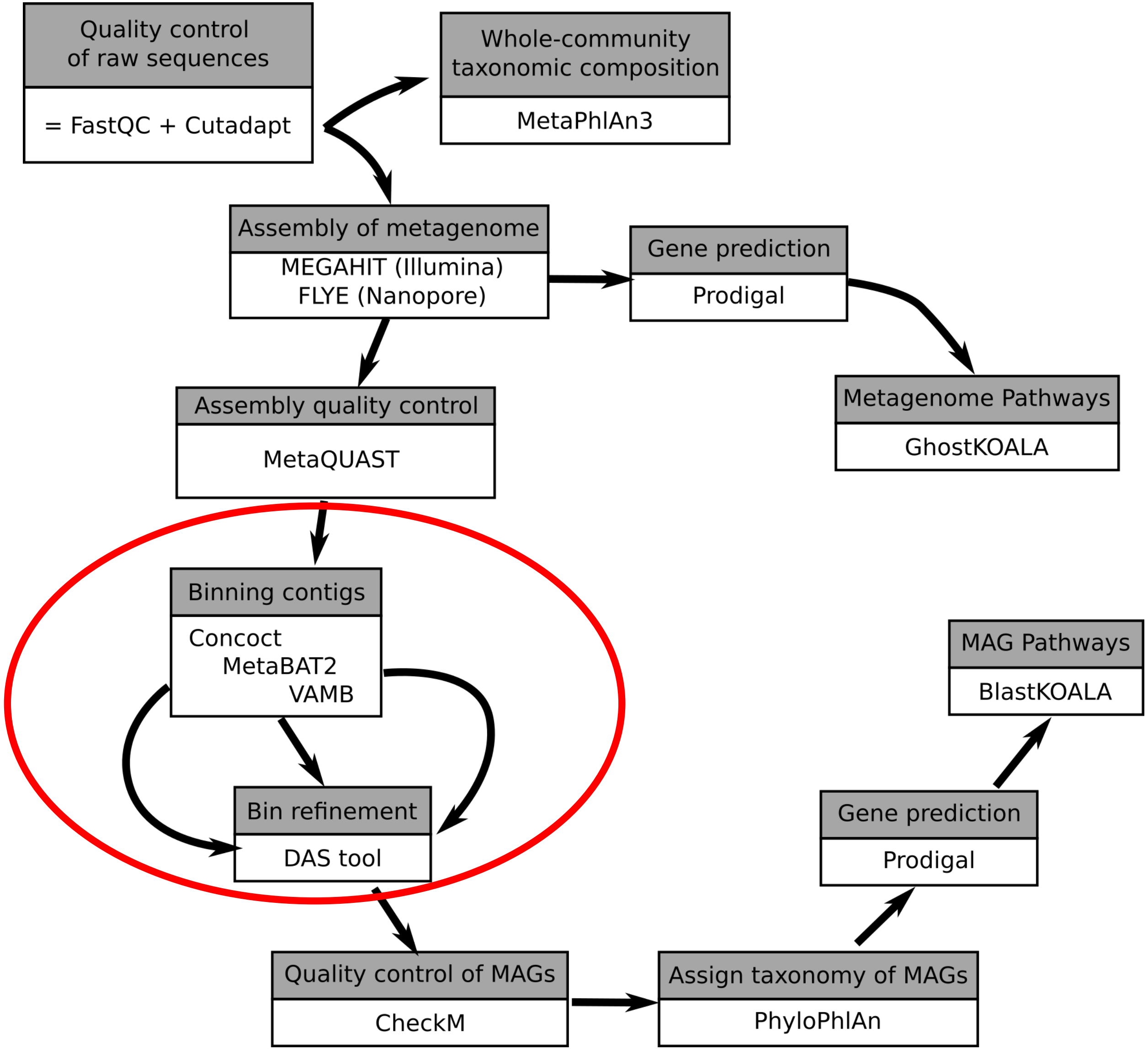
Prodigal

MAG Pathways

BlastKOALA

Day 4 – binning and refining

Day 4 – binning and refining



Quality control of raw sequences

= FastQC + Cutadapt

Whole-community taxonomic composition

MetaPhlAn3

Assembly of metagenome

MEGAHIT (Illumina)
FLYE (Nanopore)

Gene prediction

Prodigal

Metagenome Pathways

GhostKOALA

Assembly quality control

MetaQUAST

Binning contigs

Concoct
MetaBAT2
VAMB

Bin refinement

DAS tool

MAG Pathways

BlastKOALA

Gene prediction

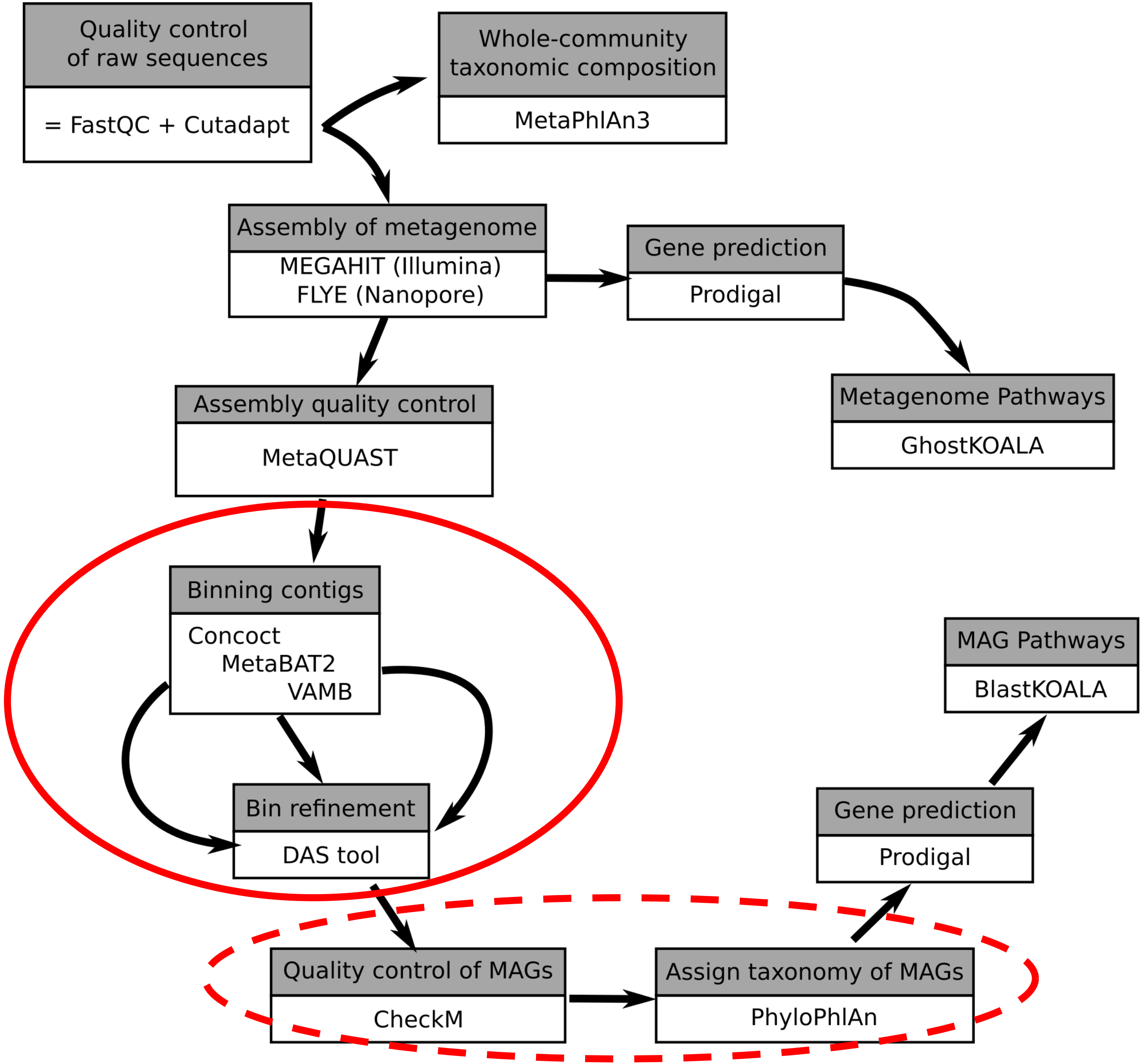
Prodigal

Quality control of MAGs

CheckM

Assign taxonomy of MAGs

PhyloPhlAn



Quality control of raw sequences

= FastQC + Cutadapt

Whole-community taxonomic composition

MetaPhlAn3

Assembly of metagenome

MEGAHIT (Illumina)
FLYE (Nanopore)

Gene prediction

Prodigal

Metagenome Pathways

GhostKOALA

Assembly quality control

MetaQUAST

Binning contigs

Concoct
MetaBAT2
VAMB

Bin refinement

DAS tool

Quality control of MAGs

CheckM

Assign taxonomy of MAGs

PhyloPhlAn

Gene prediction

Prodigal

MAG Pathways

BlastKOALA

Day 4 – binning and refining

(and maybe quality checks and taxonomy of MAGs)

You can grab today’s scripts here:

```
wget https://raw.githubusercontent.com/danchurch/FunctionalMicrobiomePractical2022/main/funmic2023/funMetagenomicScript.txt
```

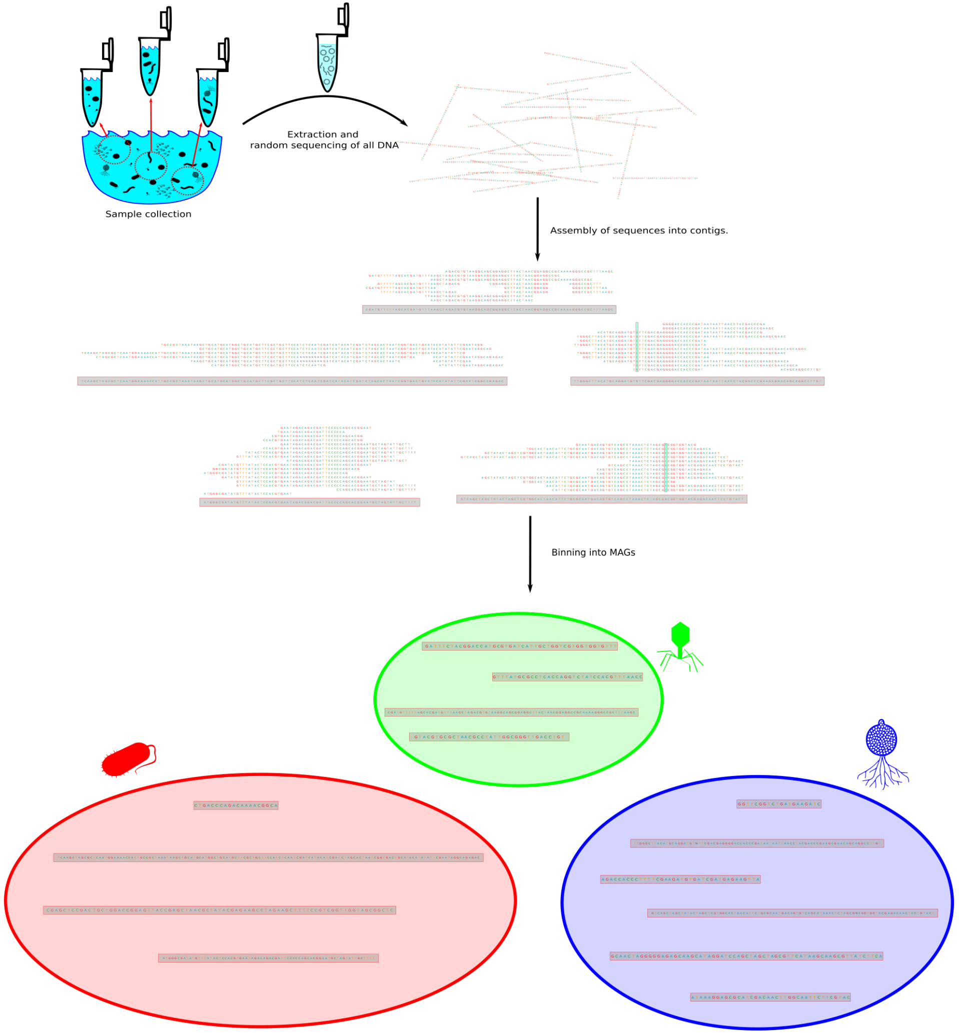
Creating MAGs from metagenomes

De novo assembly of metagenomes almost never finishes with consensus sequences representing an entire chromosome or genomes. Instead, an array of numerous shorter contigs results.

Binning algorithms take several approaches categorize contigs into “bins”, or candidate genomes.

Binning algorithms group contigs based similar:

- coverage (reads),
- K-mers, especially tetranucleotide sequences,
- GC content
- ...

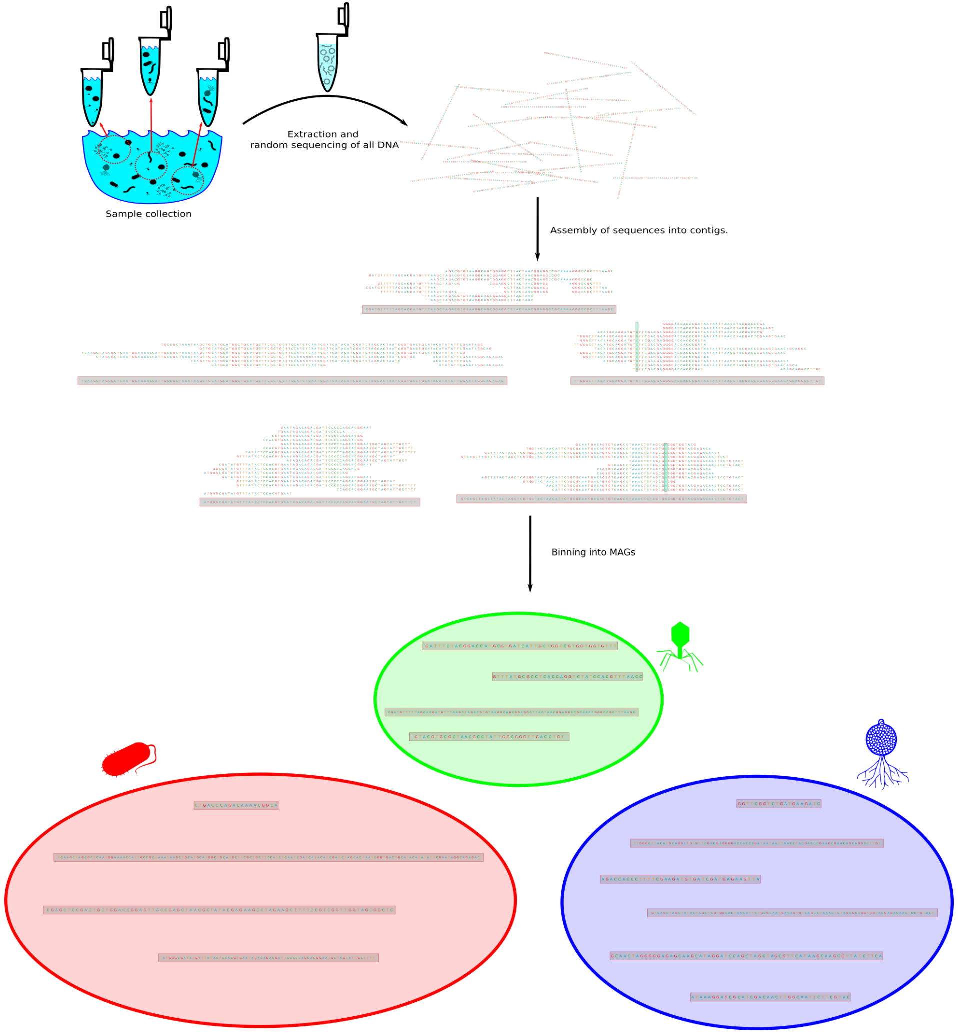

Sample collection
Extraction and
random sequencing of all DNA
Assembly of sequences into contigs.
Binning into MAGs

Creating MAGs from metagenomes

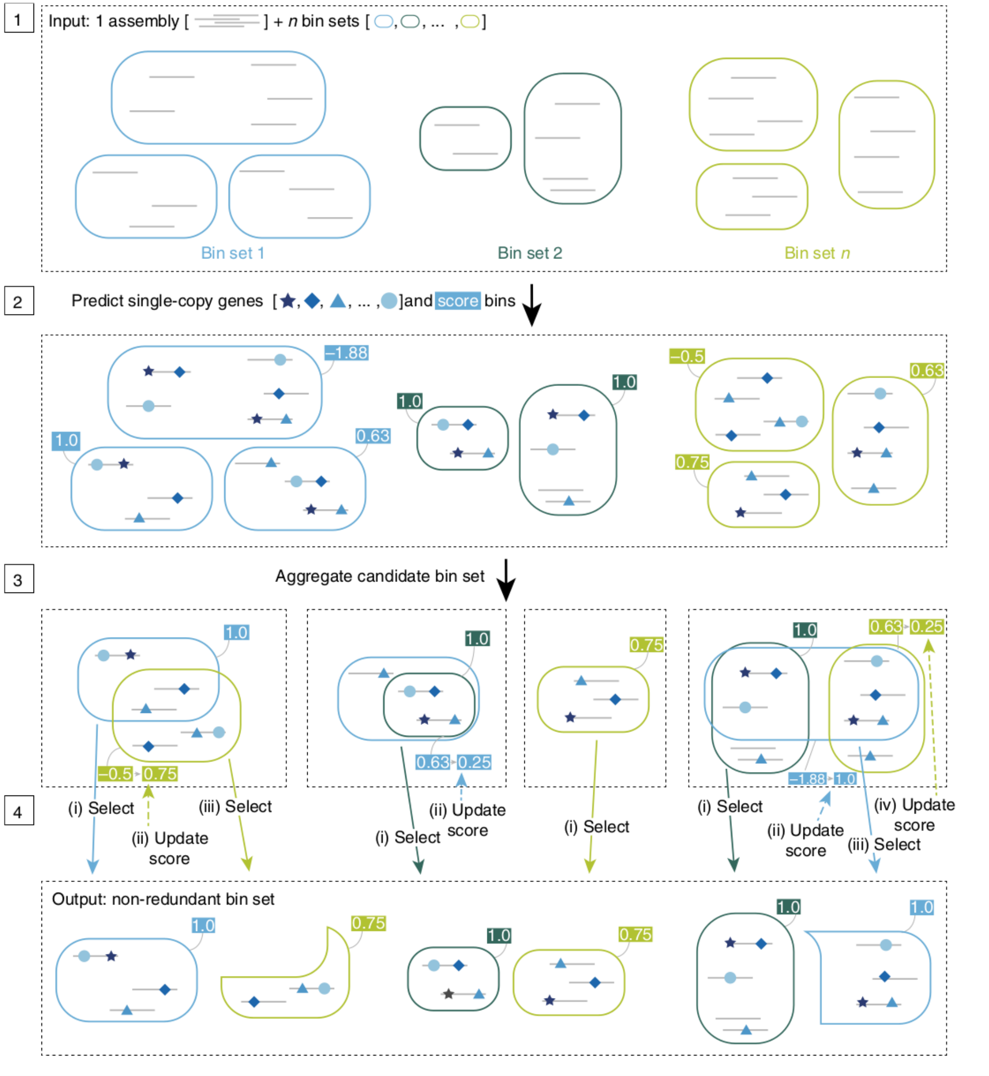
No single binning algorithm yet appears totally sufficient for getting genomes out of metagenomes. Each has strengths and weaknesses.

So we'll use three binning software packages:

1. Metabat2
2. Concoct
3. VAMB

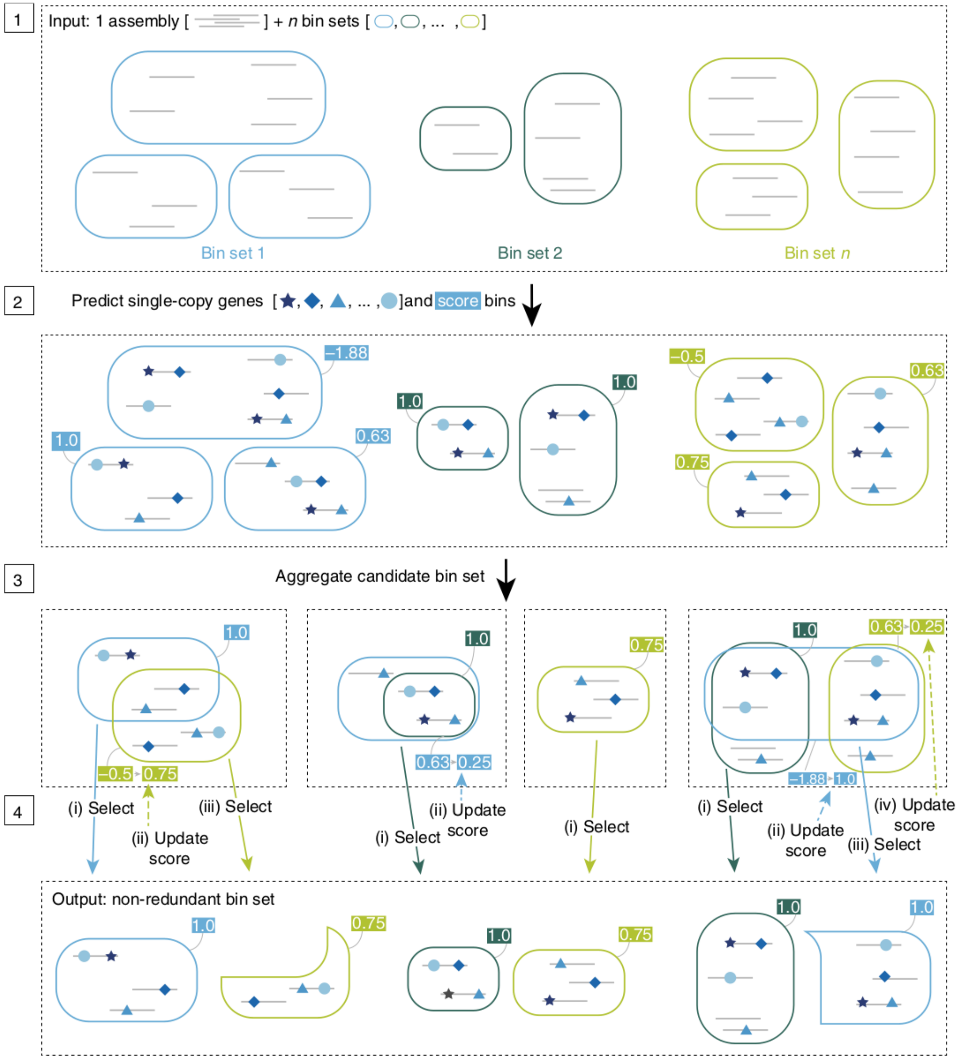

Sample collection
Extraction and
random sequencing of all DNA
Assembly of sequences into contigs.
Binning into MAGs

Creating MAGs from metagenomes



Before we can call them MAGs, these sets of bins undergo a refining process.

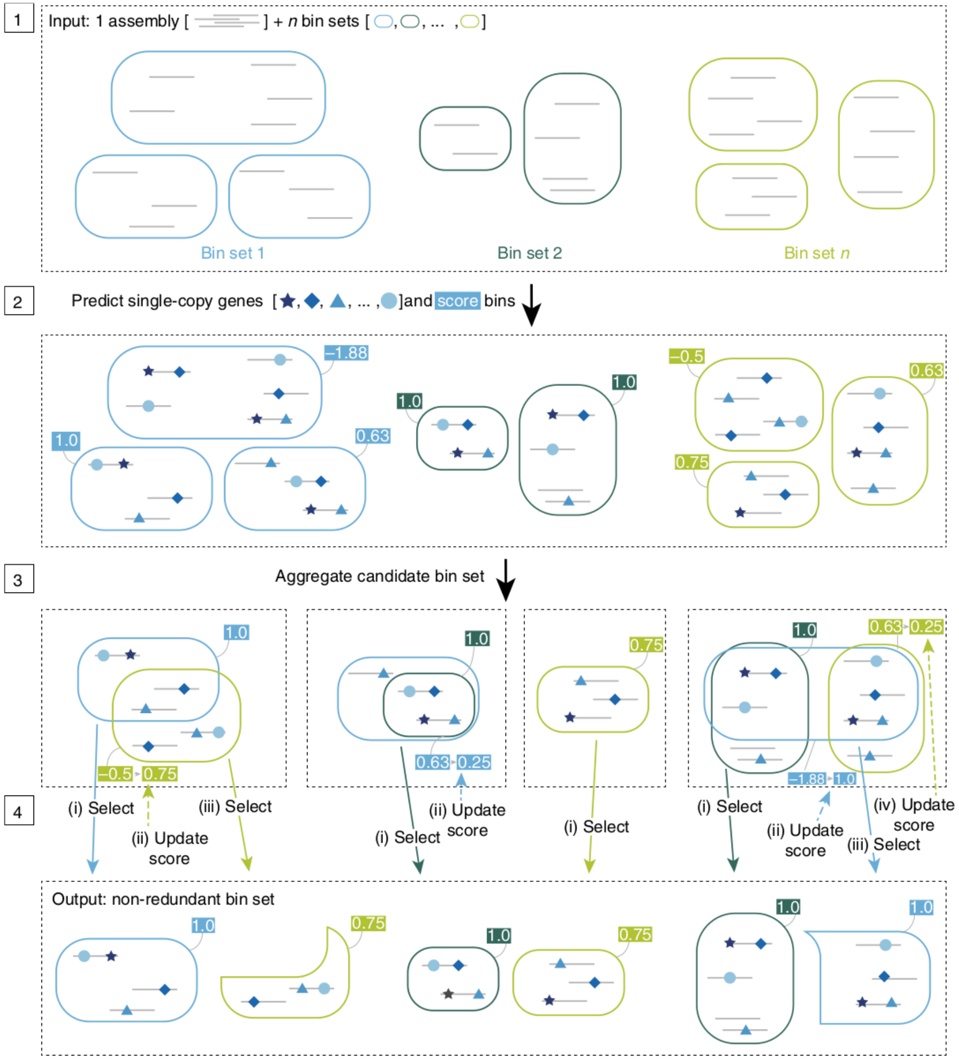
Creating MAGs from metagenomes



Before we can call them MAGs, these sets of bins undergo a refining process.

In one popular refinement software, DAS Tools (Sieber 2018) bins are scored using the predicted behaviour of single copy genes.

Creating MAGs from metagenomes



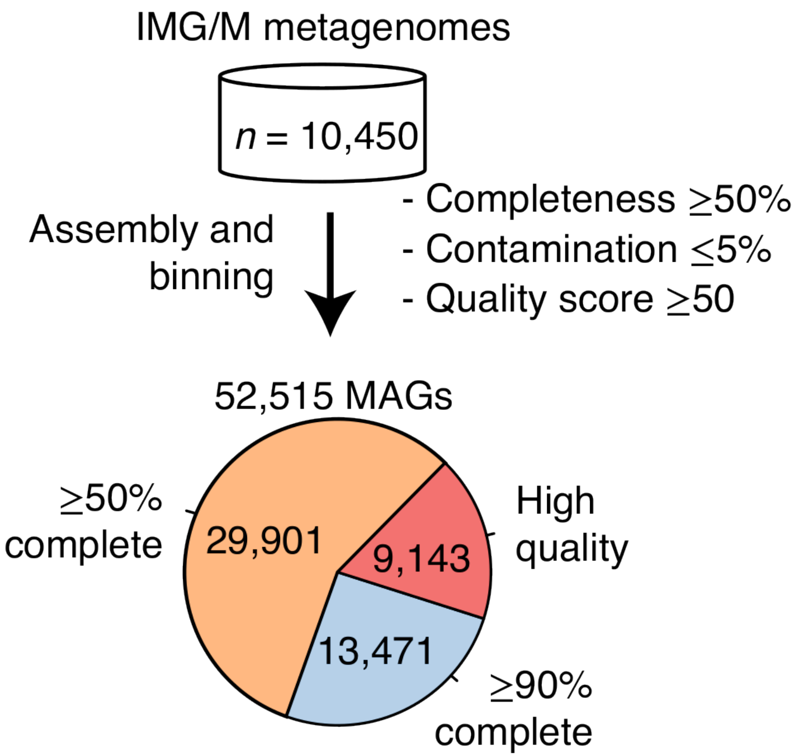
Before we can call them MAGs, these sets of bins undergo a refining process.

In one popular refinement software, DAS Tools (Sieber 2018) bins are scored using the predicted behaviour of single copy genes.

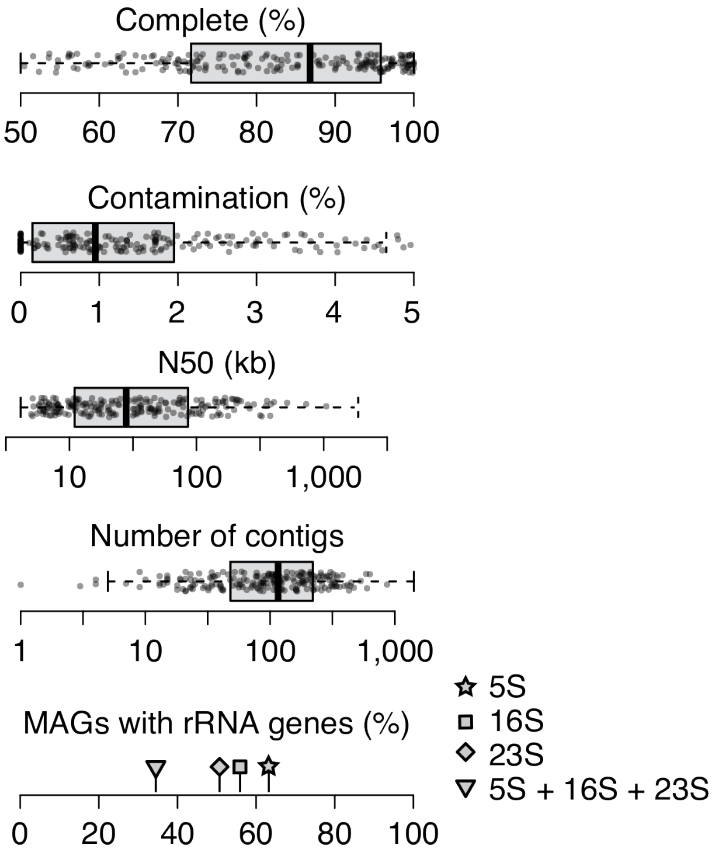
Sets of bins from multiple binning algorithms are compared, with high scoring bins are selected and edited to produce a new, aggregate set of bins with lower redundancy and higher completeness.

(Nayfach 2020)

Creating MAGs from metagenomes



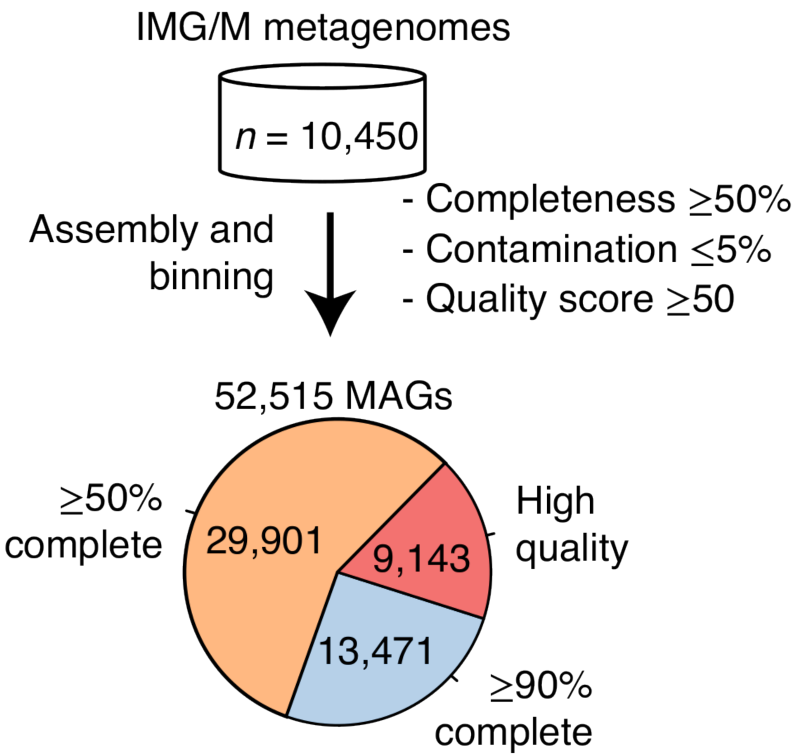
- Completeness ≥50%
- Contamination ≤5%
- Quality score ≥50



Why not just use (Meta)quast again?

(Nayfach 2020)

Creating MAGs from metagenomes



IMG/M metagenomes

n = 10,450

Assembly and binning

- Completeness ≥50%
- Contamination ≤5%
- Quality score ≥50

52,515 MAGs

≥50% complete

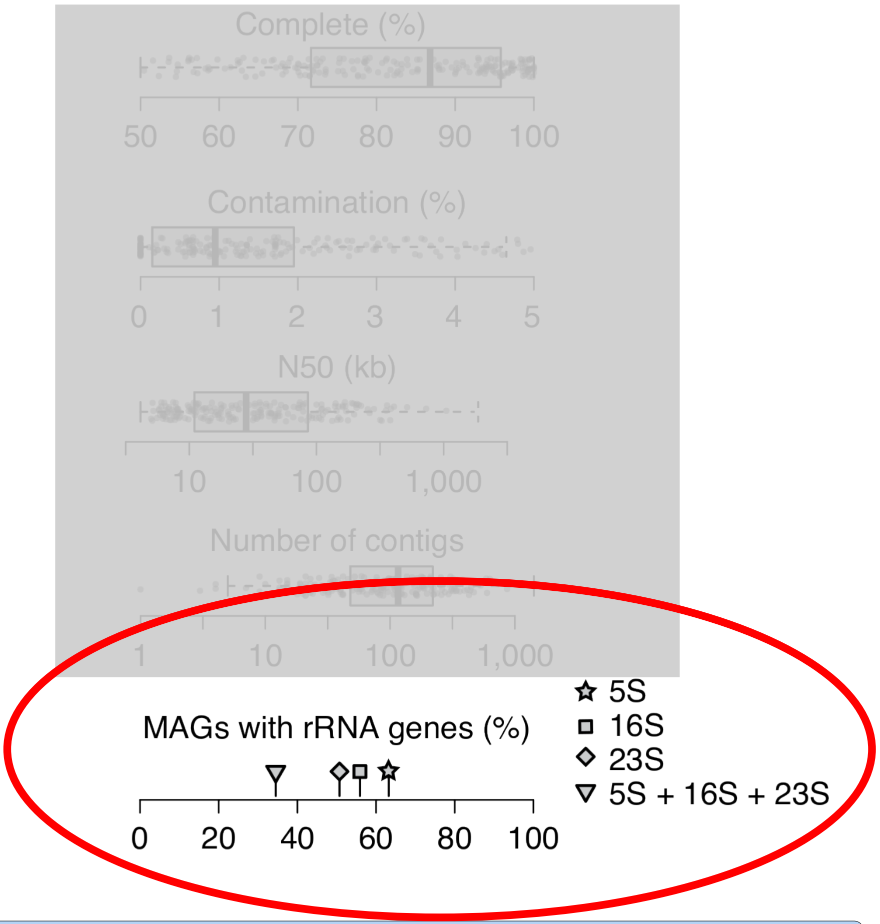
29,901

9,143

High quality

13,471

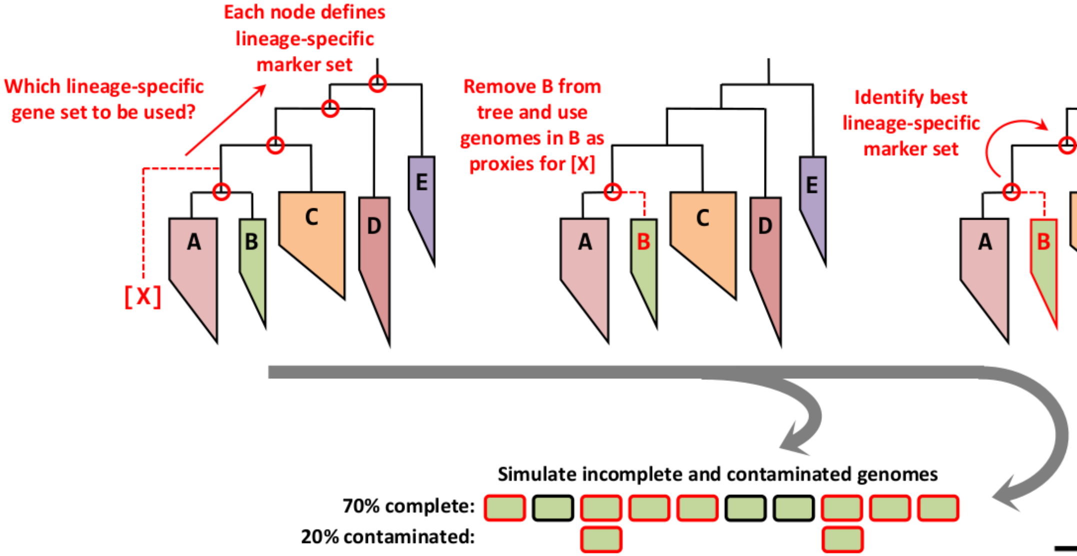
≥90% complete

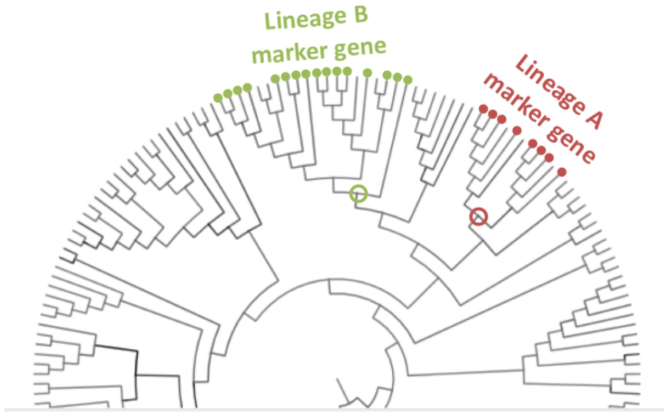


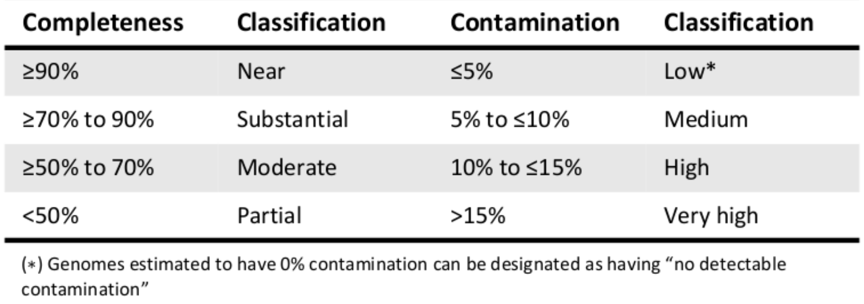
Why not just use (Meta)quast again?

(Parks 2015)

Quality check tool: CheckM


Each node defines lineage-specific marker set
Which lineage-specific gene set to be used?
[X]
A
B
C
D
E
Remove B from tree and use genomes in B as proxies for [X]
Identify best lineage-specific marker set
Simulate incomplete and contaminated genomes
70% complete:
20% contaminated:





Completeness	Classification	Contamination	Classification
≥90%	Near	≤5%	Low*
≥70% to 90%	Substantial	5% to ≤10%	Medium
≥50% to 70%	Moderate	10% to ≤15%	High
<50%	Partial	>15%	Very high

(*) Genomes estimated to have 0% contamination can be designated as having “no detectable contamination”

CheckM also uses a very large database of markers (not just 16s) to do quality control on MAGs. Using single copy marker genes that are present in all life or a target clade (e.g. Bacteria or Acidobacteriota), estimates of contamination and completion can be made.

(Asnicar 2020)

Assigning taxonomy to MAGs

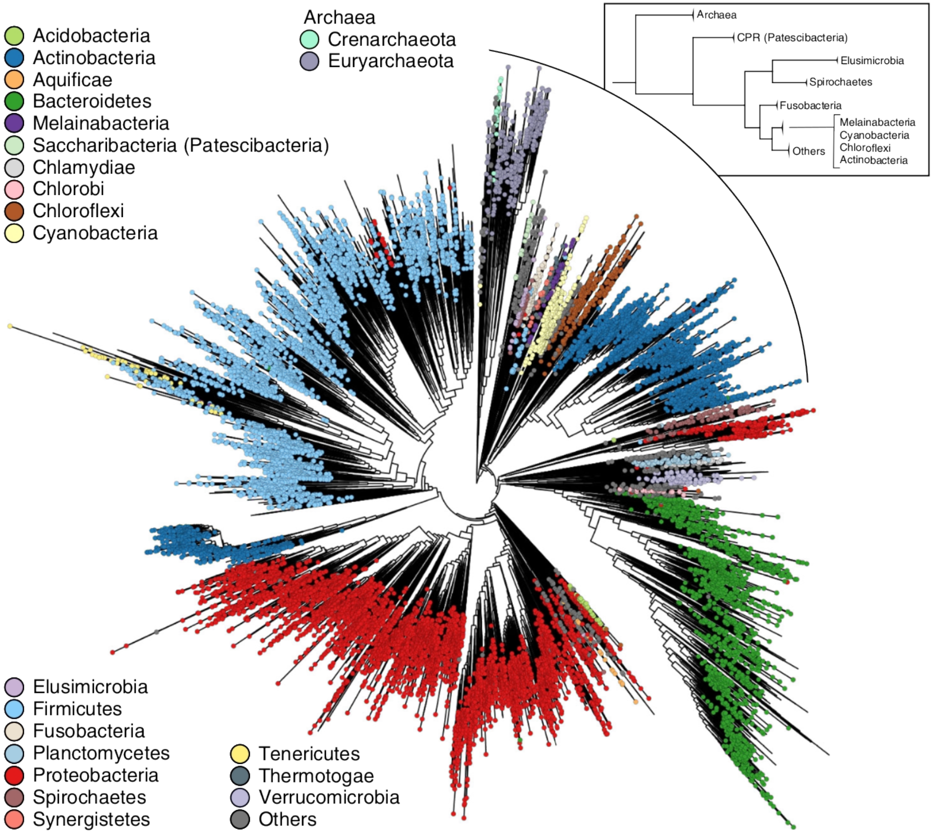


Fig. 4 PhyloPhlAn 3.0 microbial tree-of-life with 17,672 species-representative genomes from 51 known and 84 candidate phyla.

Phylogeny/Taxonomy can also be assigned to MAGs. This can be done using classic rRNA barcode approaches if the 16s or ITS region is recovered in the MAG.

(Asnicar 2020)

Assigning taxonomy to MAGs

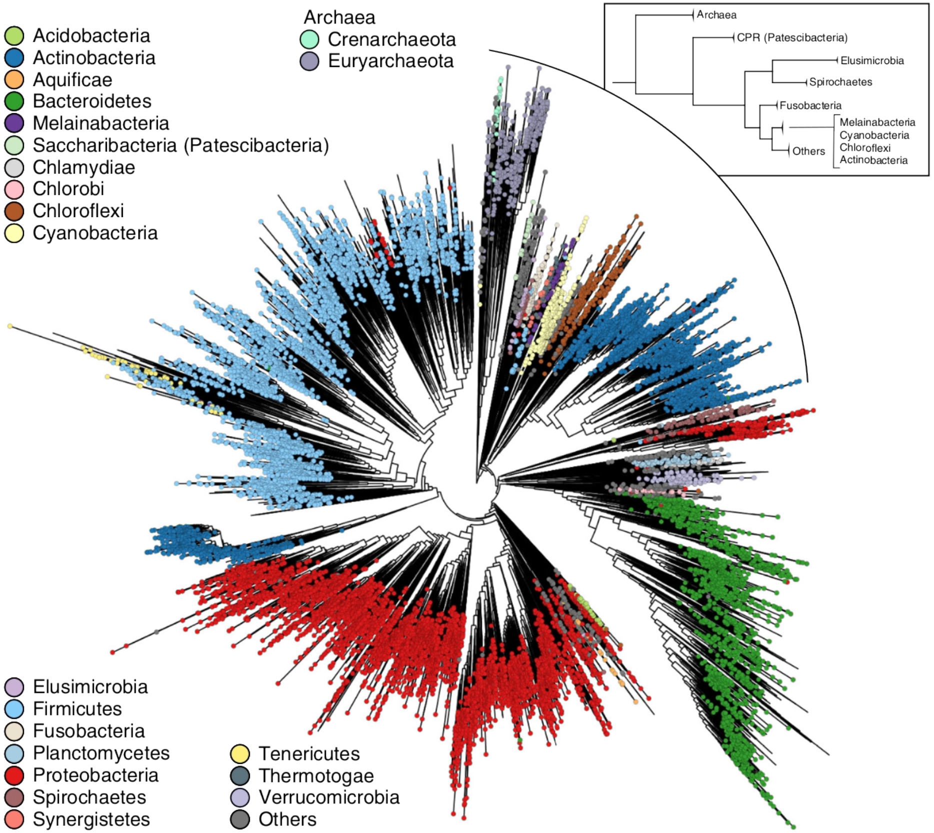


Fig. 4 PhyloPhlAn 3.0 microbial tree-of-life with 17,672 species-representative genomes from 51 known and 84 candidate phyla.

But genomes can provide much greater information than a single barcode, and barcodes can be lost (see above).

(Asnicar 2020)

Assigning taxonomy to MAGs

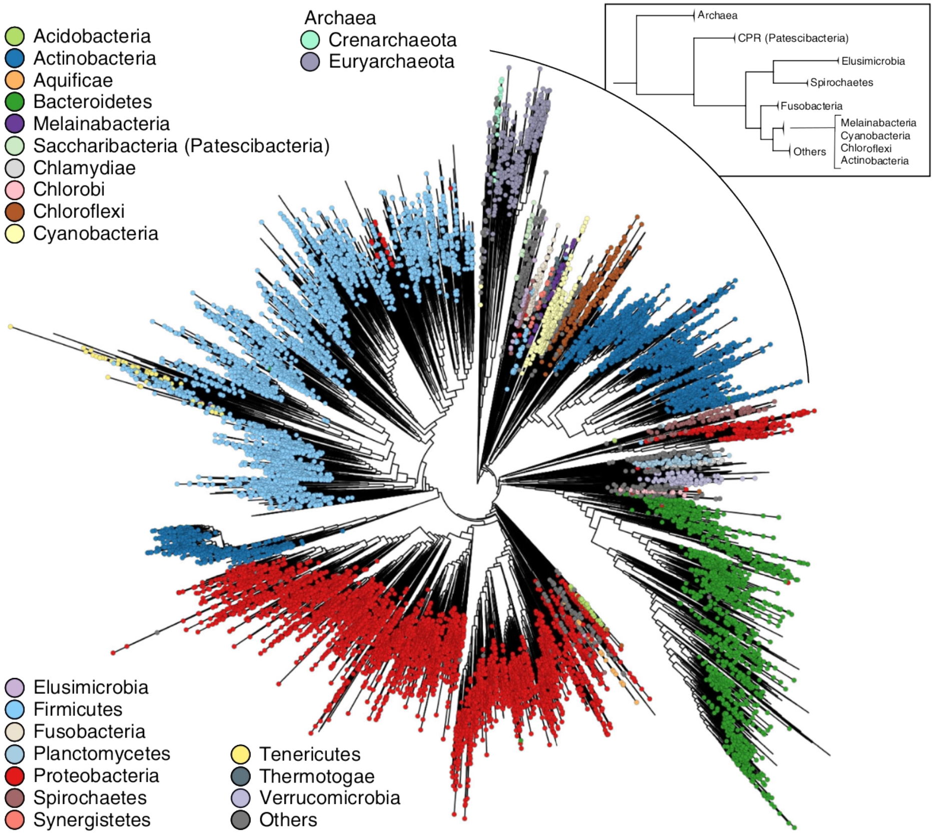
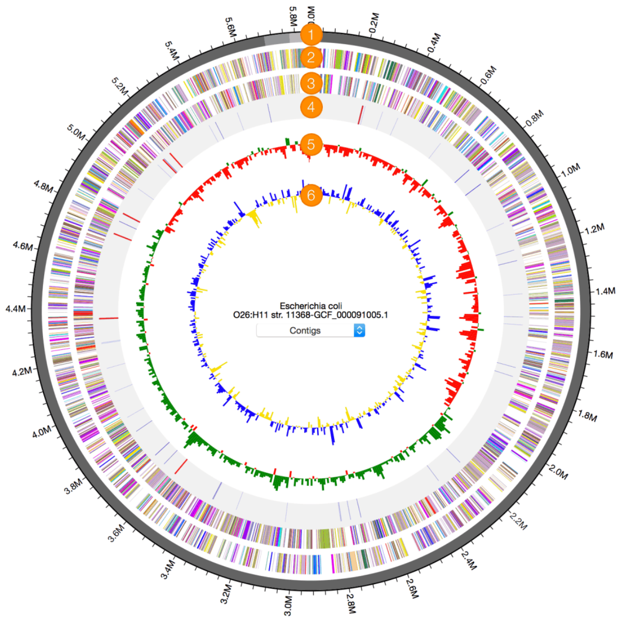


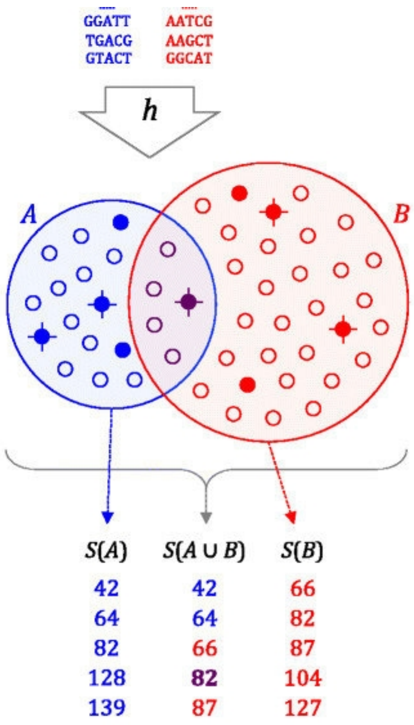
Fig. 4 PhyloPhlAn 3.0 microbial tree-of-life with 17,672 species-representative genomes from 51 known and 84 candidate phyla.

Several approaches exist, but we’ll try PhyloPhlan, from biobakery tools.

(Ondov 2016)

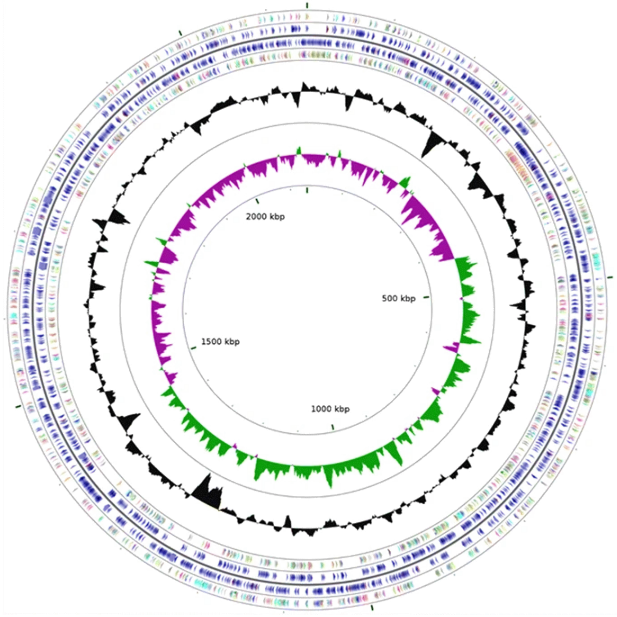
Creating MAGs from metagenomes





$S(A)$	$S(A \cup B)$	$S(B)$
42	42	66
64	64	82
82	66	87
128	82	104
139	87	127

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \approx \frac{|S(A \cup B) \cap S(A) \cap S(B)|}{|S(A \cup B)|}$$



PhyloPhlan uses minHash sketches of entire genomes that can be used to measure dissimilarity among genomes, also called Mash distances (Asnicar 2020). A MAG can be mashed and compared to the minHash sketch from other published genomes on databases like NCBI, and a closest match found.