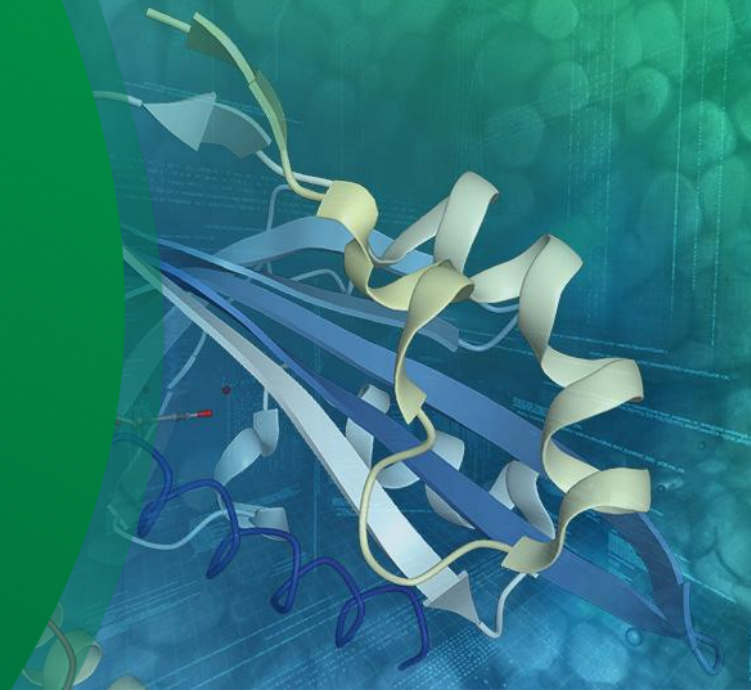
Metagenomics



Germana Baldi, Varsha Kale

Bioinformaticians at MGnify

EMBL-EBI

Who are we?

MGnify

EMBL

HoloFood

Rob Finn
Team Leader
Section Head

Lorna Richardson
Team Coordinator

Varsha Kale
Bioinformatician

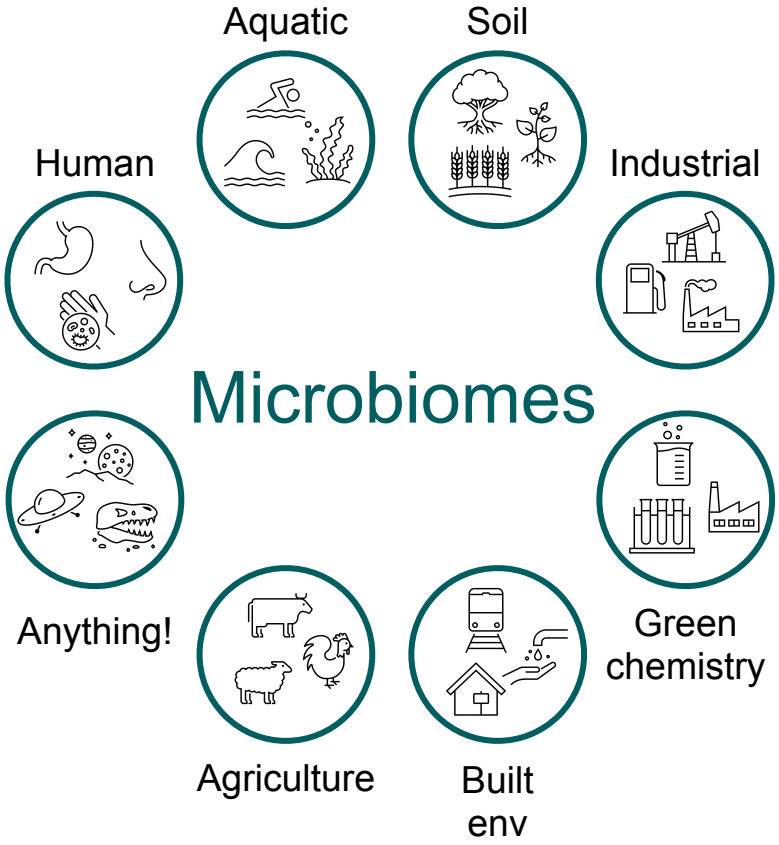
Germana Baldi
Bioinformatician

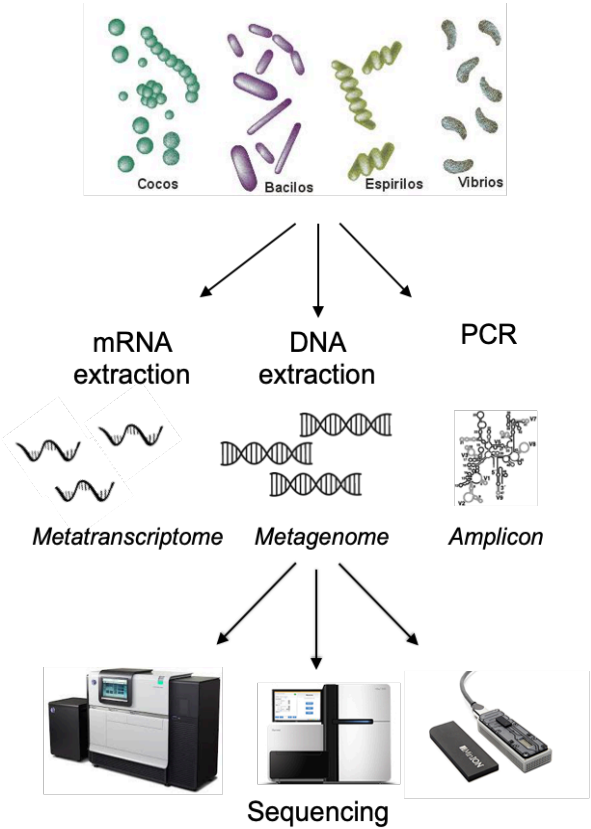
Sandy Rogers
Website Developer

What is metagenomics?

META	“transcending”, “more comprehensive”	Transcends the individual organisms to focus on the community more comprehensively
GENOMICS	“the study of genomes”	

What is metagenomics?





EMBL-EBI

Sequencing Technology

Illumina

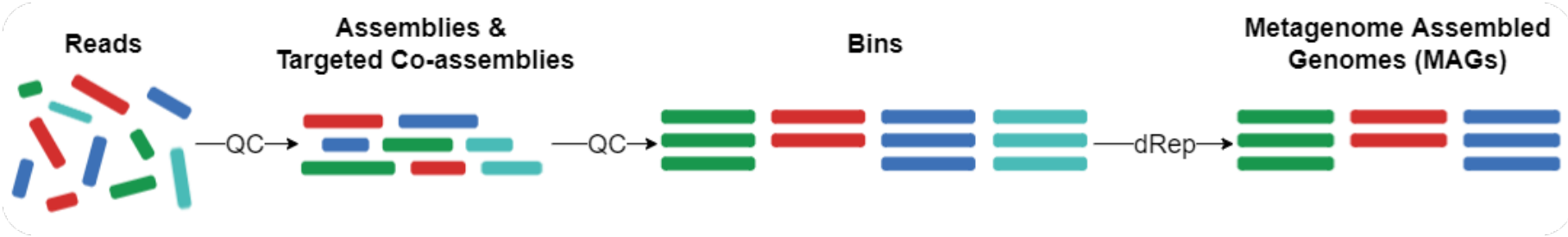
- Short-reads (50-250 bp)
- 4-20k million reads per run
- Error rate <0.1%
- Sample preparation $50 - $100
- Run cost $1k - $4k

Oxford Nanopore/PacBio SMRT

- Long-reads (>1 kbp)
- 20-100k reads per run
- Error rate 10-15%
- Sample preparation $100 - $500
- Run cost $1k - $2k

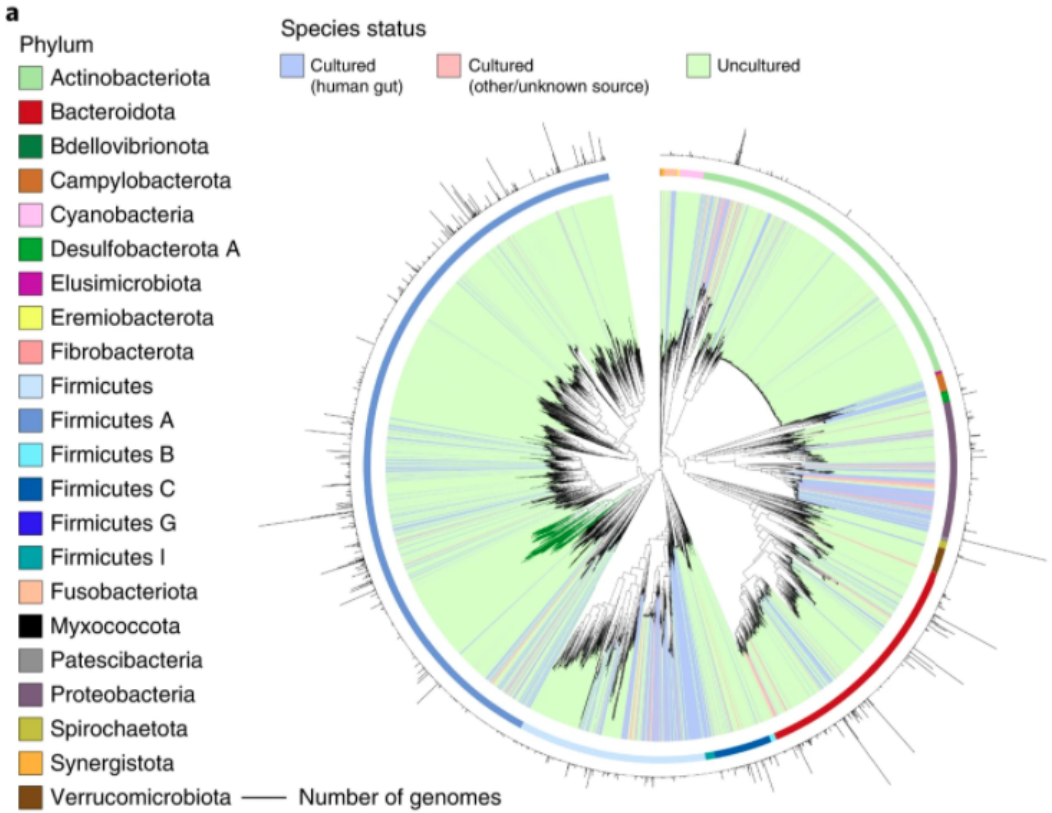
Sequencing

Overview of MAG generation workflow



EMBL-EBI

Most gut species lack isolate genomes



Almeida A. et al., Nat Biotechnol. (2021)

EMBL-EBI

Things to note

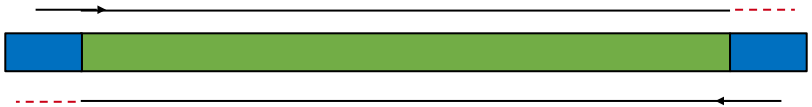
- The majority of what we will cover is prokaryotic-focused (also for historical reasons!)
- Your sample may contain a lot of variety
  - Viruses
  - Eukaryotes
  - Prophages
  - Plasmids
  - …
- There is no one correct answer to “how should I analyse my microbiome data?” (sorry!)

Methods of quality control

Cleaning raw reads reduces the risk of contamination in downstream analyses.

You might want to consider:

- GC content
- Duplicates
- Trim reads/regions by quality score
- Sequence length
- Contamination from:
  - Human/host: map samples against reference
  - Reagents: use negative controls
  - Sequencing process: clipping known primers/adapter sequences



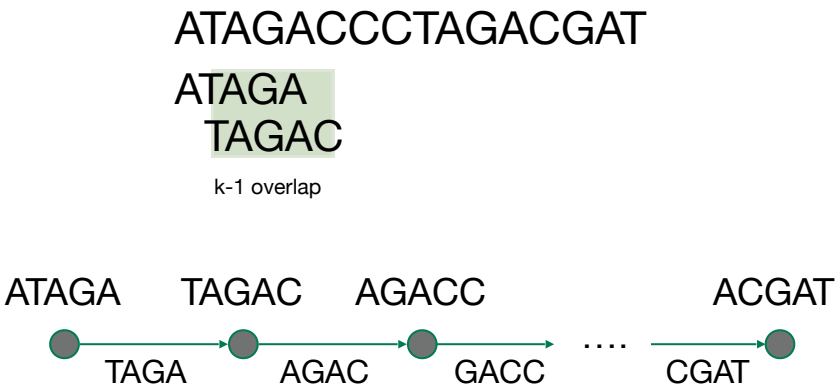
It is important to keep high quality controls throughout the whole workflow

De novo assembly

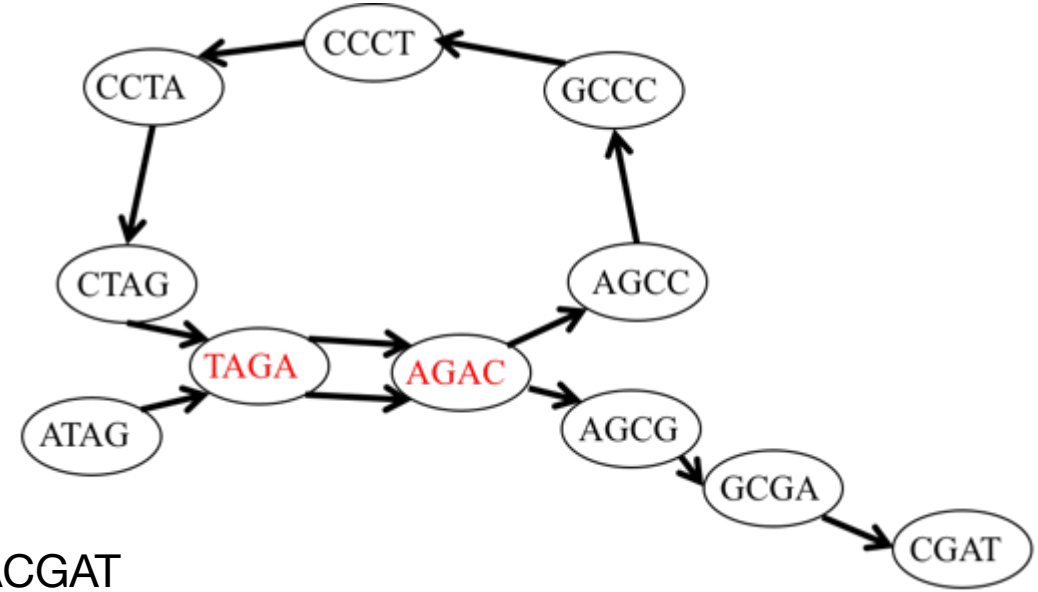
Method for reconstructing genomes from DNA/RNA fragments, with no prior knowledge of the original sequence or the order of those fragments.

De Bruijn graph

- Extract all substrings of length k from input reads (i.e. **k-mers**)
- Model relationship in a de Bruijn graph
  - nodes: k-mers
  - edges: adjacent k-mers overlapping by k-1 letters
- Visit each node exactly once through the graph (i.e. identify Eulerian path)



De novo assembly



ATAGACCCTAGACGAT

ATAGA

TAGAC

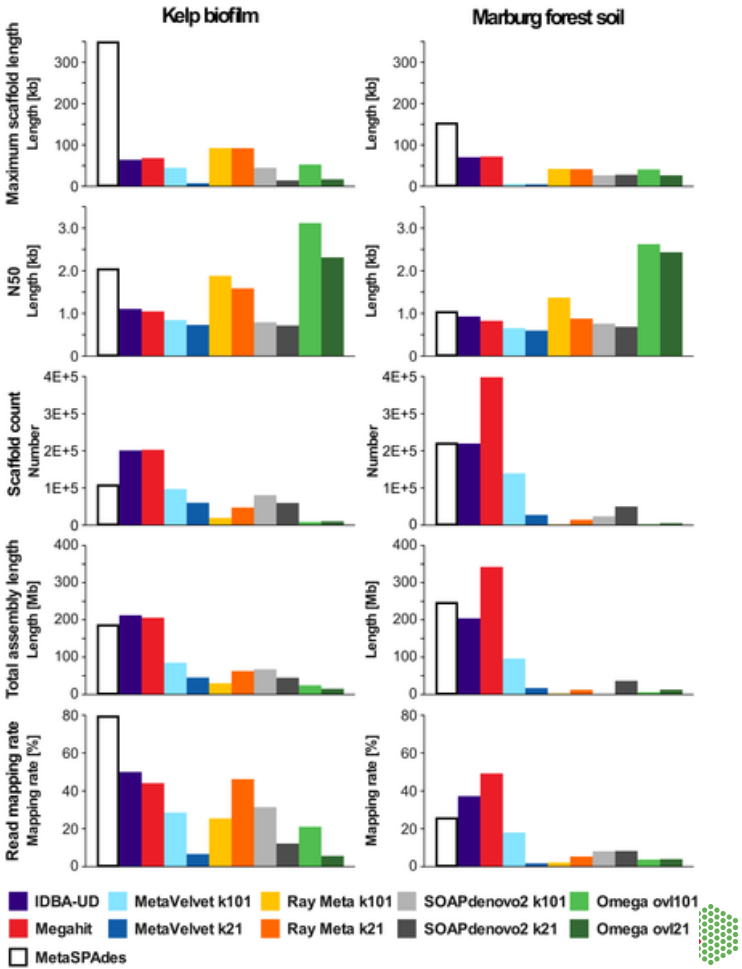
k-1 overlap

Types of assemblers

>30 different tools available

- MetaSPAdes and Megahit are the most widely used
- Different assemblers have different computational characteristics
  - Computational resources are limiting for assemblies
- Different performances on different (micro)biomes

Compromising is key



Vollmers J. et al. PLOS ONE 12(1) (2017)

Co-assembly

Merging (appending) two or more samples to be assembled together

PROS

- More data, better/longer assemblies
- Access to lower abundant organisms

CONS

- Higher computational overhead
- Risk of shattering the assembly graph by strain variations
- Risk of increased contamination

When to co-assemble?

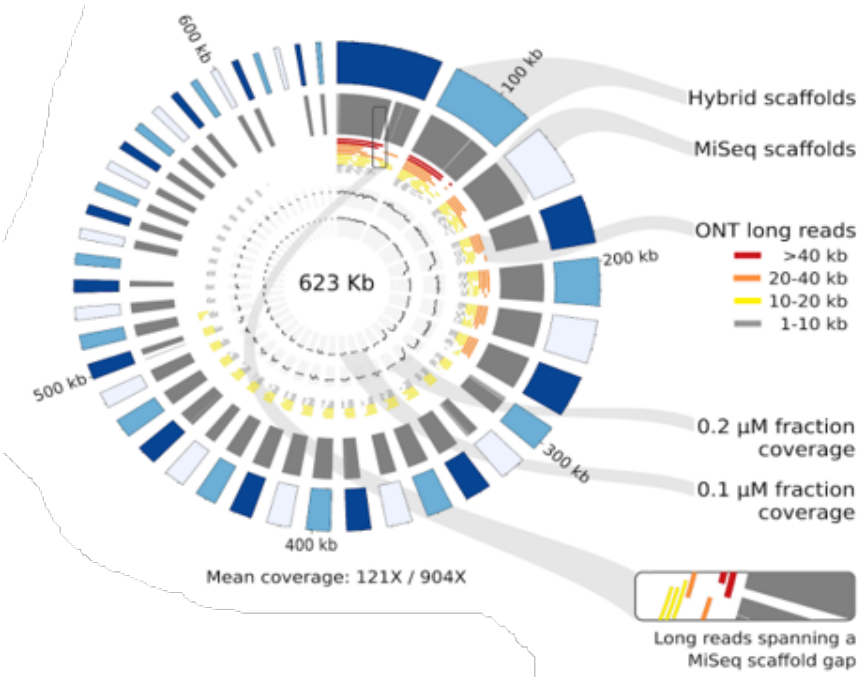
- Same sample
- Same sampling event
- Longitudinal sampling of the same individual
- Related samples

Combining sequencing technologies

Short-reads + Long-reads = better assembly

2 main strategies

- Assemble short reads, extend contigs and resolve repetitive regions with long reads (**hybrid assembly**)
- Assemble long reads, polish them with short reads



Overholt, W. A. et al., Environmental Microbiology (2020)

Binning

Supervised approach

- Relies on known reference genomes
- Uses homology or sequence composition similarity for binning

Multiple binners exist:

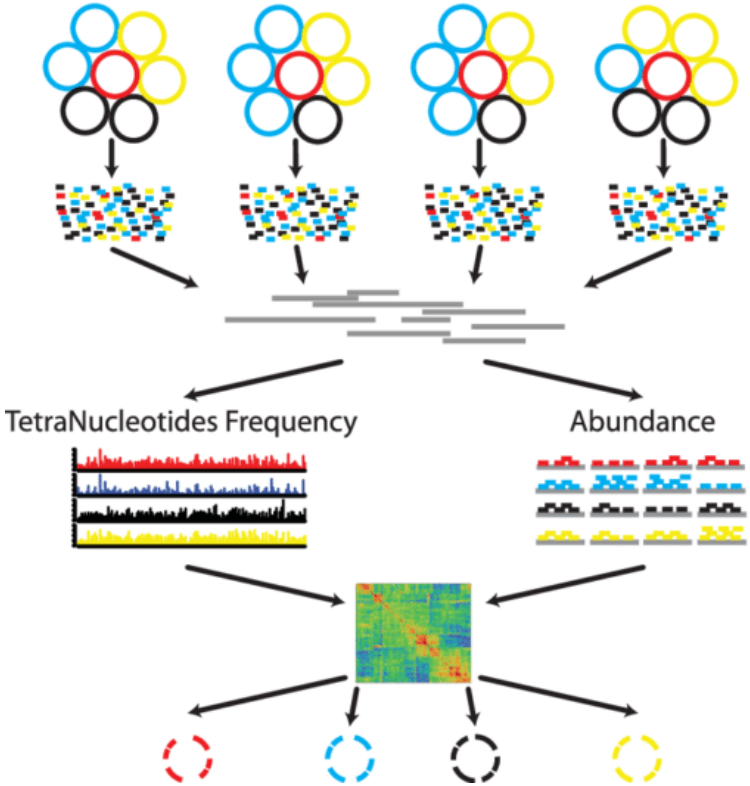
- MetaBAT
- MaxBin2
- CONCOCT
- Semibin
- …

Unsupervised approach

- Does not need a reference genome
- Relies on sequence composition similarity and/or species abundance for binning

In the practical!

MetaBAT

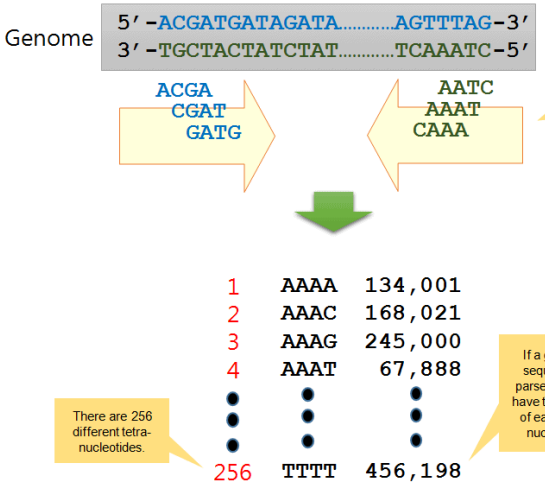


Preprocessing

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

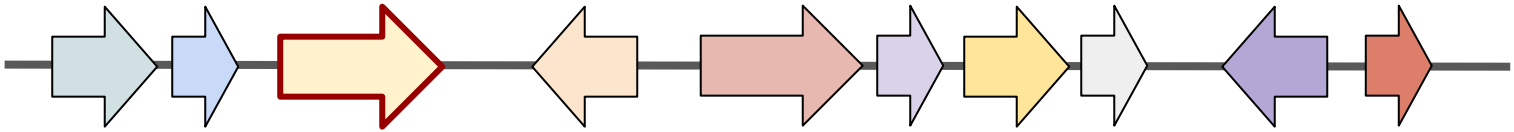


Kang D. D. et al., PeerJ (2015)

Quality assessment: CheckM

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

Reference SCMG set



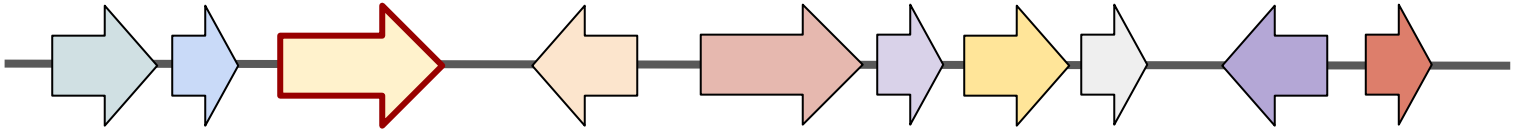
Parks D. H. et al., Genome Res. (2015)

EMBL-EBI

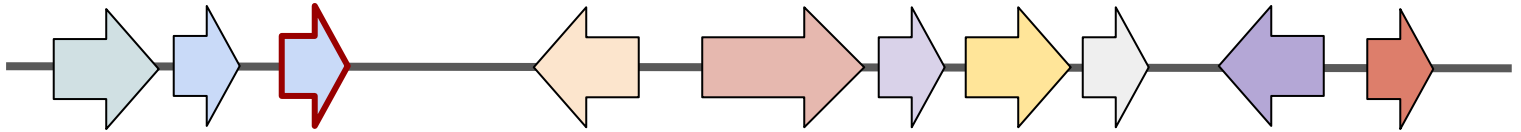
Quality assessment: CheckM

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

Reference SCMG set



New genome assembly to evaluate



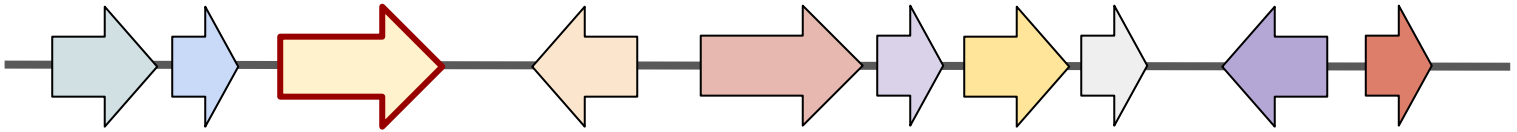
Parks D. H. et al., Genome Res. (2015)

EMBL-EBI

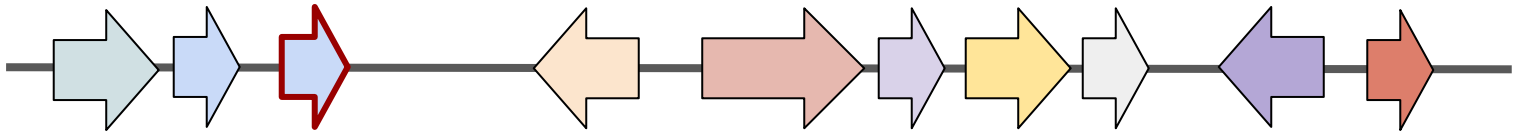
Quality assessment: CheckM

Uses a set of lineage-specific single-copy marker genes (SCMG) - genes that are present in 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 D. H. 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)

CheckM output:

Completeness: 85%

Contamination: 7%

Strain heterogeneity: 100%

Contamination is likely to come from other strains of the same species

Parks D. H. 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)

CheckM output:

Completeness: 85%

Contamination: 7%

Strain heterogeneity: 0%

Contamination is likely to come from distant species

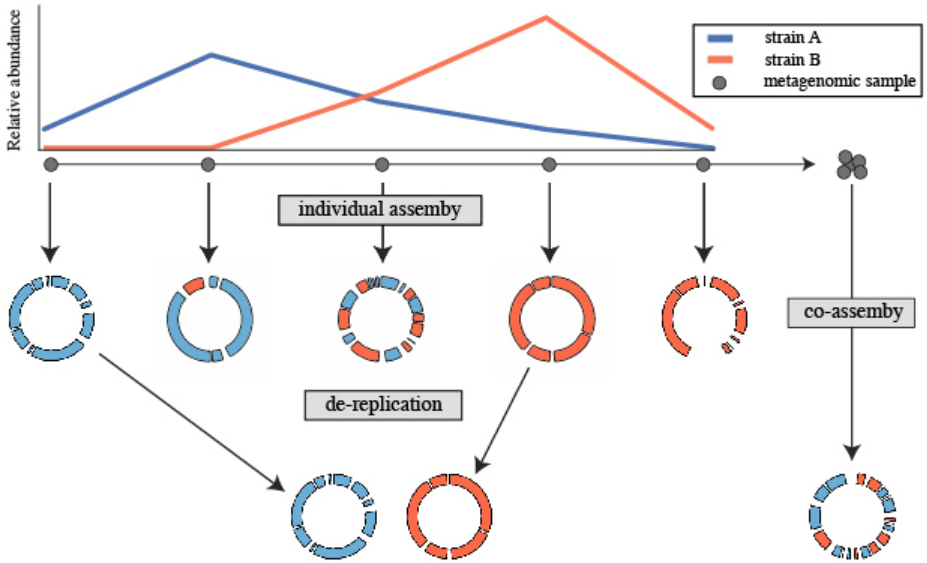
Tools to remove contamination:

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

Parks D. H. et al., Genome Res. (2015)

EMBL-EBI

Removing redundant genomes: dRep



Genomes

↓ mash

Primary clusters

↓ gANI

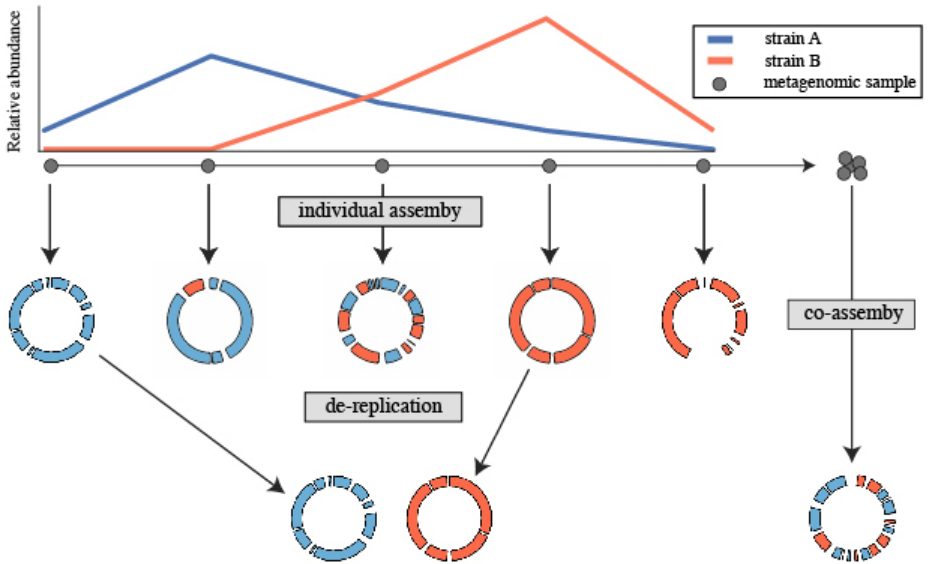
Secondary clusters

↓ cluster scoring

Representative genomes

Olm M. R. et al., ISME J (2017)

Removing redundant genomes: dRep



score = $A*Completeness - B*Contamination + C*(Contamination*(strain\ heterogeneity/100)) + D*log(N50) + E*log(size)$

Olm M. R. et al., ISME J (2017)

Genomes

↓ mash

Primary clusters

↓ gANI

Secondary clusters

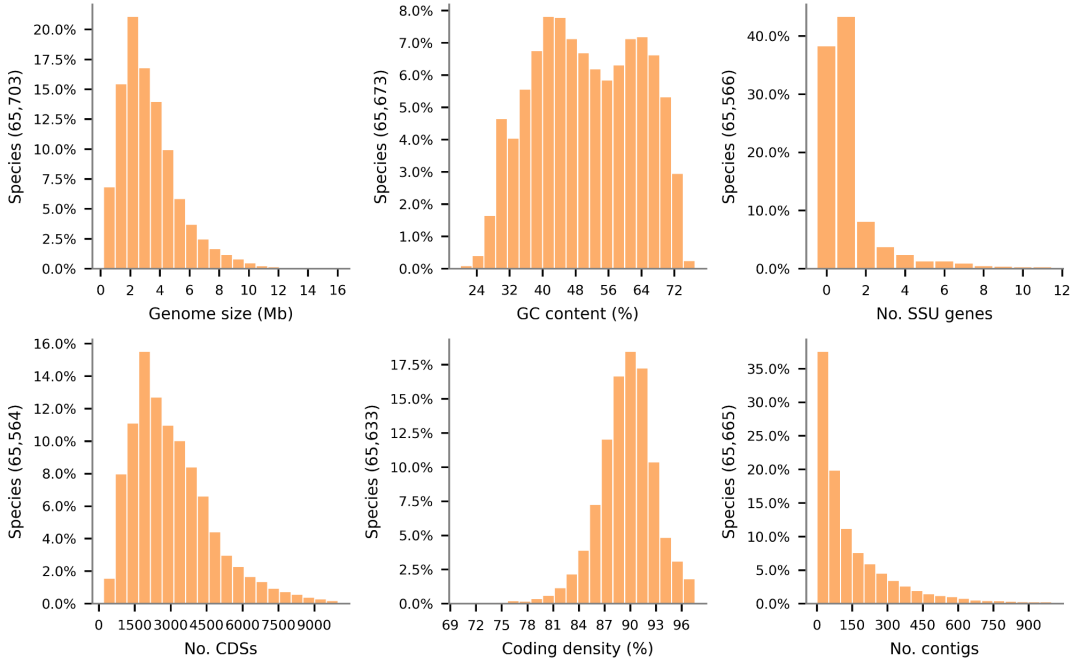
↓ cluster scoring

Representative genomes

Species-level ANI distance: 95%
Strain-level ANI distance: 99%

The Genome Taxonomy Database (GTDB)

- Different from NCBI taxonomy, which relies more on isolate genomes
- GTDB balances taxonomic groups based on number of organisms within taxonomic levels
- ~317,000 genomes (Bacteria and Archaea, MAGs and isolates) organised into ~65,000 species clusters
- GTDB-tk allows placement of your genomes within this framework



Parks D. H. et al., Nature Biotechnol. (2020)

EMBL-EBI

Acknowledgements

Rob Finn

Lorna Richardson

Varsha Kale

Sandy Rogers

Tatiana Gurbich

Juan Caballero

MGnify team