# Environmental metagenomics

MAG annotation and downstream analyses




Igor S. Pessi & Antti Karkman, University of Helsinki

# You’ve got MAGs, and now what?

# But do keep in mind

Just binning MAGs is not (shouldn't be) a research question...

...and it's not a competition

# Things you can do: taxonomic assignment

Examples of tools you can use for that:

- CheckM: https://github.com/Ecogenomics/CheckM
- GTDB-tk: https://ecogenomics.github.io/GTDBTk
- Custom phylogenetic/phylogenomic analyses

# Things you can do: functional annotation and metabolic reconstruction

Examples of tools you can use for that:

- BLAST/DIAMOND: https://github.com/bbuchfink/diamond
- HMMER: http://hmmer.org
- Prokka: https://github.com/tseemann/prokka
- RAST: https://rast.nmpdr.org
- GraftM: https://github.com/geronimp/graftM
- METABOLIC: https://github.com/AnantharamanLab/METABOLIC

# Things you can do: functional annotation and metabolic reconstruction

Examples of databases you can use:

| | |
|---|---|
| **KEGG** | Collection of databases dealing with genomes, biological pathways, diseases, drugs and chemical substances |
| **UniProt** | Aggregate of two databases: SwissProt with functional annotations obtained from the literature and subjected to human review and TrEMBL with functional annotations computationally assigned |
| **Pfam** | Curated database of protein families |
| **Interpro** | Curated database of protein families |
| **Metacyc** | Highly curated metabolic database that contains metabolic pathways, enzymes, metabolites, and reactions from all domains of life |
| **GO** | The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism. Three structured, controlled vocabularies (ontologies): biological processes, cellular components and molecular functions |
| **SEED** | A comparative genomics environment consisting of databases of protein families (FIGfam) and metabolic pathways (Subsystems) |

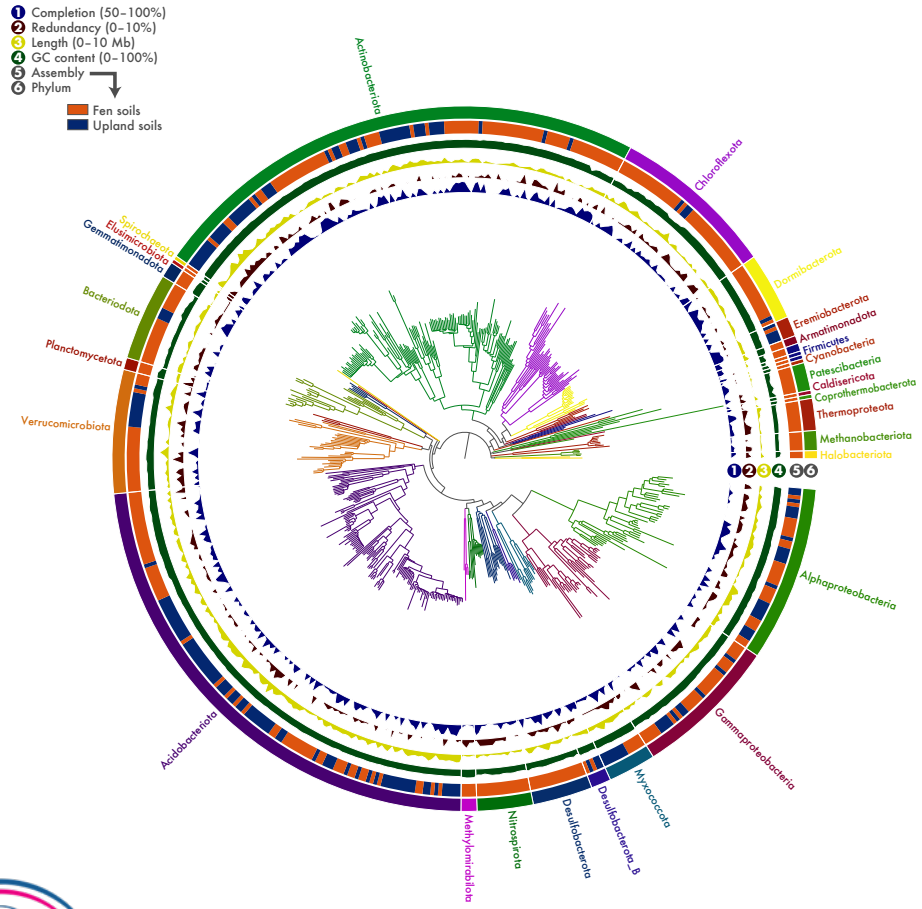
# Things you can do: abundance/distribution analyses

Examples of tools you can use for that:

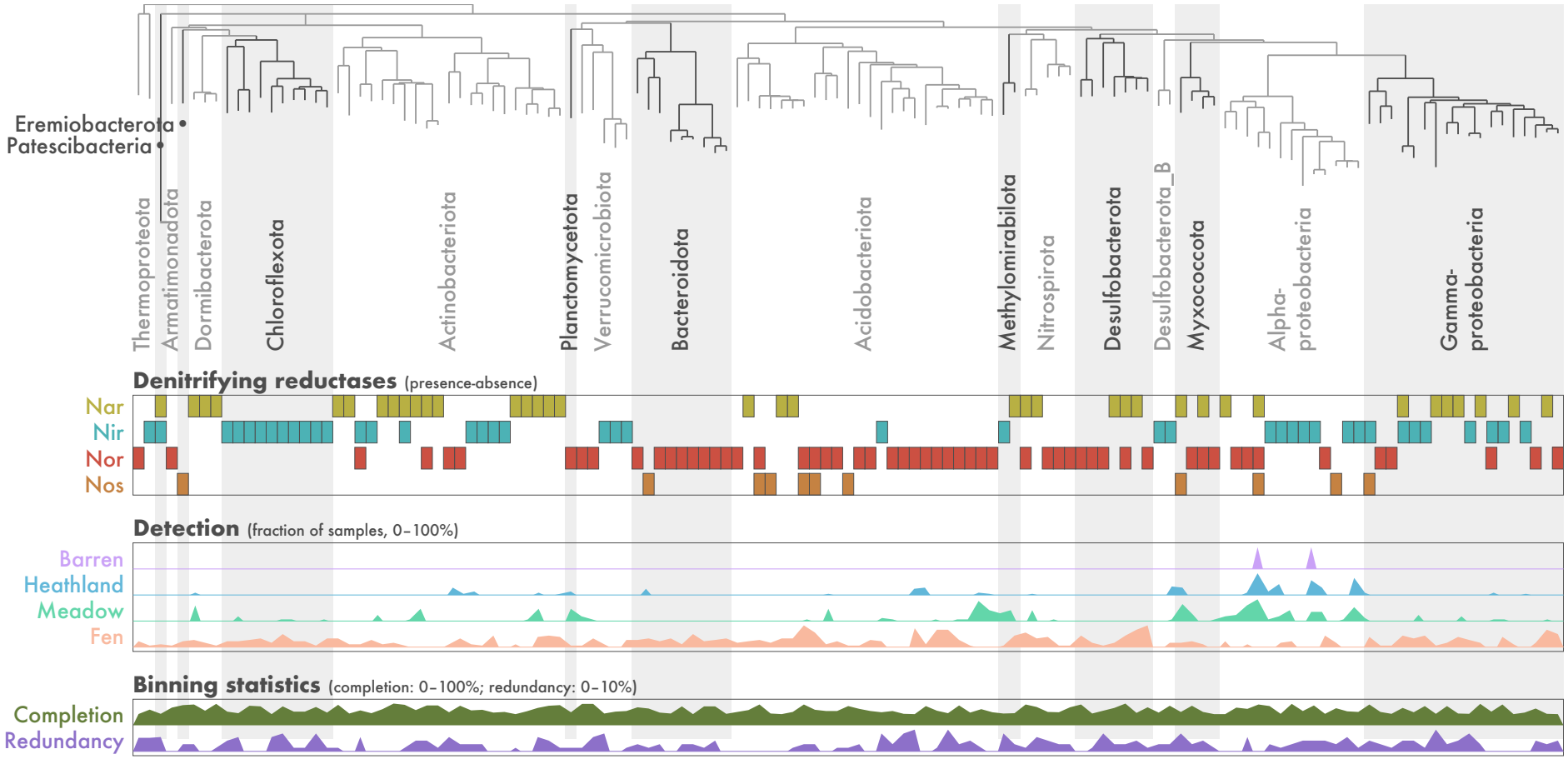
- Anvi'o: https://merenlab.org/software/anvio
- CoverM: https://github.com/wwood/CoverM

# A real-life example:

## Pessi et al., 2020: https://www.biorxiv.org/content/10.1101/2020.12.21.419267v1

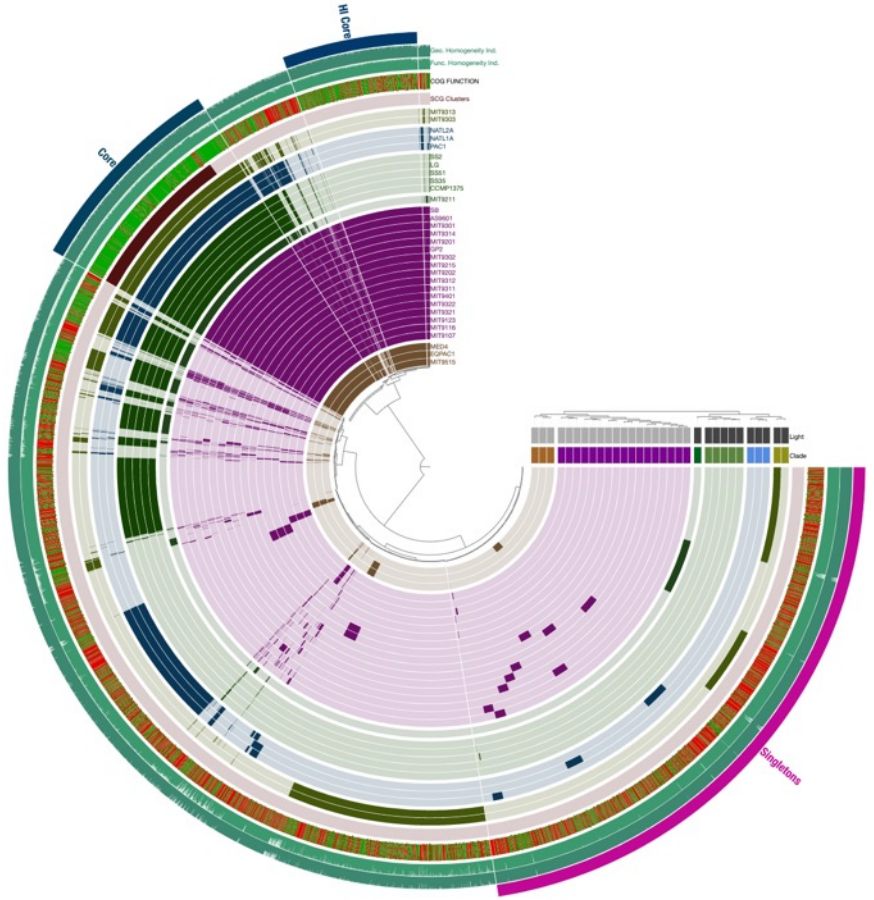



# Denitrifying communities in tundra soils are dominated by truncated denitrifiers



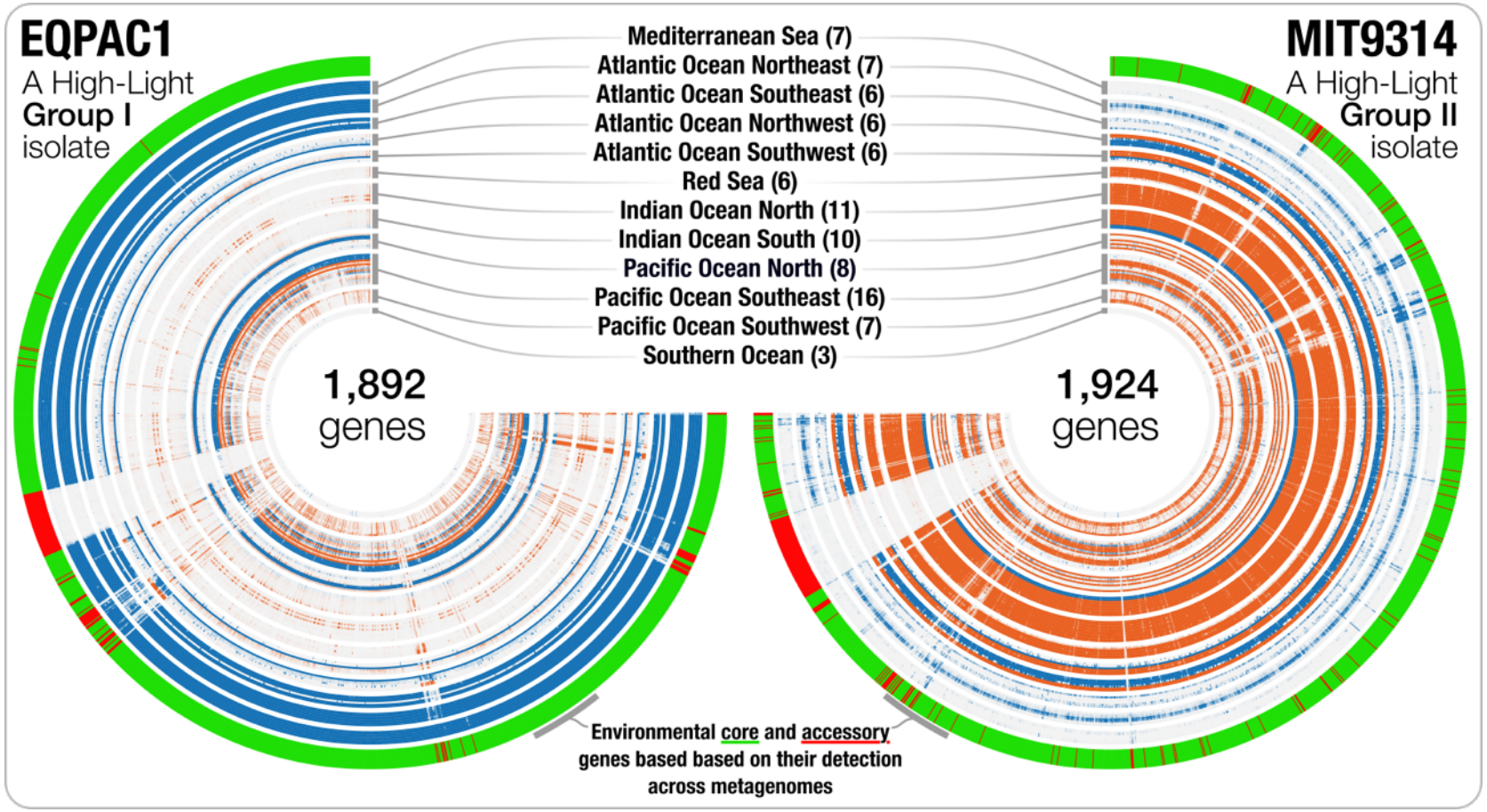
# Some other things you could do: pangenomics

merenlab.org/2016/11/08/pangenomics-v2/



# Some other things you could do: (meta)pangenomics

merenlab.org/data/prochlorococcus-metapangenome



# A note on MAG dereplication

During this week you have assembled and binned:

- Four samples with Illumina

But while you were sleeping, we were:

- Assembling and binning with Nanopore
- Assembling and binning with Nanopore + Illumina (hybrid assembly)

Taking together all these samples and assemblies, it is very likely that we have obtained the same MAG more than once

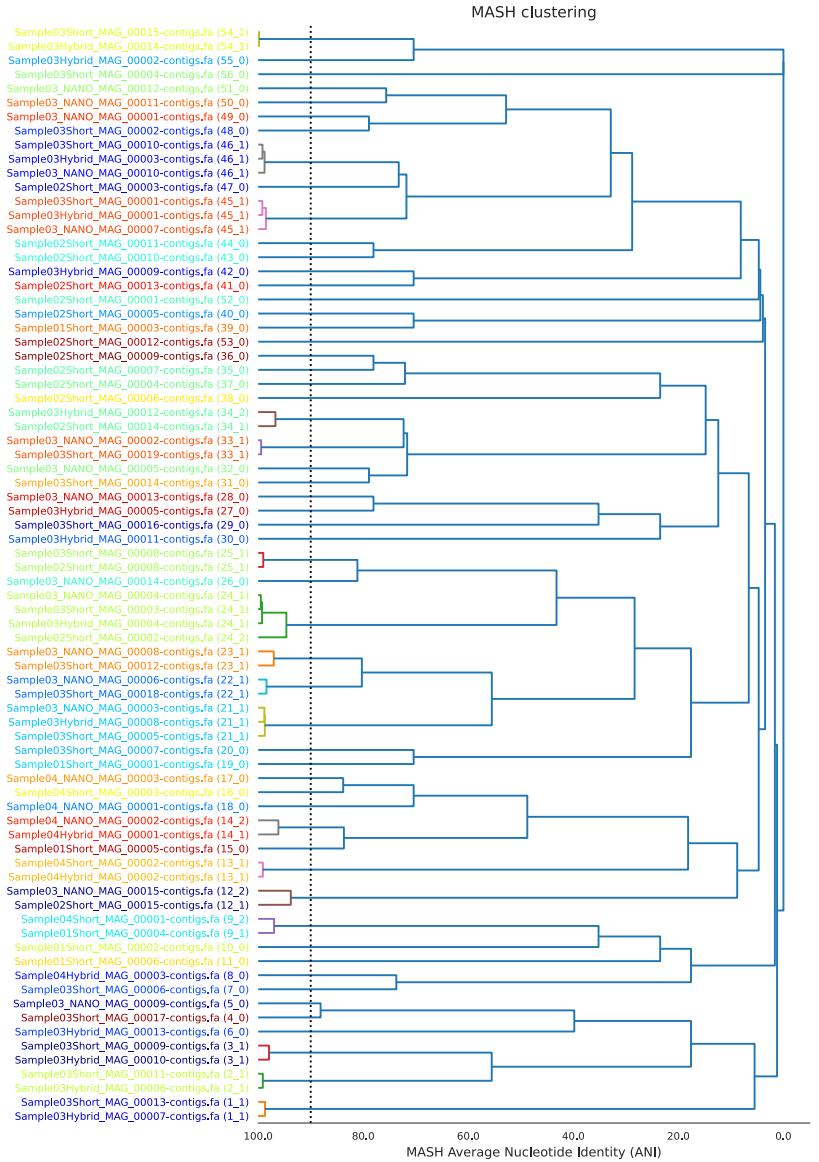
# MAG dereplication

To remove redundancy:

- i.e. copies of the same MAG

You can do that using, e.g.:

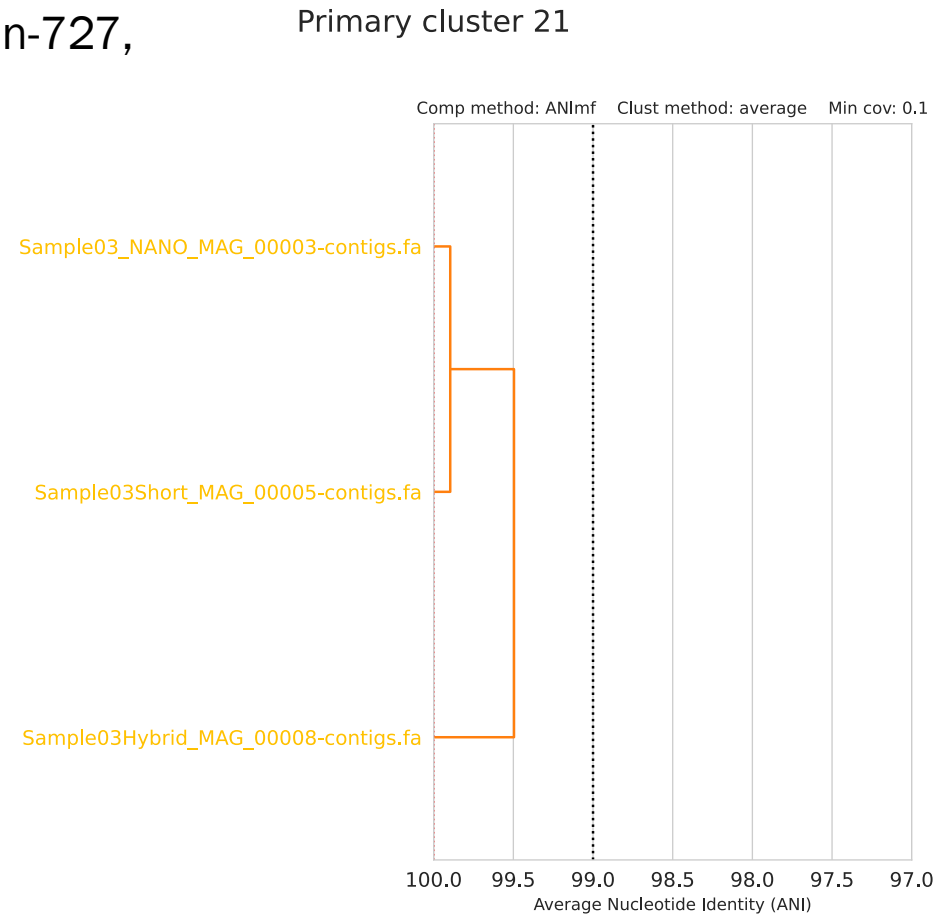
- Anvi'o
- dRep: https://drep.readthedocs.io



# MAG dereplication

d_Bacteria, p_Actinobacteriota, c_RBG-13-55-18, o_Fen-727, f_Fen-727, g_FEN-680, s_FEN-680 sp003157385

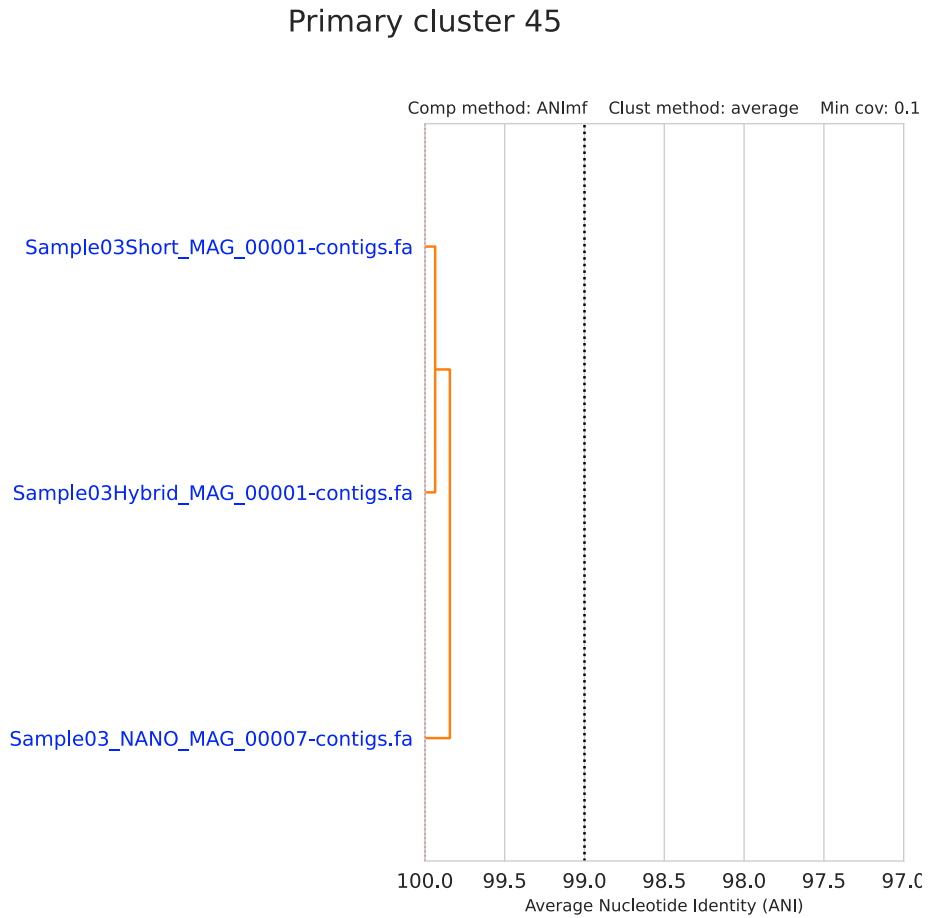
| Assembly | Completion | Redundancy | Contigs | 16S rRNA |
|---|---|---|---|---|
| Illumina | 93.0 % | 0.0 % | 62 | 1142 bp |
| Nanopore | 94.4 % | 1.4 % | 10 | 1527 bp |
| Hybrid | 87.3 % | 0.0 % | 44 | Nope |



# MAG dereplication

No taxonomic assignment (for now)

| Assembly | Completion | Redundancy | Contigs | 16S rRNA |
|---|---|---|---|---|
| Illumina | 100.0 % | 1.4 % | 85 | 1314 bp |
| Nanopore | 84.5 % | 0.0 % | 45 | 1559 bp |
| Hybrid | 100.0 % | 1.4 % | 53 | Nope |



# MAG dereplication

d_Bacteria, p_Chloroflexota, c_Ellin6529, o_CSP1-4, f_CSP1-4, g_Fen-1039

| Assembly | Completion | Redundancy | Contigs | 16S rRNA |
|---|---|---|---|---|
| Illumina | 81.7 % | 7.0 % | 72 | 904 bp |
| Nanopore | 95.8 % | 1.4 % | 1 | 1484 bp |