

Introduction to the

Ocean Microbiomics Database (OMD v2)

and the reference database for

metagenomic Operational Taxonomic Units (mOTUS v4)

Get your data:

```
cp -R /nfs/nas22/fs2202/biol_micro_teaching/551-1119-00L-2024/s111213_OMD ./
```

OMD and mOTUs resources:

- Publication OMD v1: <https://www.nature.com/articles/s41586-022-04862-3>
- Companion website (OMD v1 and v2): <https://microbiomics.io/ocean/>
- Publication mOTUs v3: <https://doi.org/10.1186/s40168-022-01410-z>
- Companion website 9mOTUs v4): <https://motus-db.org/>

What is the OMD v2?

- A compilation of ~274,000 marine genomes from ~12,000 samples, including metagenomic samples from:
 - Tara Oceans, Malaspina and Biogeotraces expeditions
 - HOT and BATS time series

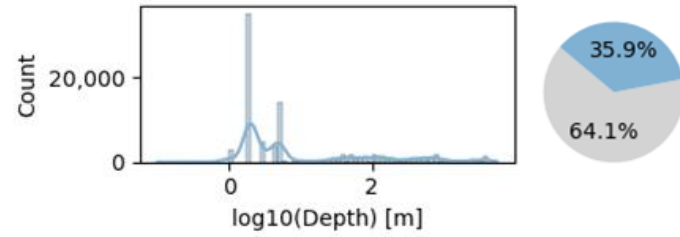
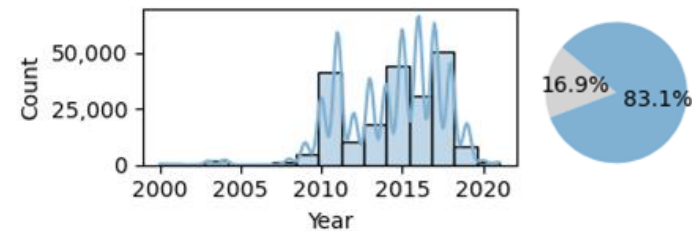
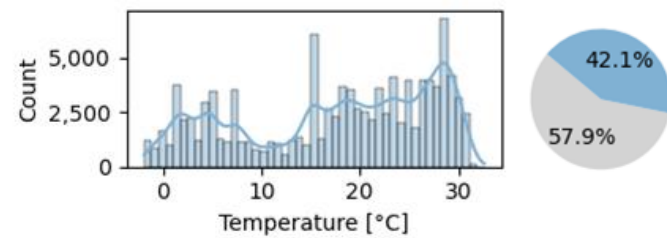
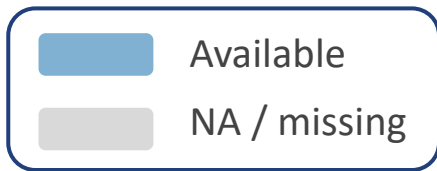
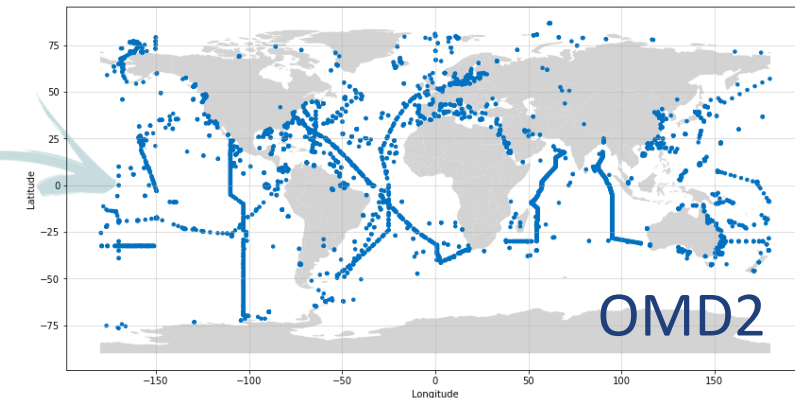
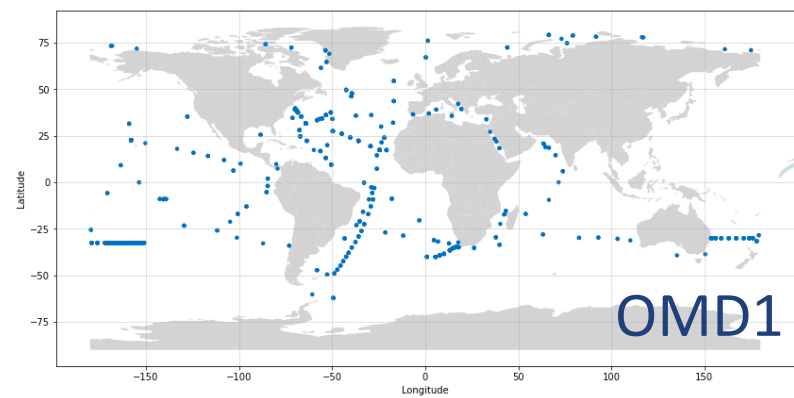
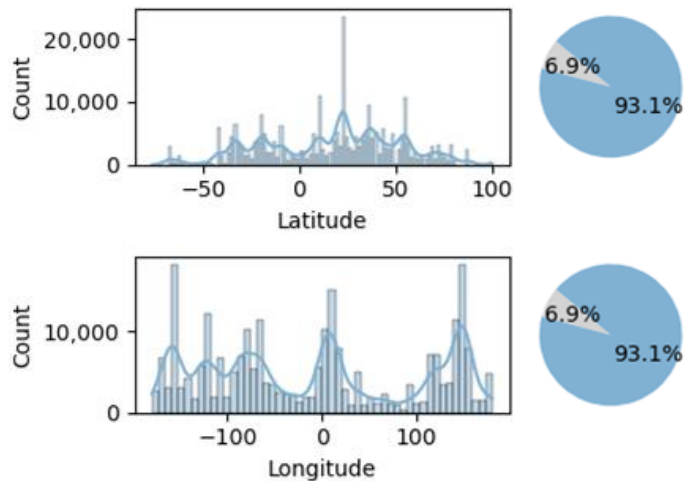
What is the mOTUS v4?

- A superset of the OMD v2, with ~3,700,000 genomes from ~118,000 samples from all biomes



Integrative global metadata exploration

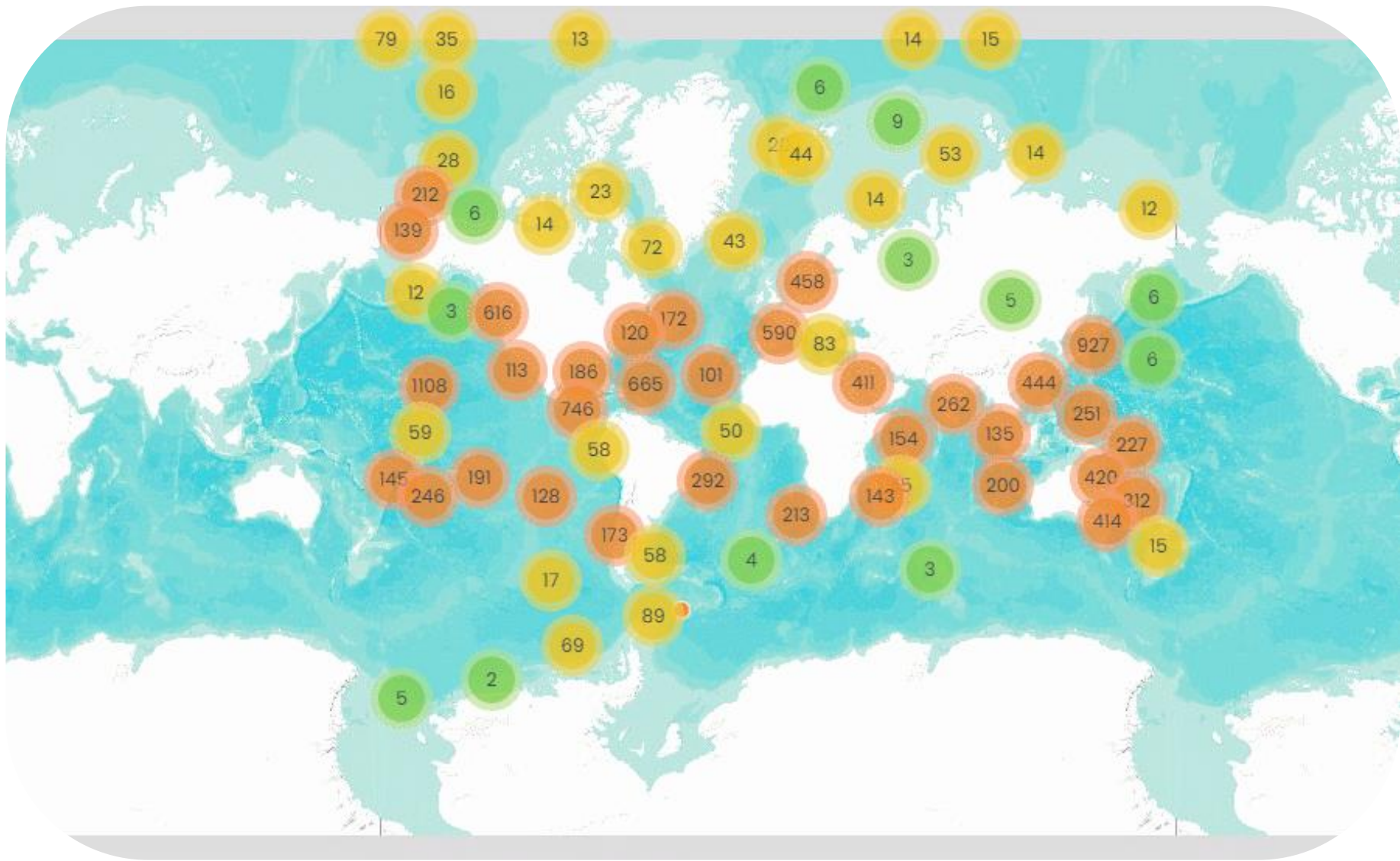
Retrieval and curation of metadata from diverse sources under constant development



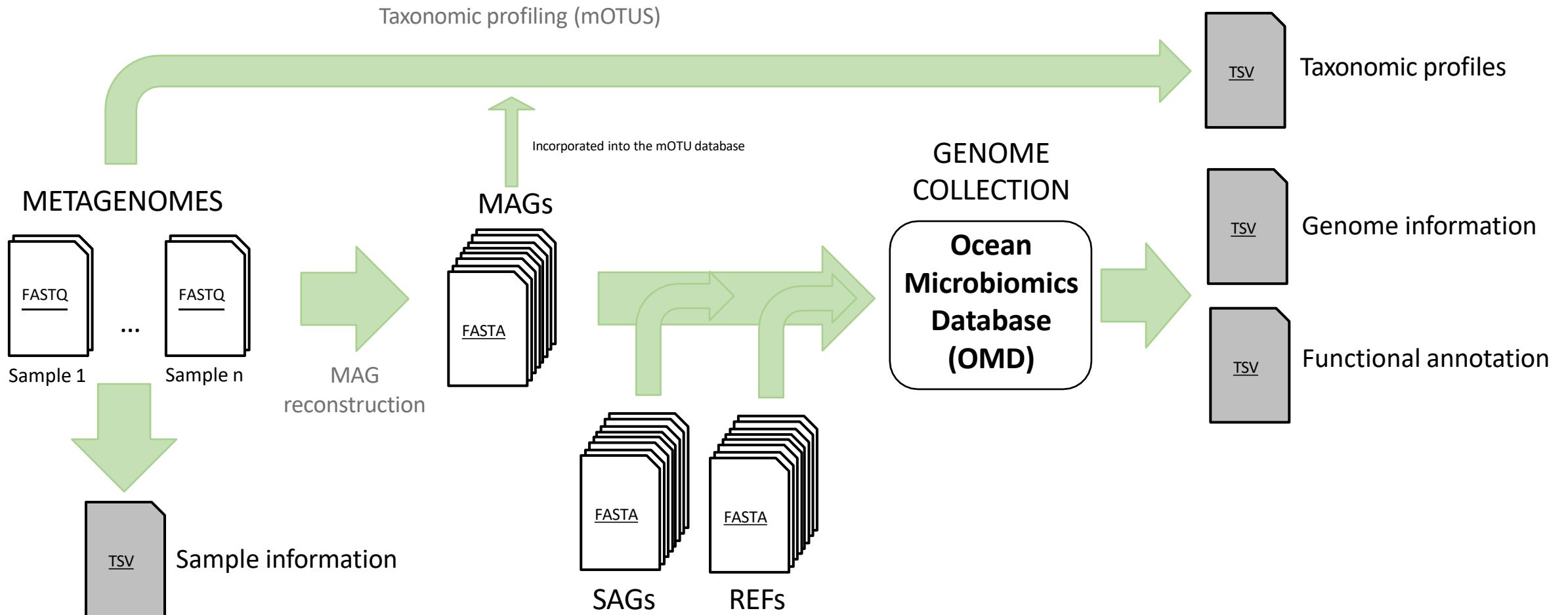
- ENA
- MarineMetagenomeDB
- Planet Microbe
- MarDB/MarRef
- BioPlatform Australian Microbiomes
- Pangaea



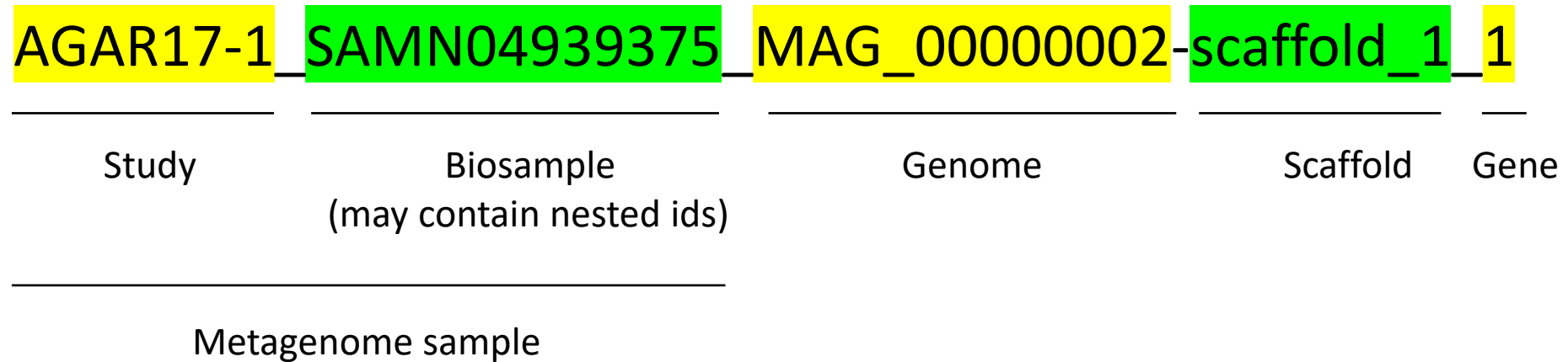
<https://microbiomics.io/ocean2>



Overview of the OMD data



The identifier



Genes in a genome

- Every scaffold contains multiple genes
- For each scaffold, the numbering of genes starts with 1

```
>AGAR17-1_SAMN04939375_MAG_00000002-scaffold_1_1 # 66 # 713 # 1 # ID=1_1;partial=00;start_type=ATG;rbs_motif=None;rbs_spacer=None;gc_cont=0.685
ATGTCGGTTGTGGTACCACCGACCGTCCGGCGGAGAACAGCGCCGCATTGGGCAAGAAGCGGATTTGAAC
TCTGGCCGGGACACCCAGCCTGCTGCGAGACGAATACGAGCACGTCGTTCCCTTGGGTAACCGTGATCGA
AGCCGACAGCCCGGTCGTTTGCCAAAGTGTGGGGTAAAATCCCGCCAACTTGCCCGTTGCAGTTCCT
GAGACAACCTGCACGATACCCGTCGGACCTGCCGGTCCCGTAGGTCCCGTCGGGCTGCGGGTCCAATCG
GTCCAGTAGATCCCGTCGGGCCTGCGGGTCCAATCGGTCCAGTAGGTCCCGCAGGACCCGCAACGCCACG
GAAGCCGCGCGGGCCGCGTGCCCCCGCGGGCCGCGTGCCCCCGCGGGCCCGTTGCACCGACATCCCCC
GCAAGCCCCTGCGGGCCGGTCTCTCCGGCAGGCCCTTGCGGACCCGTCCTCTCCGGCAGGCCCTGTGGAC
CGGTCCCTCCGGCGGGCAGGTTCTGTTAGCCTCGACCCATCGCCGACGAACGCCGTTGCGGTGACTGTCCC
CGCCAGATGCGTCTCGTCGTGCGTCGCCGACGTCGGATCCGGATCTTGCCCGTCTCGTCCGACGACCG
CCGACGTTCCCGATGTAG
>AGAR17-1_SAMN04939375_MAG_00000002-scaffold_1_2 # 1392 # 1847 # 1 # ID=1_2;partial=00;start_type=ATG;rbs_motif=None;rbs_spacer=None;gc_cont=0.680
ATGTACGTCCCGCGCTCTGCCCGTCGGCAGGGCCATCGACACCTCCGTGCGCGTCTGACTCCGATACAG
TCAACGCCGTGTTCCAAGGTGGGCTGTTGGTGCCGATGTCCCAGGTTACCCCCGTGACTACCAGCGT
CGTTGCGTCGGCGCTCACCTGGGCCGACAGGATCGCAGGCGTCGCATCCGACGACGACCCAGCCGACATT
GCCCAGACGGCAGGCACCCACAGCACCAGACCTGCCAATAGAATCCGTGTTCTTGTCCGCATCATGTACC
TCCTTGCCCTCAAGCGCGGGCGCATCATGCCACCGCCCGGAAACGTGCGCCCGATCCCTGCCGCTGGGC
GTACGGCCGCGTCCGCCGCCGCGCCTCGGCATCCAGAGCCGACCCGACCCGAGACCCGCGCATCACTC
ACCGGCGCGTCCGGACACAGTCCAGACCGCTACTGA
>AGAR17-1_SAMN04939375_MAG_00000002-scaffold_1_3 # 2063 # 2683 # -1 # ID=1_3;partial=00;start_type=GTG;rbs_motif=GGTGG;rbs_spacer=3bp;gc_cont=0.729
GTGTCGGTGGGCTGGGCGGCGGCTTCGCGCTGGCCACAGCCCTGGCGGCGCGGCGCTCTGCGAACGACG
AGATCCAGTCCACCATGGAGGCGGGCCGGGTGACTCTGGTCGCGACCGGGCGCGGTTGACGGACGTGCT
CGCGGAGTGGTCGCGCTGGGGGTACGCGCTTCGTGGACACGGAGGCGATGGCCGGACAGTCGGTCCAG
CTGCACGTCGTCGACGTCGCCGAGTCTGAAGCGTTGCAGGTTCTGCTGCGCCCCGGCGGTGCGCTACGTTG
CGGCTCCGCGCCGCCGCGGCTCGACGGGCGCTTCGCGTTACGATCGCGTCAAGATCCTGGGCACAGGCCG
CCTTGCCGCTGCGACGACCGGTACGGGCGGGCGCATCTCGTGGCGACGAACCGGCTGGCGCGCGGAGG
CGGGCGGGGGCCGATGCCGCTTGAGGACATGCAGCGGCTCCTGGACGCGGTCTCTGGTGCCTCGGGCG
GCGTGACGGCCGCCACCACGACCCCGGTGCAGCCAGCAACGCCAATCCGGTGACGACGTGGGGAT
GGTCCGTCCAGCGCCGACACACCGTTTCCGGGGATGGTGGTCAACCCGGCGTGCCTGA
```


Functional annotations -- BGCs

antiSMASH (biosynthetic gene clusters – BGCs)

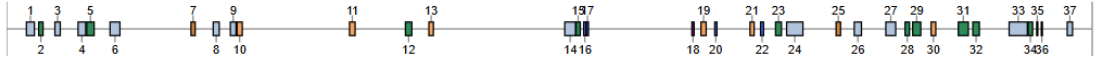
antiSMASH version 7.1.0 Download About Help Contact

Select genomic region:

Overview 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 1.10 1.11 1.12 1.13 1.14 1.15 1.16 1.17 1.18 1.19 1.20 1.21 1.22 1.23
1.24 1.25 1.26 1.27 1.28 1.29 1.30 1.31 1.32 1.33 1.34 1.35 1.36 1.37

Identified secondary metabolite regions using strictness 'relaxed'

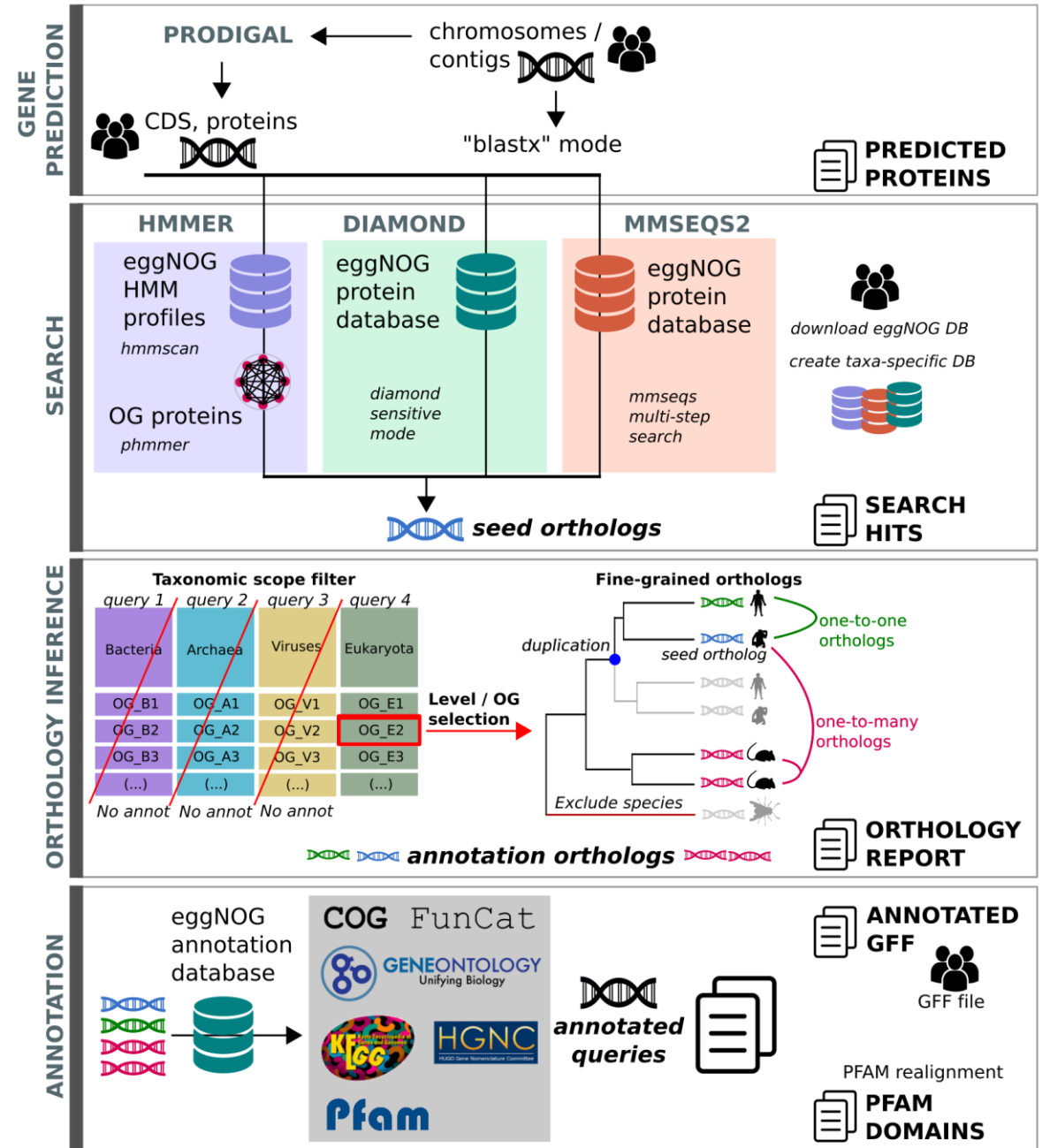
contig_1



Region	Type	From	To	Most similar known cluster	Similarity
Region 1	NRPS ↗ , T1PKS ↗	195,659	274,544		
Region 2	NRPS ↗	316,334	359,486		
Region 3	NRPS-like ↗ , T1PKS ↗	476,432	529,335		
Region 4	NRPS ↗ , T1PKS ↗ , thiopeptide ↗	704,620	779,795		
Region 5	NRPS ↗	789,717	863,165		
Region 6	NRPS ↗ , T1PKS ↗	1,019,945	1,117,239	crochelin A ↗	NRP+Polyketide 8%
Region 7	arylpolyene ↗	1,817,597	1,858,823	APE Vf ↗	Other 15%
Region 8	NRPS ↗ , T1PKS ↗	2,036,622	2,096,062		
Region 9	T1PKS ↗ , NRPS ↗	2,205,652	2,261,565		
Region 10	hglE-KS ↗ , T1PKS ↗	2,271,356	2,322,927		
Region 11	hglE-KS ↗	3,379,882	3,434,972		
Region 12	NRPS ↗	3,931,122	3,993,968		
Region 13	T1PKS ↗	4,162,290	4,206,804		
Region 14	transAT-PKS ↗ , NRPS ↗	5,501,206	5,607,834	sorangicin A ↗	Polyketide:Trans-AT type I polyketide 8%
Region 15	NRPS-like ↗	5,612,181	5,654,343	1-nonadecene/(14Z)-1,14-nonadecadiene ↗	Polyketide:Modular type I polyketide 100%
Region 16	lassopeptide ↗ , RRE-containing ↗	5,687,532	5,711,466		
Region 17	RRE-containing ↗	5,714,782	5,735,078		

Functional annotations eggNOG mapper

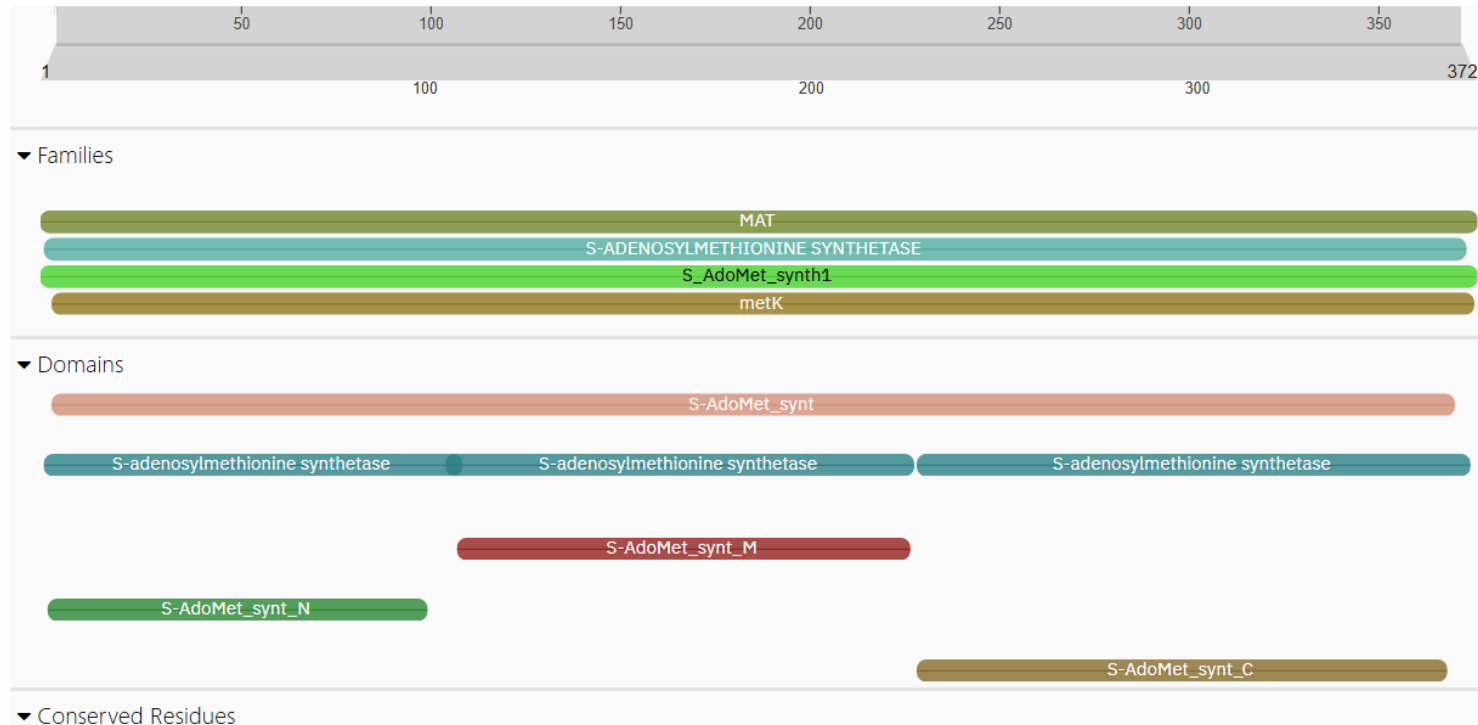
Can produce PFAMs, KEGG, CAZy



Functional annotations PFAMs – Protein domains

InterPro -- host of PFAM DB
<https://www.ebi.ac.uk/interpro/search/sequence/>

Entry matches to this protein¹



InterPro
Classification of protein families

Search / Sequence

Scan your sequences

>P081MF_15550 methionine_adenosyltransferase
MTHLFSSEVTEGRPKISEQLDVALEALGDFRGRVACETFTTTLIVVSGEITTSQGLDIANKVWRD1GYTDSAMSFACVCAWVYELKGGPDIAMHVDVDSAGDQGLMFGYACRETELMPFFIDLA

S-AdoMet_synthetase - IPR002133
PIRSF: MAT - PIRSF000497
PANTHER: S-ADENOSYLMETHIONINE SYNTHETAS
HAMAP: S-AdoMet_synth1 - MF_00086
NCBIFAM: metK - TIGR01034

Representative domains

- S-AdoMet_synthetase_sfam** - IPR022636
SSF: S-adenosylmethionine synthetase - SSF55973
- S-AdoMet_synt_central** - IPR022629
PFAM: S-AdoMet_synt_M - PF02772
- S-AdoMet_synt_N** - IPR022628
PFAM: S-AdoMet_synt_N - PF00438
- S-AdoMet_synt_C** - IPR022630
PFAM: S-AdoMet_synt_C - PF02773

PFAMs

Functional annotations

KEGG KOs – Metabolic genes

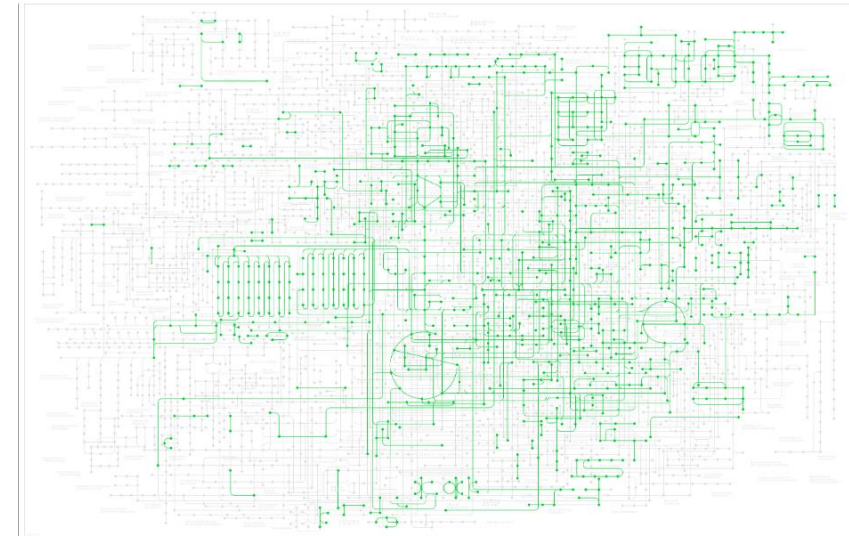
<https://www.kegg.jp/>



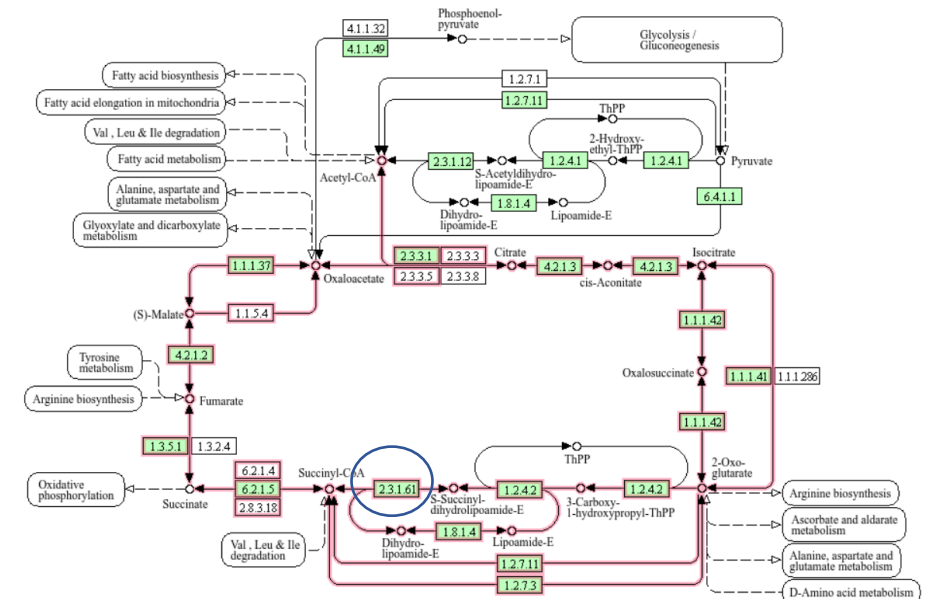
Sulfidibacter corallicola: J3U87_01635

[Help](#)

Entry	J3U87_01635	CDS	T07864
Symbol	odhB		
Name	(GenBank) 2-oxoglutarate dehydrogenase complex dihydrolipoyllysine-residue succinyltransferase		
KO	K00658 2-oxoglutarate dehydrogenase E2 component (dihydrolipoamide succinyltransferase) [EC:2.3.1.61]		
Organism	scor Sulfidibacter corallicola		
Pathway	<p>scor00020 Citrate cycle (TCA cycle)</p> <p>scor00310 Lysine degradation</p> <p>scor00380 Tryptophan metabolism</p> <p>scor00785 Lipoic acid metabolism</p> <p>scor01100 Metabolic pathways</p> <p>scor01110 Biosynthesis of secondary metabolites</p> <p>scor01120 Microbial metabolism in diverse environments</p> <p>scor01200 Carbon metabolism</p> <p>scor01210 2-Oxocarboxylic acid metabolism</p>		
Module	<p>scor_M00009 Citrate cycle (TCA cycle, Krebs cycle)</p> <p>scor_M00011 Citrate cycle, second carbon oxidation, 2-oxoglutarate => oxaloacetate</p>		
Brite	<p>KEGG Orthology (KO) [BR:scor00001]</p> <p>09100 Metabolism</p> <p>09101 Carbohydrate metabolism</p> <p>00020 Citrate cycle (TCA cycle)</p> <p>J3U87_01635 (odhB)</p> <p>09105 Amino acid metabolism</p> <p>00310 Lysine degradation</p> <p>J3U87_01635 (odhB)</p>		



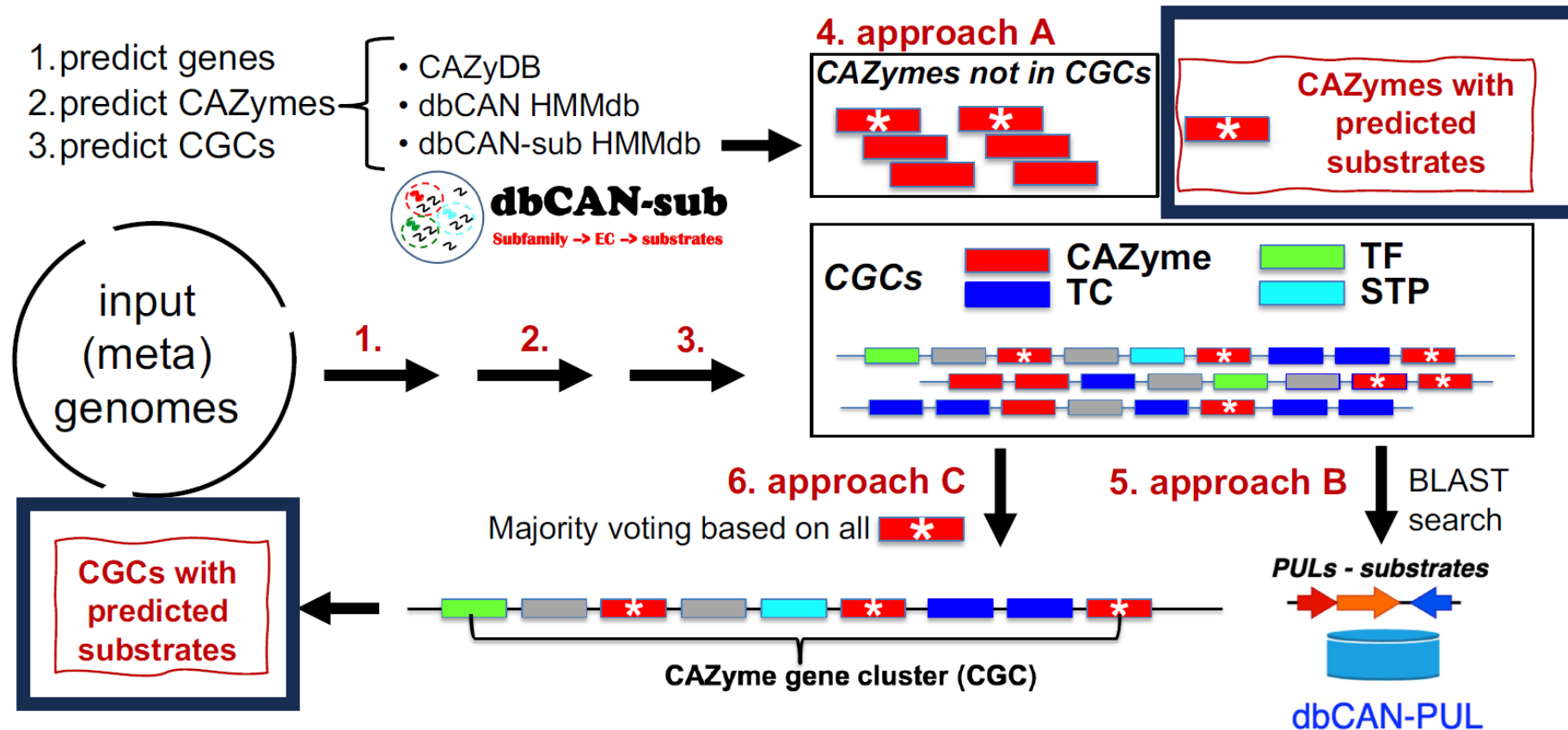
CITRATE CYCLE (TCA CYCLE)



Functional annotations dbcan – Complex sugar degradation

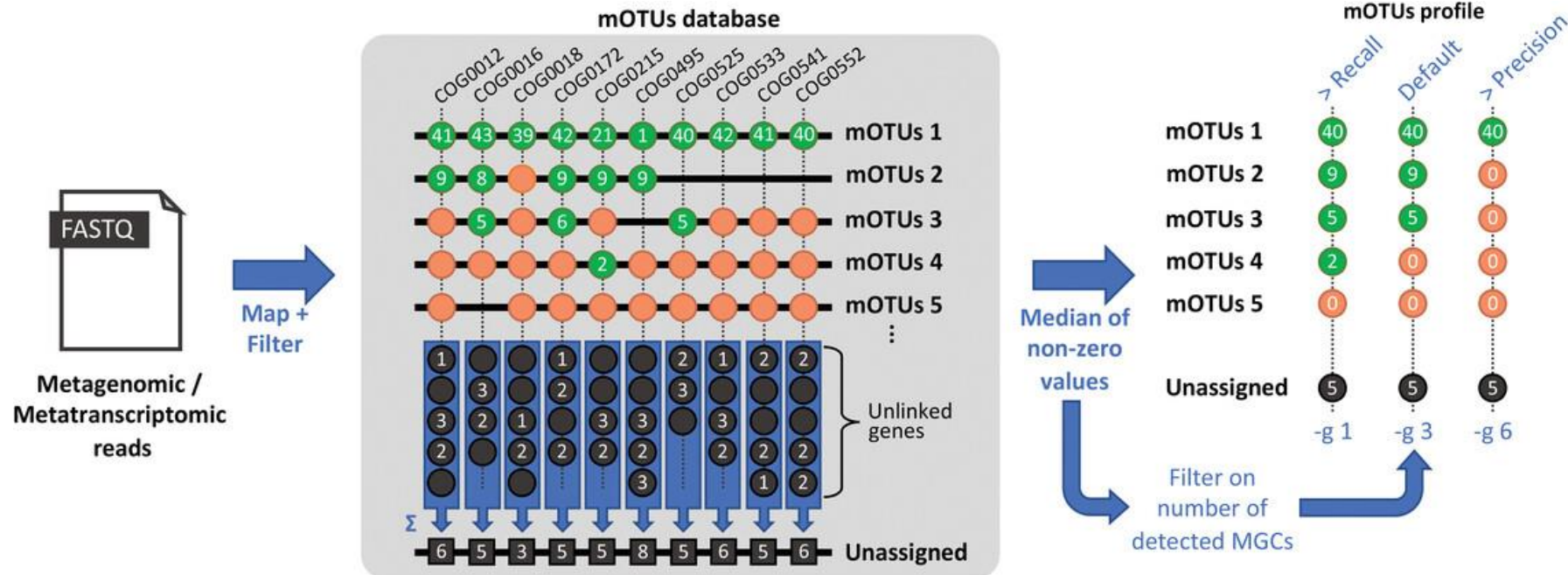
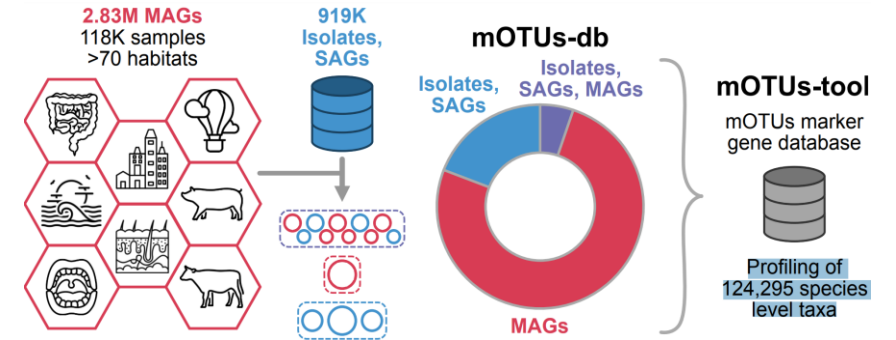
<https://bcb.unl.edu/dbCAN2/>

Can produce CAZy, CAZyme gene clusters (CGCs) and predicts substrates



Functional annotations

Taxonomic profiling with mOTUs



Data exploration exercises

General recommendations:

- You can try to answer the questions by any means: making plots, tables, numerical summaries, etc.
- The questions don't have a yes/no answer. The goal is to get familiar with the data and practicing with python
- Apart from producing results spend some time exploring them and understanding the meaning.
- If you don't know how to compute some step – google! (or ask us)
- Be organized and create a script performing the entire task. Annotate your script so it can be understood later.

Presentation

- At the end of the day, everyone will present 3 results of their data exploration exercises, and the approach to get there (we will open a Zoom and you can share your desktop)

1) Explore samples of OMDv2 database

Table:

Sample metadata: [OMDv2.sample.meta.tsv.gz](#)

Exercises:

- How many samples are in the OMDv2?
- How many samples were collected from each study? (column: dataset)
- Familiarize with the origin of the samples (columns: 11 – 15, and 24)
- Make a world map with the location of all samples
- Differentiate studies in the map in some way (color, shape, etc.)
- Are the different studies covering different ranges of temperature and depth?

2) Explore genomes of OMDv2 database

Table:

Genomes: OMDv2.genome.meta.tsv.gz

mOTU abundance per sample: OMDv2.taxa.abundance.tsv.gz (LARGE!!)

- How many genomes?
- From which study are most of the genomes?
- How many genomes are from seawater, and how many from corals?
- How many genomes are associated with an mOTU? (column: MOTU4)
- How many different mOTUs are represented?
- Plot the distribution of genomes per mOTU
- What is the distribution of QSCORE, COMPLETENESS and CONTAMINATION for the genome collection?
- Plot the GC content colored by genus
- Are the values different depending on the genome type? (column: IS_MAG)?
- Which bacterial Phylum is the most abundant in the open ocean seawater?

3) Explore functional annotations, using a subset of ~7k Acidobacteriota (subset from mOTUs v4)

Tables:

- Genomes (~7k Acidos): [acidus.motus4.taxa.tsv.gz](#)
- Environment (for whole mOTUs DB): [motus4.environments.tsv.gz](#)
- BGC counts (~7k Acidos): [acidus.motus4.antiSMASH.count.tsv.gz](#)
- CGC counts (~7k Acidos): [acidus.motus4.dbcan.CGC.counts.tsv.gz](#)
- CAZy substrates (~7k Acidos): [acidus.motus4.dbcan.CAZy.substrate.tsv.gz](#)
- PFAM annotations (for a single Acidobacteriota): [acidus.motus4.genomes.eggnoG.emapper.annotations_PFAMs_filter.tsv.gz](#)
- PFAM map (global): [Pfam-A.hmm.tsv.gz](#)

Exercises:

- Which Acidobacteriota family has the most BGCs (Biosynthetic Gene Clusters) per genome?
- Which Acidobacteriota has the most CGCs (CAZyme gene clusters) per genome, and what are potential substrates?
- Is there a correlation between BGCs and CGCs? You can group by taxonomic unit (e.g. family or genus level), or environment.
- What are the PFAMs that occur in the direct neighborhood of CAZys with predicted substrates (e.g. distance of 1 to 5 genes, left and right of a CAZy).

If you have an idea for another exercise, go for it!