Introduction to the Ocean Microbiomics Database (OMD)

OMD Resources

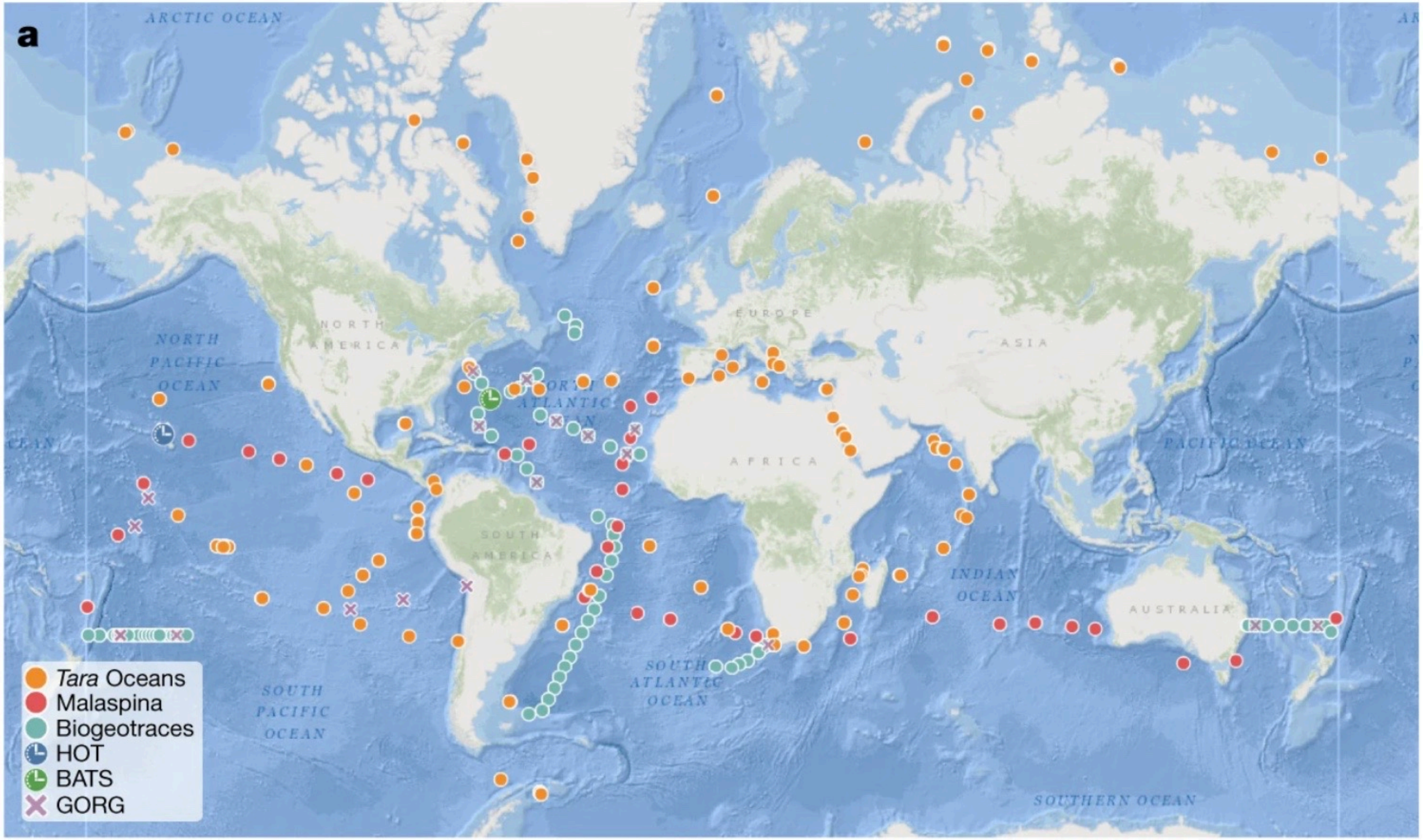
- Publication: https://doi.org/10.1038/s41586-022-04862-3
- Companion website: https://microbiomics.io/ocean/
- Data location in the servers: */nfs/nas22/fs2202/biol_micro_teaching/551-1119-00L/masterdata*

What is the OMD?

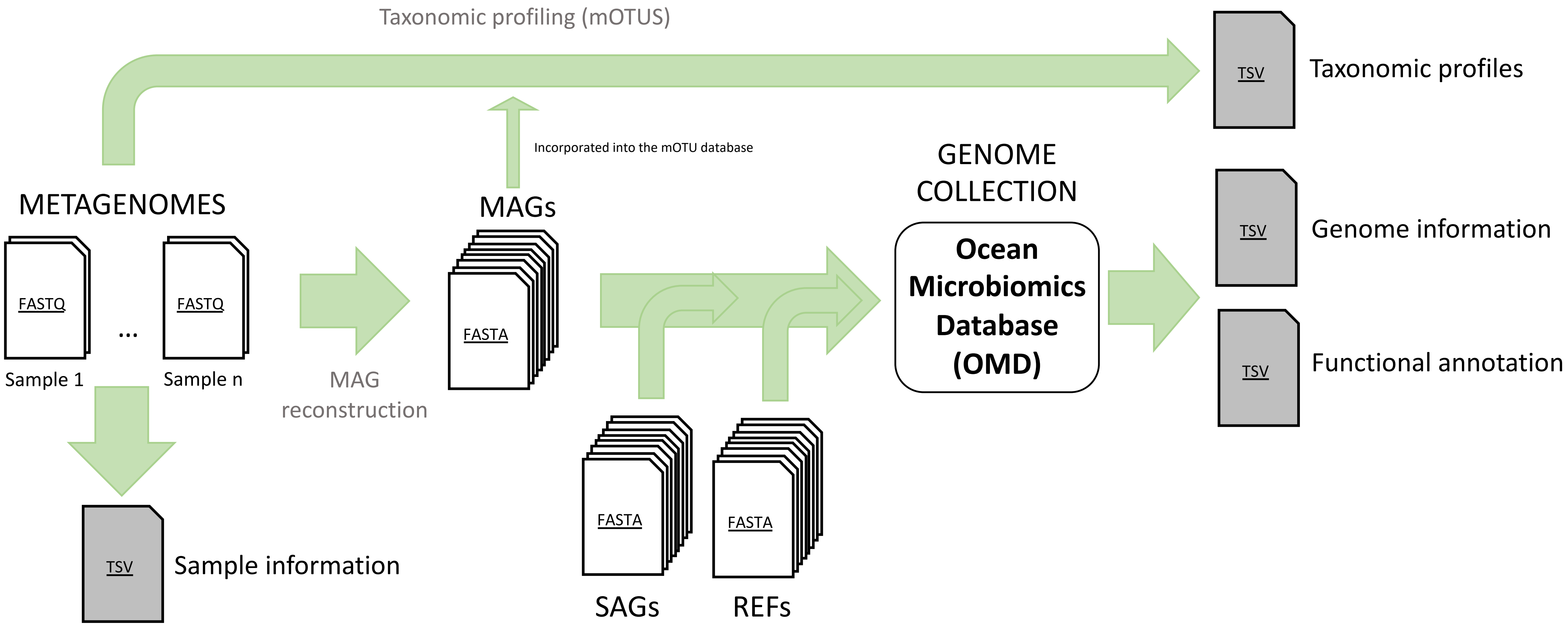
- A compilation of ~35,000 marine genome:
  - ~27,000 metagenome assembled genomes (MAGs)
  - ~5,900 single amplified genomes (SAGs)
  - ~1,700 reference genomes (REFs)

- MAGs were reconstructed from a set of ~1,000 metagenomic samples:
  - Tara Oceans, Malaspina and Biogeotraces expeditions
  - HOT and BATS time series

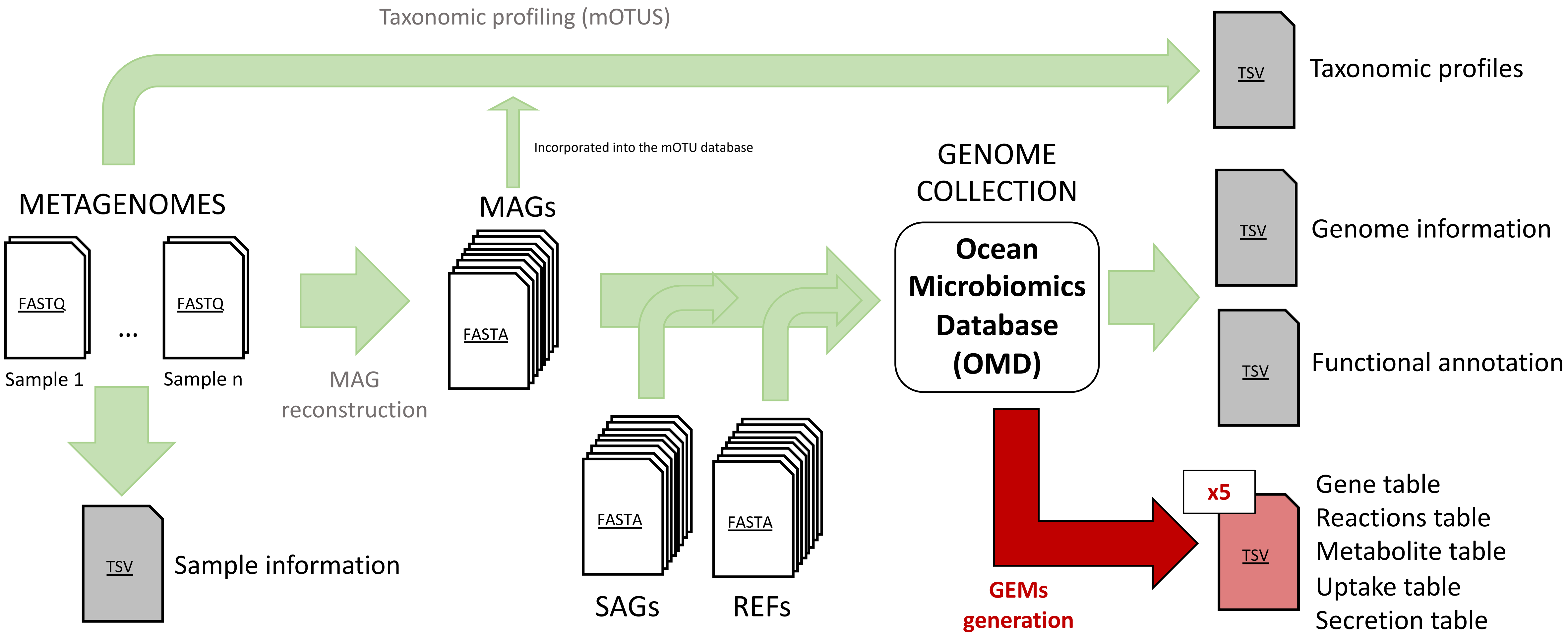
What is the OMD?



Overview of the OMD data



Overview of the OMD data



Available data for projects

TSV

Sample information → sample_info_marine_v31.tsv

- Sample origin (study, coordinates, depth, etc.)
- Environmental variables (temperature, chlorophyll, nutrient conc., etc.)

TSV

Taxonomic profiles → motus_profiles_marine_v31.tsv

- Abundance of each mOTU in each sample

TSV

Functional annotation → genome_kegg_marine_v31.tsv.gz

- KEGG annotation (KO presence per genome)

TSV

Genome information → genome_info_marine_v31.tsv

- Genome-mOTU link
- Source
- Quality stats
- Features (GC%, Genome size, etc.)

Available data for projects

TSV

Sample information → sample_info_marine_v31.tsv

- Sample origin (study, coordinates, depth, etc.)
- Environmental variables (temperature, chlorophyll, nutrient conc., etc.)

TSV

Taxonomic profiles → motus_profiles_marine_v31.tsv

- Abundance of each mOTU in each sample

TSV

Functional annotation → genome_kegg_marine_v31.tsv.gz

- KEGG annotation (KO presence per genome)

TSV

Genome information → genome_info_marine_v31.tsv

- Genome-mOTU link
- Source
- Quality stats
- Features (GC%, Genome size, etc.)

~2k xml

GEM models → GEMs/*.xml

- GEM models in the form of xml files

TSV

Gene table → genes_reps.tsv

- Presence/absence of each gene in each genome

TSV

Reaction table → reactions_reps.tsv

- Presence/absence of each reaction in each genome
- You'll need **bigg_models_reactions.txt** from BiGG

TSV

Metabolite table → metabolites_reps.tsv

- Presence/absence of each metabolite in each genome
- You'll need **bigg_models_metabolites.txt** from BiGG

TSV

Uptake table → uptake_reps.tsv

- Presence/absence of each compound in each genome

TSV

Secretion table → secretion_reps.tsv

- Presence/absence of each compound in each genome

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 R.
- Apart from producing results spend some time exploring them and understanding the meaning.
- If you don't know how to compute some step google!; if it does not work ask us!

- Be organized and create a script performing the entire task. Annotate your script so it can be understood later.
- Use the `fread()` (from the data.table package) to load the files.

Data exploration exercise

Task 1: Summarize number of genomes:

1. How many genomes do we have available?
2. How many from each type (SAGs, REFs & MAGs)?
3. How many genomes can we associate with a mOTU?
4. How many mOTUs do we have? What proportion of those correspond to mOTUs for which we have at least a genome?
5. What is the distribution of genomes per mOTU?

Data exploration exercise

Task 2: Explore genome quality statistics:

1. What is the distribution of completeness and contamination for the genome collection?
2. Is it different depending on the genome type (SAGs, MAGs, REFs)?
3. How comparable are the values computed with CheckM and ANVIO?
4. Is there any relationship between completeness and genome size? Is it the same for all three types of genomes? What may it mean?

Data exploration exercise

Task 3: Explore the sample information:

1. How many samples do we have from each study?
2. Make a world map with the location of all samples. Differentiate studies in the map in some way (color, shape, etc.) → Will probably need to google how to make a map with ggplot
3. Are the different studies covering different depths?
4. Are the different studies covering different ranges of temperature?

Data exploration exercise

Task 4: Explore the GEM information:

1. What is the average number of reactions per genome? What is the minimum and maximum? And the average/minimum/maximum number of compounds?
2. What is the organism capable of growing with the simplest environment? And the genome the most complex secretion?
3. List all genomes for which GEM predicts that nitrification is taking place
    1. You'll have to find which is the reaction responsible for nitrification. Something as basic as Wikipedia might help: https://en.wikipedia.org/wiki/Nitrification
    2. You'll have to find the corresponding reaction code in the BiGG database: http://bigg.ucsd.edu