Exploring Online Resources and Repositories (II)

Miguel Semedo

2024/09/05

Funding

Sponsorship

Support

Online Resources and Repositories

MG-RAST

- Upload **raw and assembled reads** and perform taxonomic and functional annotation

- **Download stored data** (512,000 metagenomes)

- Explore stored data (find certain genes and/or functions)

- Export analysis results

- *Meyer et al., 2008, BMC Bioinformatics*

- Website link: https://www.mg-rast.org

IMG/MER

- Upload **assembled reads** and perform taxonomic and functional annotation

- **Download stored data** (24,000 metagenomes + 25,000 genomes, etc.)

- Explore stored data (find certain genes and/or functions)

- Export analysis results

- Chen et al., 2019, Nucleic Acids Res

- Website link: https://img.jgi.doe.gov/cgi-bin/mer/main.cgi

- Available workshops (next: April 28th – May 2nd, 2025)

KBase

- Upload **any metagenome input (reads, contigs, bins)** and perform taxonomic and functional annotation

- **Highly customized pipeline.** Most current tools available

- **Share "narratives"**

- Explore stored data (find certain genes and/or functions)

- Export analysis results

- *Arkin et al., 2018, Nat Biotechnol*

- Website link: https://www.kbase.us

2

Online Resources and Repositories

MGnify

- Upload **raw and assembled reads (DNA, RNA, proteins)** and perform taxonomic and functional annotation

- Metabarcoding and metagenomics data

- All bioinformatic tools available, independently

- Download stored data – <u>linked to ENA</u> as well

- Explore stored data (find certain genes and/or functions)

- Export analysis results

- *Richardson et al., 2023, Nucleic Acids Res*

- Website link: https://www.ebi.ac.uk/metagenomics

ENA
European Nucleotide Archive

- Upload and make public **raw, assembled, or functionally annotated reads (DNA, RNA). Not used to perform analysis.**

- Metabarcoding and metagenomics data

- Explore and download stored data **using their metadata**.

- Website link: https://www.ebi.ac.uk/ena/

All of these are free, but require registration (sometimes with different access levels). Check webpage documentation.

Online Resources and Repositories

Sample metadata is essential, but can be very difficult to acess..

- Querying public metagenome databases for **raw shotgun sequencing data** from **deep-sea marine sediments**, 277 metagenome entries (samples) were obtained: 79 from **MG-RAST**, 198 from **ENA**, and 0 from **Mgnify**

- **Available raw reads for download:** 20 MG-RAST entries and 46 ENA entries (66 metagenomes in total)

- **After our metagenomics full workflow**: 81 MQ or HQ MAGs found – **still extremely worth it** the hard work

Semedo, Valquaresma, et al (*in prep*)
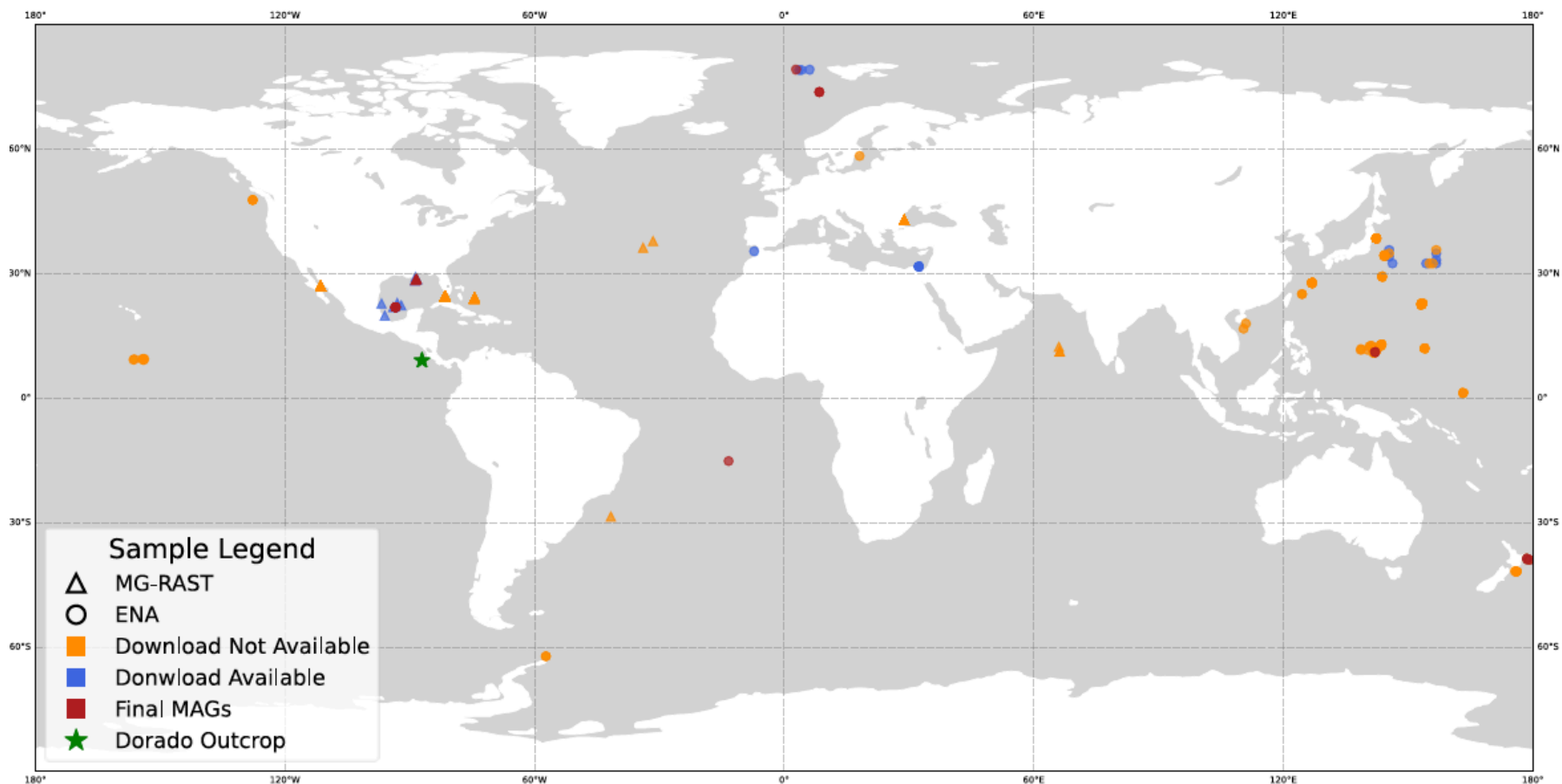
Online Resources and Repositories

Zone	Habitat	Sample ID	Class	Genus	CMPL %	CONT %	Nitrogen Cycle Genes												
							narG	napA	narB	nasA	nrfA	nirK	nirS	norB	nosZ	nifH	pmoA	hao	hzs
Bathypelagic (1000m – 4000m)	Cold Seep	SAMEA5663119	Methanosarcinia	UBA204	72	0.66		■											
			Spirochaetia	JAHOYX01	88	4.85		■											
			UBA6098	Unclassified	84	2.20													
	Continental Slope and Abyssal Plain	mgm4510162.3	Alphaproteobacteria	Amylibacter	52	0.86	■	■		■	■								
		mgm4510164.3	Alphaproteobacteria	Amylibacter	78	1.68	■	■		■									
				Sneathiella	55	5.33													
			Gammaproteobacteria	50-400-T64	61	8.10													
		mgm4510168.3	Gammaproteobacteria	Agaribacterium	93	3.79			■										
		mgm4510171.3	Gammaproteobacteria	Colwellia	56	8.62		■							■		■		
				UBA3067	87	9.99		■					■		■		■		
		mgm4510172.3	UBA6911	Unclassified	69	3.04	■												■
	Hydrothermal Vent	SAMN03002195	Alphaproteobacteria	Aurantimonas	96	1.61		■					■		■				■
			Bacteroidia	SM23-62	66	2.15				■									
			Gammaproteobacteria	GCA-002733105	91	0.42	■	■		■									■
				Halomonas	98	2.01		■		■									
				Halopseudomonas	99	1.06		■		■									
				Marinobacter	98	0.45				■									
				Methylophaga	79	8.62				■									■
				Pseudohongiella	97	0.06				■									
		SAMN03002196	Alphaproteobacteria	Aurantimonas	100	7.59		■					■		■				■
			Aminicenantia	SOIV01	78	4.84	■												
			Aquicultoria	Unclassified	63	1.81	■												
			Gammaproteobacteria	Halomonas	100	2.44	■			■									
				Halopseudomonas	94	0.73				■									
				Pseudoalteromonas	98	6.55				■									
				Pseudohongiella	71	0.06				■									
			Phycisphaerae	HyVt-337	99	5.88	■								■				
	Subduction Zone and Trenches	SAMN19461850	Anaerolineae	Unclassified	92	4.55	■								■				
		SAMN19461852	Aerophobia	SOJT01	75	1.12	■												
		SAMN19461860	Actinomycetia	Mycobacterium	55	0.63	■												
			Anaerolineae	E44-bin32	85	5.61	■												
			Desulfobacteria	Unclassified	82	0.97													
Abyssopelagic (4000m – 6000m)	Hydrothermal Vent	SAMN05571518	Alphaproteobacteria	Unclassified	67	1.03													
			Bipolaricaulia	DRJF01	83	0.00	■												■
			Gammaproteobacteria	HyVt-429	86	0.76											■	■	
			SZUA-79	QIJE01	87	0.86									■	■			
			Thermodesulfovibrionia	BMS3Bbin07	94	2.73													■
			UBA6919	CAJXGR01	66	2.27	■	■							■				
			UBA7883	DRJW01	89	4.44									■				
	Unclassified	SAMEA5744179	Gammaproteobacteria	GCA-2729495	84	6.65	■	■											
				JAACFB01	51	8.55		■											
Hadopelagic (>6000m)	Subduction Zone and Trenches	SAMN16252685	Gammaproteobacteria	SMWN01	55	1.10										■			
			Nitrospinia	SZUA-226	86	2.42			■										
			UBA1144	CR02bin9	53	0.00	■												
		SAMN16252686	Nitrospinia	SZUA-226	64	0.91			■										
			Terriglobia	JACZYA01	58	1.78	■												

Metagenome assembled genomes (MAGs) from public metagenomes

Functional profiling

Associated taxonomic classification

Presence of genes of interest in poorly known habitats

Semedo, Valquaresma, et al (*in prep*)

Brief tour on IMG/JGI and KBase

IMG/MER

Better for exploring data (lots of genomes, annotations, and relatively good metadata).

https://img.jgi.doe.gov/cgi-bin/mer/main.cgi

- Show around
- Specific task:
  - Find the nosZ gene (K00376) in all available bacterial genomes?
  - Isolated from deep-sea marine sediments?
  - What are the gene neighborhoods in selected genes?
- Anything else you would like to do? Thousands of genes from thousands of genes and metagenomes from different environments (with metadata)

KBase

Better for analyzing your own data (or from collaborators). Many current tools and highly flexible and modular narratives. However, computational speed is limited, since this is a public server.

https://www.kbase.us

Questions?