

MGnify: Free and open microbiome data, analysis and resources at EMBL-EBI

Lorna Richardson

Microbiome Informatics Coordinator

Planetary Protection Metagenomics in Spaceflight Workshop
19-22 November 2024



The European Molecular Biology Laboratory



EMBL-EBI

Bioinformatics



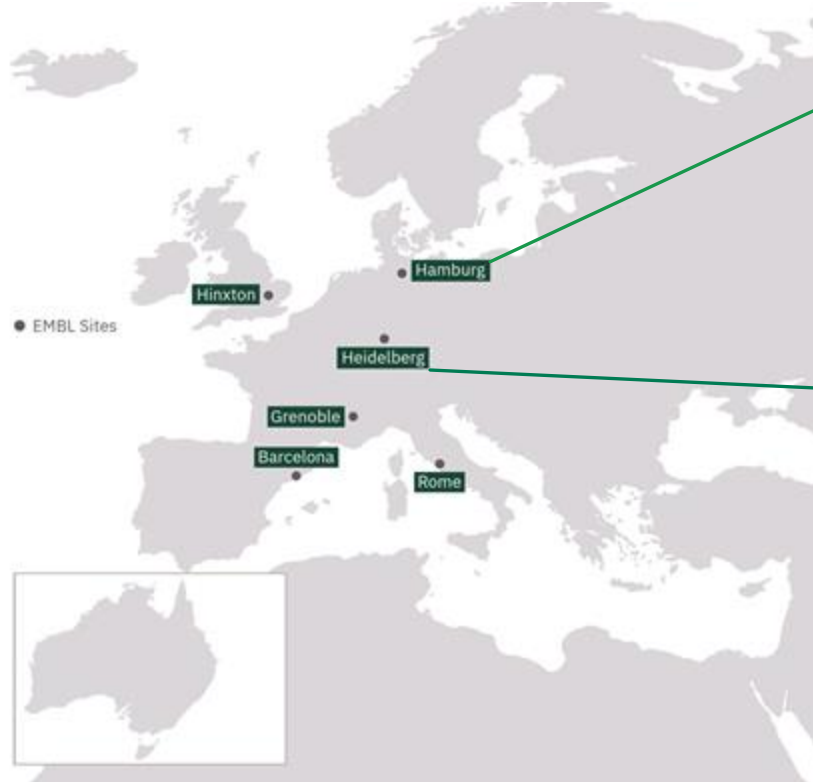
Grenoble

Structural biology



Barcelona

Tissue biology
and disease
modelling



Hamburg

Structural biology



Heidelberg

Life sciences



Rome

Epigenetics
and neurobiology



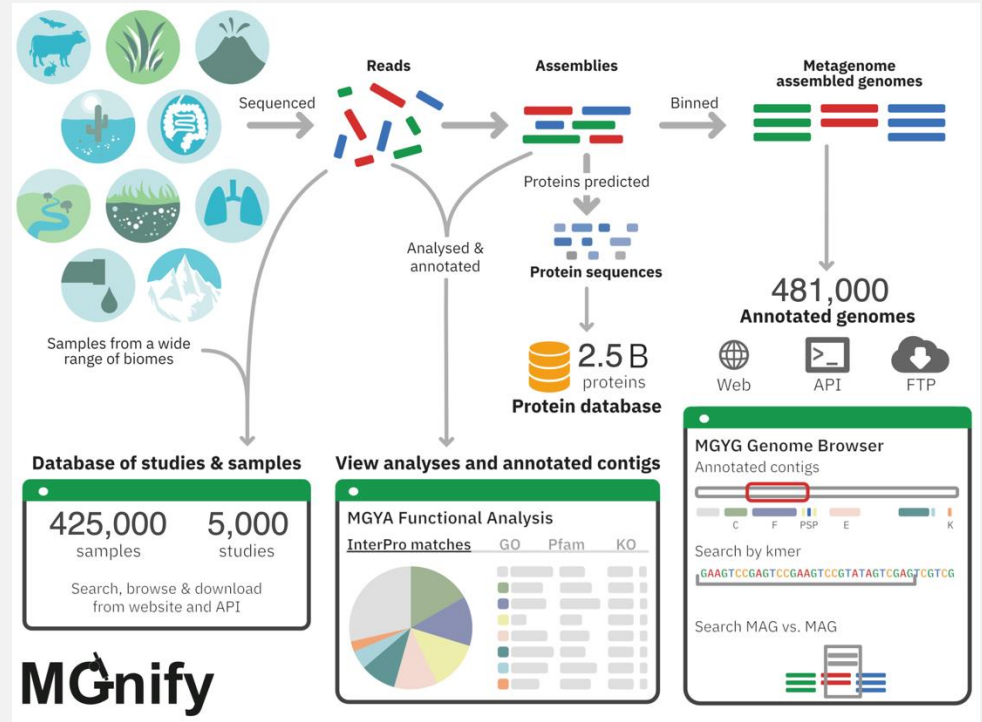
A not-for-profit research organisation funded by over 20 member states

Data resources at EMBL-EBI

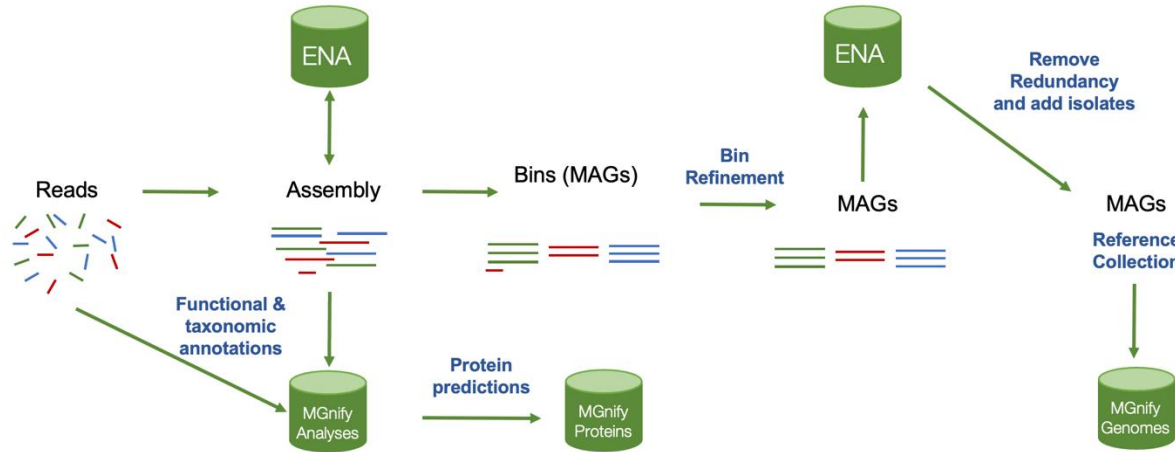


Introducing MGnify

- EMBL-EBI's microbiome data and analysis resource
- Free analysis of metabarcoding, metagenomic, and metatranscriptomic data
- Assembly of raw-reads as a community service
- Biome-specific genome catalogues as references
- Database of 2.5bn proteins from microbiome analysis

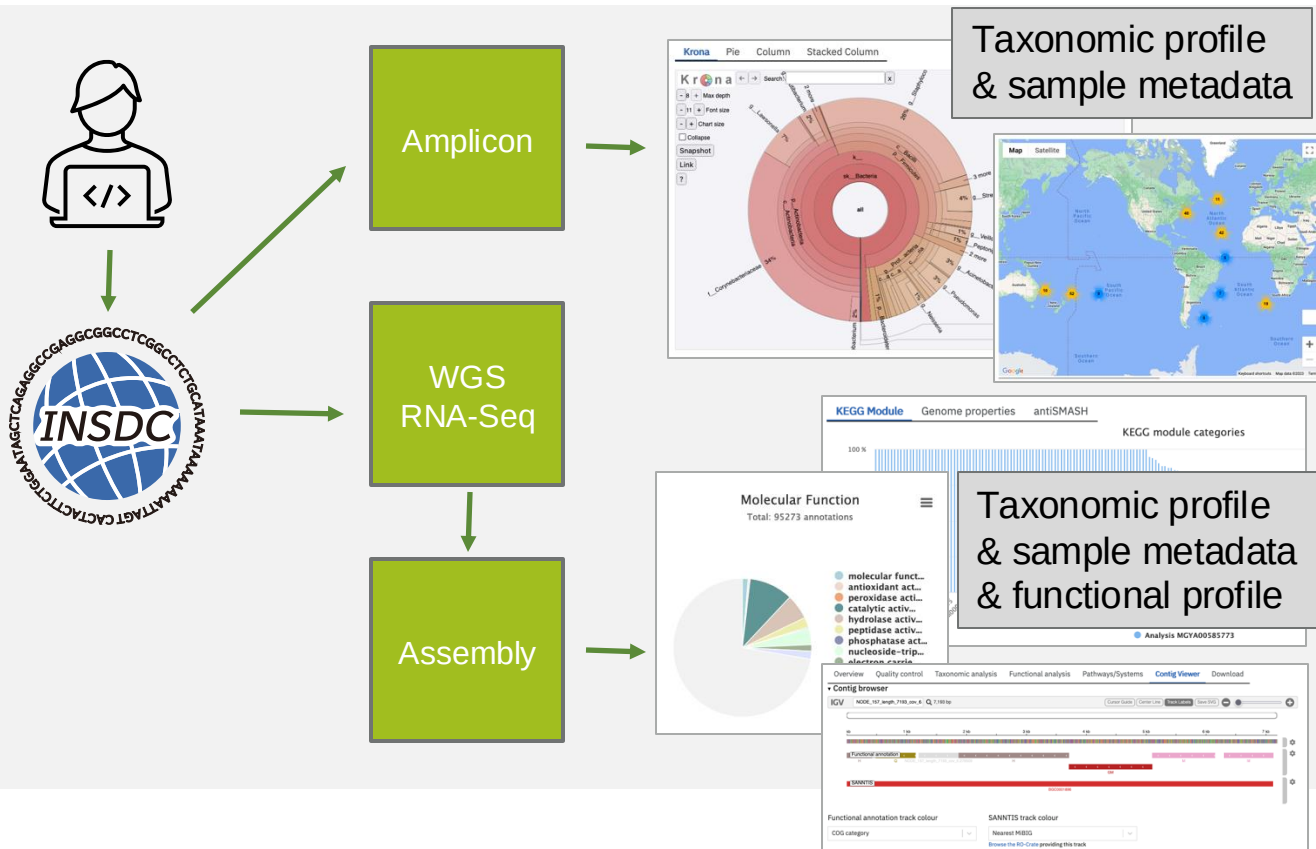


Interface with INSDC (NCBI, ENA, DDBJ)



- Rely on INSDC for data provenance and adherence to metadata standards
 - Fetching raw data
 - Submitting assemblies
 - Submitting MAGs
 - Inheriting metadata
- MGnify provides analysis of microbiome-derived data on request (subject to fair-use)
- Private (pre-publication) data also supported

MGnify analysis pipelines



- Stable versioned analysis pipelines
- New versions cover newer tools and data types
- All results available to download and access via MGnify API

Computational access to data

Programmatic access

The  API endpoint for this Study is

<https://www.ebi.ac.uk/metagenomics/api/v1/studies/MGYS00006491>

Analysing data using R or Python

The  MGnify Jupyter Lab server hosts examples of data analysis using R and Python. These are live examples that you can modify without downloading or installing any software.

 [Open Study in R](#)

 [Open Study in Python](#)

MGnify API

</metagenomics/api/schema>

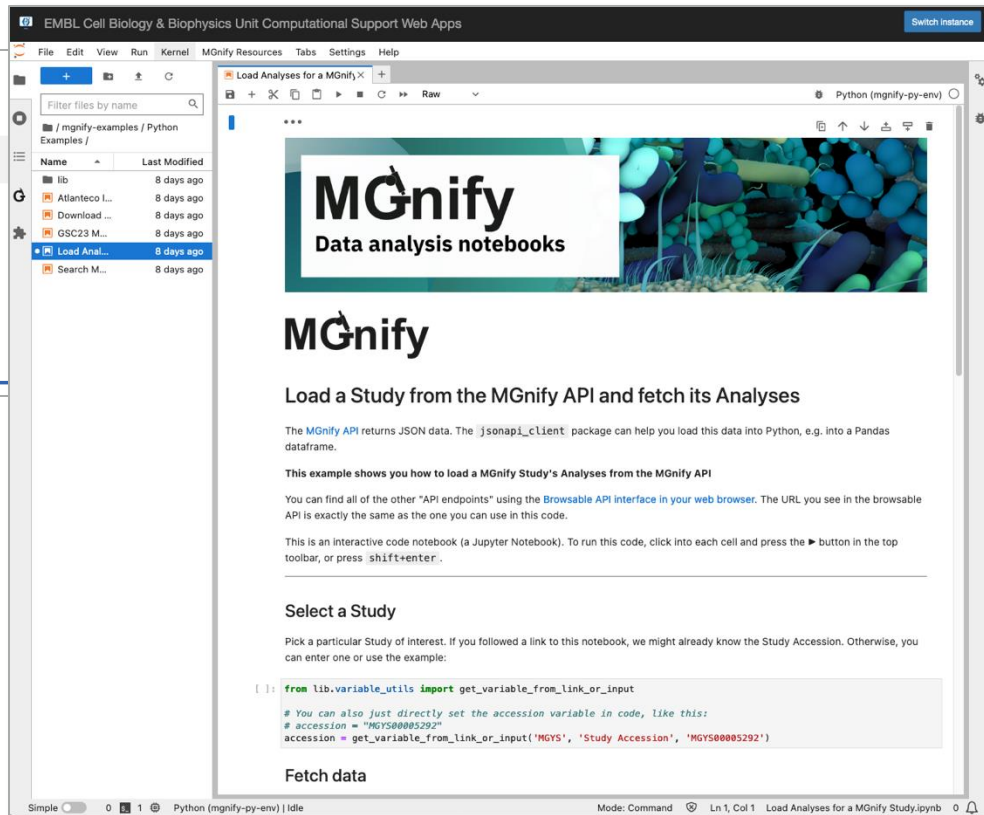
General MGnify documentation is available on Read the Docs: <https://emg-docs.readthedocs.io/>

[MGnify documentation – API section](#)

Servers

</metagenomics/api> - MGnify API 

Authorize



EMBL Cell Biology & Biophysics Unit Computational Support Web Apps

File Edit View Run Kernel MGnify Resources Tabs Settings Help

Filter files by name

/ mgify-examples / Python

Examples /

Name	Last Modified
lib	8 days ago
Atlanteco L...	8 days ago
Download ...	8 days ago
GSC23 M...	8 days ago
Load Anal...	8 days ago
Search M...	8 days ago

MGnify
Data analysis notebooks

MGnify

Load a Study from the MGnify API and fetch its Analyses

The [MGnify API](#) returns JSON data. The `jsonapi_client` package can help you load this data into Python, e.g. into a Pandas dataframe.

This example shows you how to load a MGnify Study's Analyses from the MGnify API

You can find all of the other "API endpoints" using the [Browsable API interface in your web browser](#). The URL you see in the browsable API is exactly the same as the one you can use in this code.

This is an interactive code notebook (a Jupyter Notebook). To run this code, click into each cell and press the ► button in the top toolbar, or press `shift+enter`.

Select a Study

Pick a particular Study of interest. If you followed a link to this notebook, we might already know the Study Accession. Otherwise, you can enter one or use the example:

```
[ ]: from lib.variable_utils import get_variable_from_link_or_input

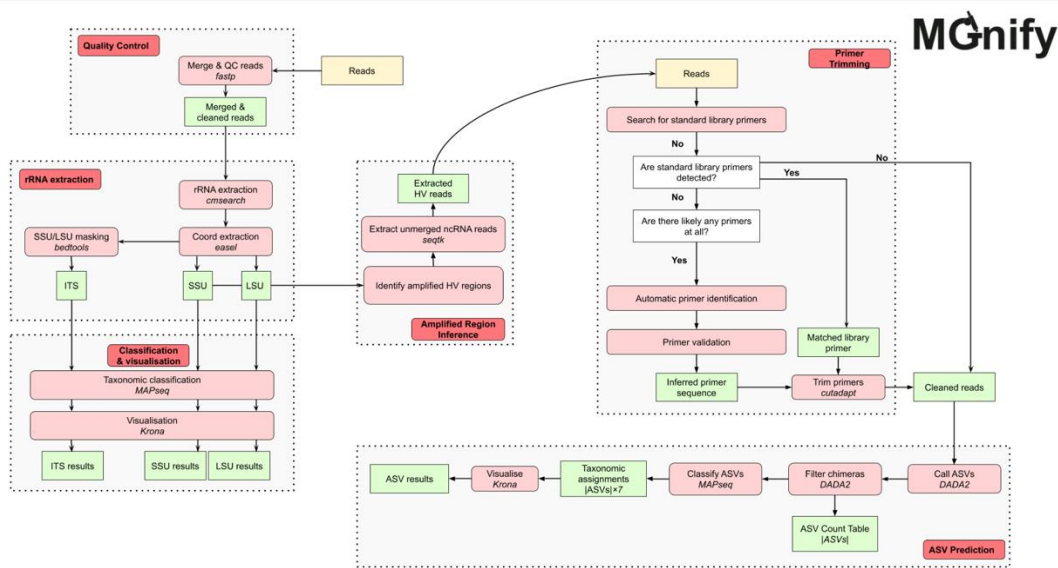
# You can also just directly set the accession variable in code, like this:
# accession = "MGYS00005292"
accession = get_variable_from_link_or_input('MGYS', 'Study Accession', 'MGYS00005292')
```

Fetch data

Simple 0 1 Python (mgify-py-env) | Idle

Mode: Command Ln 1, Col 1 Load Analyses for a MGnify Study.ipynb 0

V6 analysis pipelines



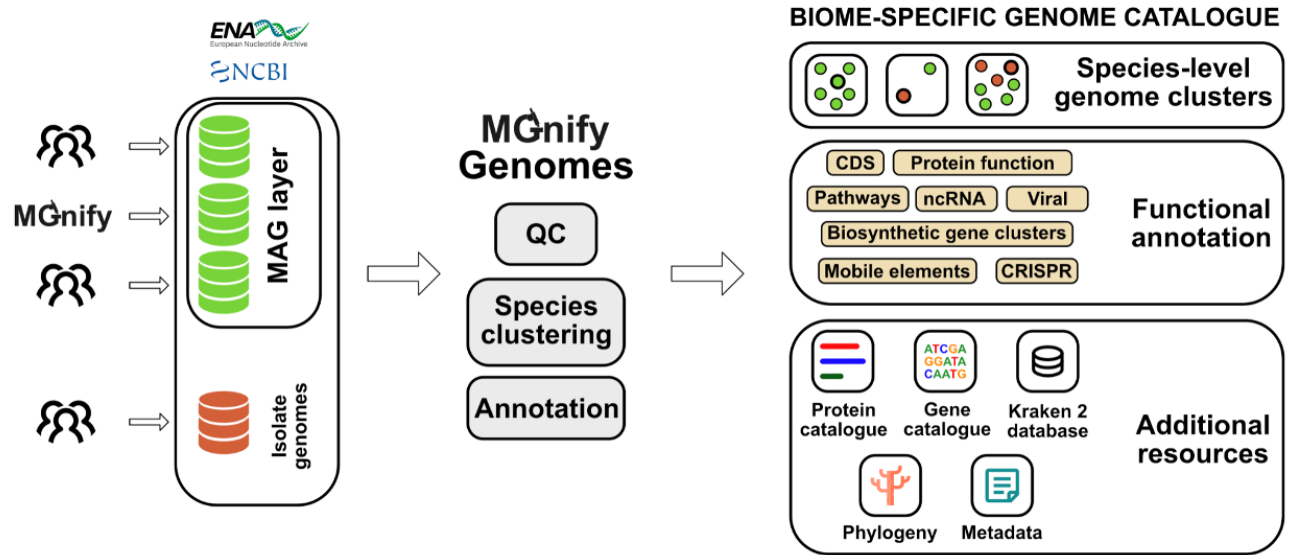
Amplicon analysis pipeline now in beta

- Amplicon
 - Addition of ASVs
 - Improved eukaryotic taxonomy
 - Darwin-core format outputs
- Assembly
 - Viral annotations
 - Mobilome annotations
 - Improved BGC predictions
 - Addition of CAZy
 - Addition of Rhea (and ChEBI)
 - Improved summary outputs
- Raw-read
 - mOTUs and Pfam only

MGnify Genomes

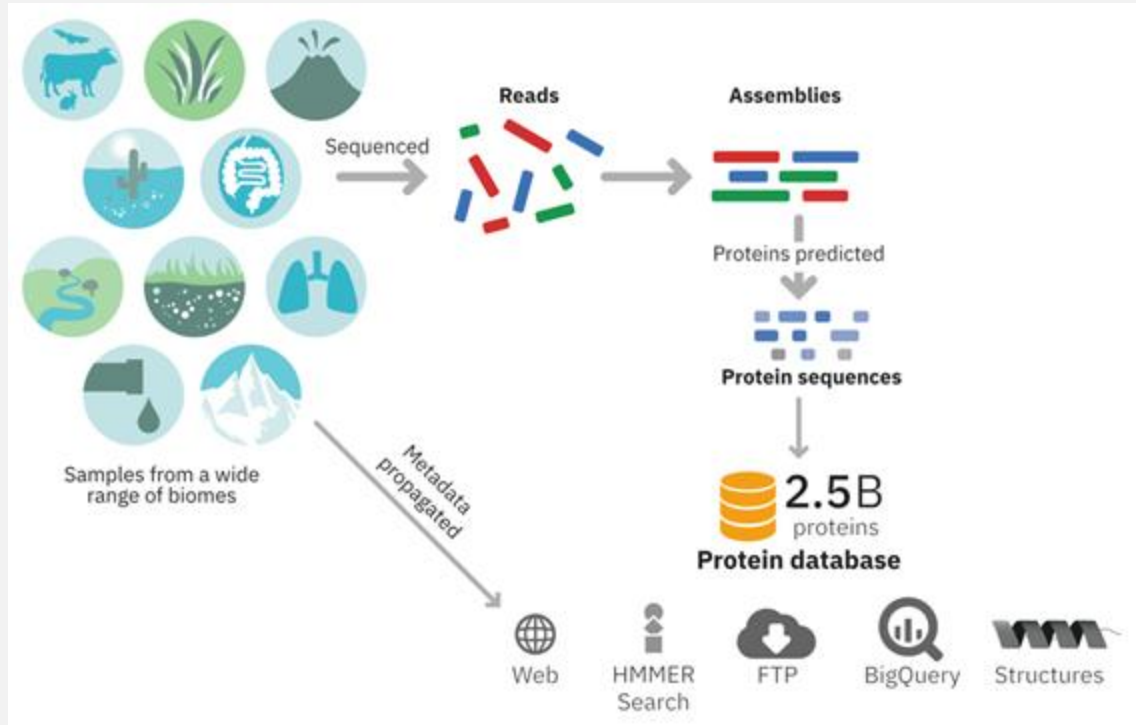
- Environmental
 - Marine
- Host-associated
 - Human gut
 - Human oral
 - Human vaginal
 - Honey-bee gut
 - Cow rumen
 - Sheep gut
 - Pig gut
 - Fish gut
 - Mouse gut
 - Zebrafish fecal

Biome-specific genome catalogues comprised of MAGs and isolates



Use in shallow-shotgun read-mapping process – option for low-biomass data

MGnify Proteins



- *de novo* gene prediction on MGnify contigs
- over 2.5bn non-redundant sequences in latest release
- clustered at 90% coverage and sequence identity into 717 million clusters
- <1% have an identical match in UniProtKB
- ~50% have a match to Pfam
- structural prediction for all cluster reps from ESMAtlas

MGnify Proteins

MGnify Proteins BETA

Explore proteins predicted from metagenomic assemblies.

Example: MGYP000261684433

[Home](#) [MGnify](#) [Sequence Search](#) [Protein Database Release](#) [Documentation](#) [About](#)

Protein: MGYP000261684433

Information

Cluster size 1 protein

Full length ORF ☒

Biomes



Marine
(1 time)

Pfam annotations

Showing 1 to 1 of 1 entries

Pfam Acc	Pfam Name	Clan	Colour	HMM Start	HMM End	Env Start	Env End	Bit Score	Evalue
PF04659	Archaeal flagella protein	-		2	93	325	412	88.1	3e-23

Show entries

Previous ☒ Next



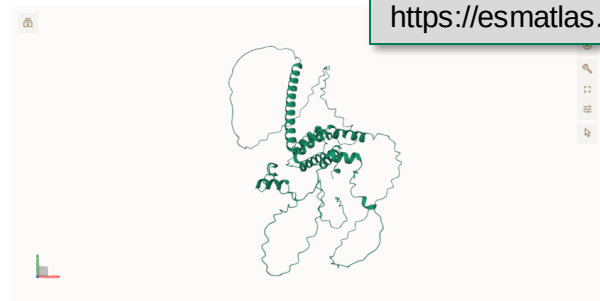
ESM Atlas 3D structure

Predicted 3D protein structure through the Meta AI ESMFold model. ESMFold uses the representations from a large language model (ESM2) to generate an accurate structure prediction from the sequence of a protein.

For more information visit:

[ESM Metagenomic Atlas](#)

Evolutionary-scale prediction of atomic-level protein structure with a language model



Structures pulled from
<https://esmatlas.com/>

Studies - Assemblies - Contigs

Show entries

Search:

Study	Assembly	Contig	Contig Start	Contig End	Strand
ERP108403	ERZ534239	MGYC000564781877	931	2196	-

Showing 1 to 1 of 1 entries

Previous ☒ Next

Sequence

```
EEESGTVRAVQAGDEVPGVEFFKGVVEETLRGGPAGEAAAGDTEAATAETDDHGLGSVGRNEKTSFAELKAYESGDLWVDDRETGPDGTGESEHSKT  
SFADLKAYESGDLWVDDGPGDEAAARTTGGSDERTAAASVETATTEVPVGGEEERTAGEADEGEWDDGEGDEETGDLADELDELFEDEIWDG  
DDNTPEESLARAGEEAGERAEAAEDERTVEDERAAEDERGVSADTADGATATAEDSSADAPTADQRPDDAAGSRGQGEAPSPAASTSAGESPSAD  
REGVPEAASGGATKRRPPGESGEGKPYLETLPQGHWADLLVMEWLEFLVEEGGTQAATRALEYERIGRIDGGVTLELRYLAGFEGDGDGALSIDHHR  
SLSYVDQLGDGDSVERLL
```

Copy

Acknowledgements

Rob Finn (Group Leader)

Chris Atallah

Germana Baldi

Martin Beracochea

Ales Escobar

Santiago Fragoso

Vikas Gupta

Tanya Gurbich

Varsha Kale

Vangelis Karatzas

Jennifer Lu

Jennifer Mattock

Khalimat Murtazalieva

Sonya Ochkalova

Aca Rajkovic

Sandy Rogers

Kate Sakharova

Mahfouz Shehu

Johanna von Wachsmann

Jiawei Wang



EMBL



BIOcean5D

eOSC

Blue-Cloud2026



DTO-BioFlow
Integration of biodiversity monitoring
data into the Digital Twin Ocean

