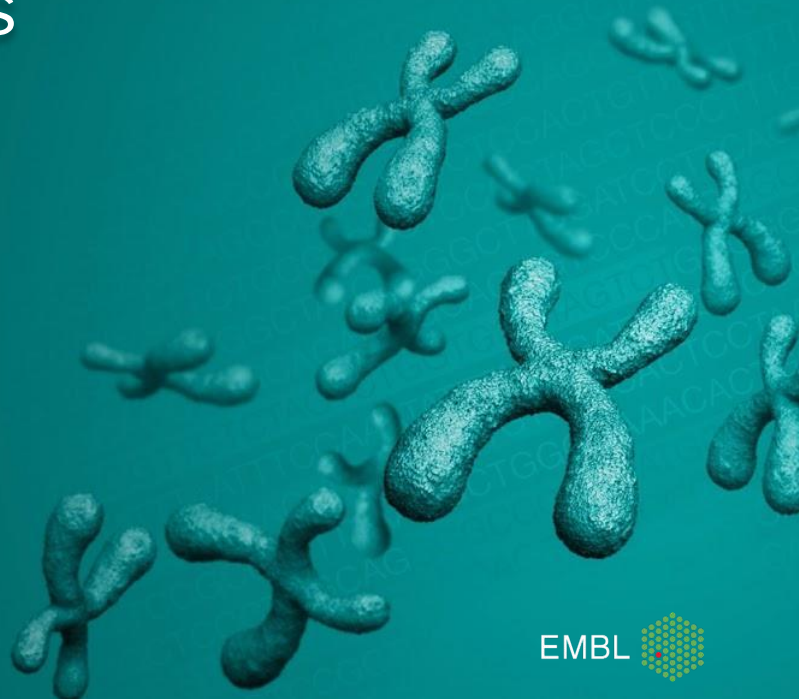
A machine learning framework for discovering and enriching metagenomics metadata from open access research articles

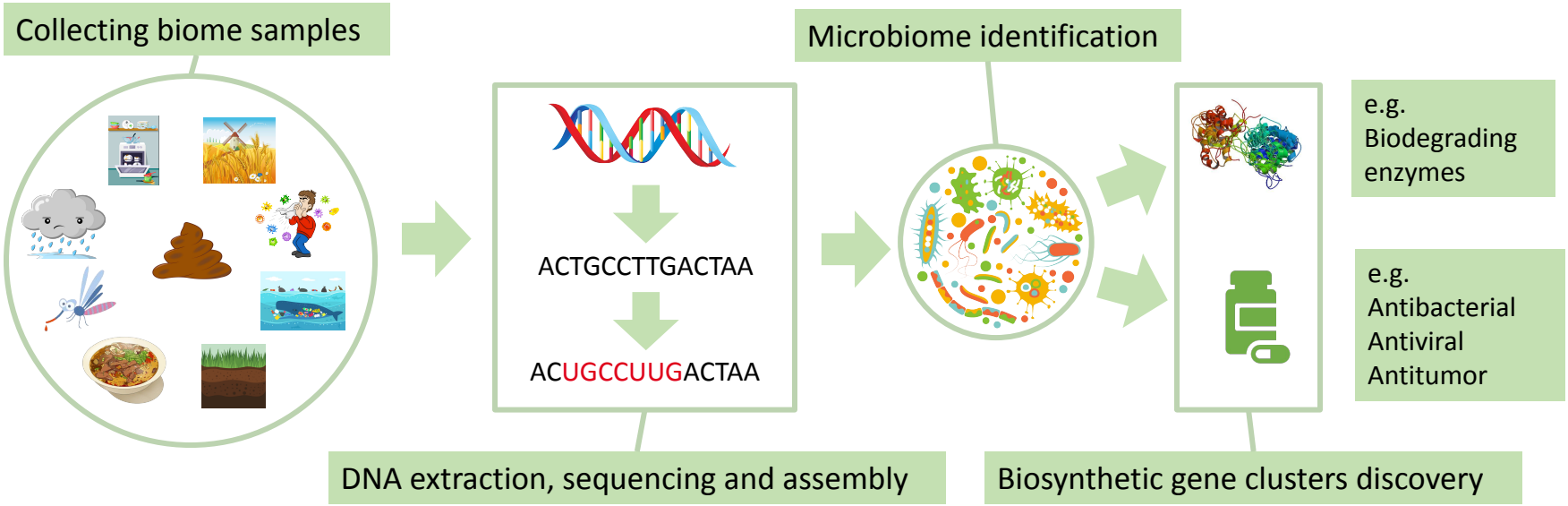
Maaly Nassar, Robert Finn and
Johanna McEntyre



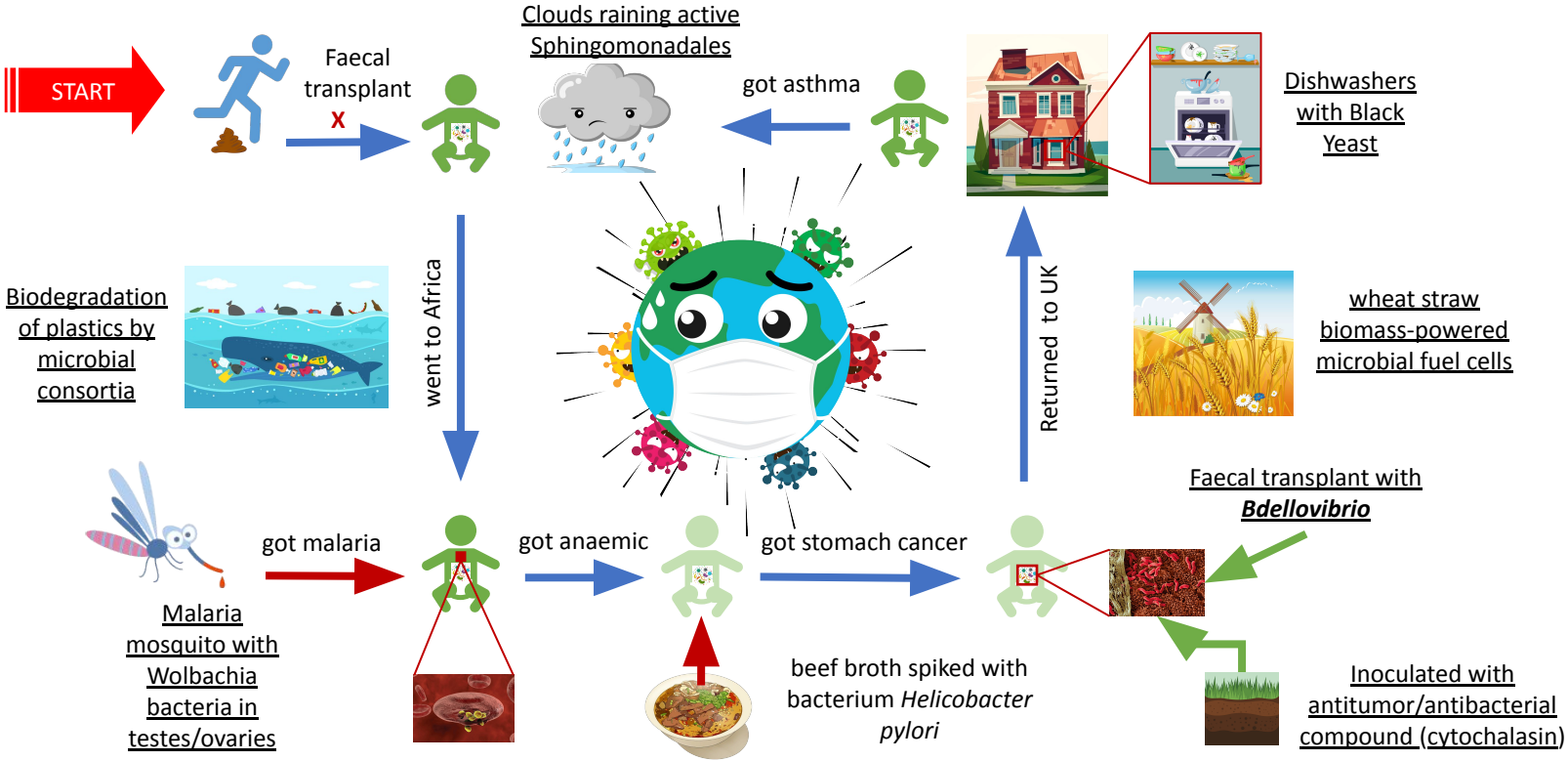
EMBL

Background: Metagenomics

Metagenomics involves the use of non-culture methods to identify microbial consortia in environmental and organisms samples, aiming for uncovering their implications in a variety of environments as well as in health and disease



Microbiomes are everywhere and uncovering/annotating their environments is very challenging



Background: Metagenomics Data Resources

- Metagenomics databases, such as **MGnify**, rely on the manual curation of microbiome sequences, using metagenomics ontologies, such as **GOLD** and **ENVO**.
- However, these ontologies cannot cover the wide variety of microbiome environments and their hierarchical levels can lack important data for metagenomics analysis.

What is missing? main host? host state? exact sample site? host treatment? DNA extraction kit? place? … etc.

Aim and Challenges

- Hence, the need for an automated Named Entity Recognition (NER) that can extract the missing data about microbiome sample & environment from literature.
- But, whereas machine/deep learning (ML/DL) pioneers offered unprecedented NER models to train and implement, there were challenges:
  - ***Database Enrichment.*** How can we enrich databases, such MGnify with data from literature?
  - ***Named-Entity Recognition.*** Which ML/DL models to train for recognizing entities?
  - ***Entities types and biocuration.*** What type of entities do we need to extract, curate and train to have more conclusive metagenomics results?
  - ***Literature triage.*** How can we triage literature to provide a wide variety of microbiome environments contexts for biocuration and models training?

Metagenomics ML/DL Framework

Accordingly, a machine learning (ML) framework was developed to enrich metagenomics studies with data about **microbiome samples**, **environments, living organisms** and **experimental methods** from Europe PMC open access articles**.** This framework includes the following components:

1. **Literature Classification and Triage**
2. **Named Entity Recognition (NER)** that comprises**:**
   - Entity types and biocuration
   - NER models training
3. **Databases Enrichment**

Metagenomics ML/DL Framework: Literature Classification

- A multiclass random forest classifier was trained to classify publications into: 1) **Host-associated**, 2) **Environmental**, 3) **Engineered**.
- Classifier training dataset was constructed by mapping the GOLD biome annotations assigned to the metagenomics studies in MGnify to their corresponding publications in Europe PMC.

Class	Precision	Recall	F1-Score
Engineered	0.88	0.94	0.91
Environmental	1	0.94	0.97
Host-associated	0.94	0.94	0.94

Metagenomics ML/DL Framework: Literature Triage

Using trained classifiers, ~ 14,776 ENA cross-referenced metagenomics papers were categorized into the 3 categories and 140 papers (45-50 papers per category) was randomly selected and manually validated as literature triage:

- Environmental (n=48)
- Engineered (n=46)
- Host-associated (n=46)

Metagenomics ML/DL Framework: Named Entity Recognition

(1) Ecoregion	**(4) Sample-Material** **(5) Date** **(6) Place** **(7) Site** **(8) Body-Site** **(9) State** **(10) Treatment** **(11) Kit** **(12) Gene** **(13) Primer** **(14) LS** **(15) LCM** **(16) Sequencing**
(2) Engineered	
(3) Host	

Entity types and biocuration

- A total of 16 novel metagenomics entities, covering biome and experimental data were curated in 140 ENA cross-referenced metagenomics papers, using Hypothes.is.

Metagenomics ML/DL Framework: Named Entity Recognition

NER models training

- Annotations were mapped to their corresponding sentences and tagged with the language processing annotation scheme BIO .
- 16 individual training datasets were constructed to train each of the metagenomics entities, separately, using BioBERT models
- For each model, grid search was performed over 5 learning rates and 7 epochs to select the best hyperparameters.

Metagenomics ML/DL Framework: Named Entity Recognition

Entity	Learning Rate	Epoch	Recall	Precision	F1-Score
Ecoregion	4e-5	50	0.95	1	0.98
Host	2e-5	90	0.89	0.93	0.9
Engineered	2e-5	10	0.65	0.93	0.75
Date	4e-5	90	0.78	0.91	0.83
Place	3e-5	90	0.78	0.86	0.82
Site	4e-5	10	0.71	0.85	0.77
Body-Site	4e-5	90	0.98	0.95	0.97
Sample-Material	5e-5	110	0.8	0.9	0.85

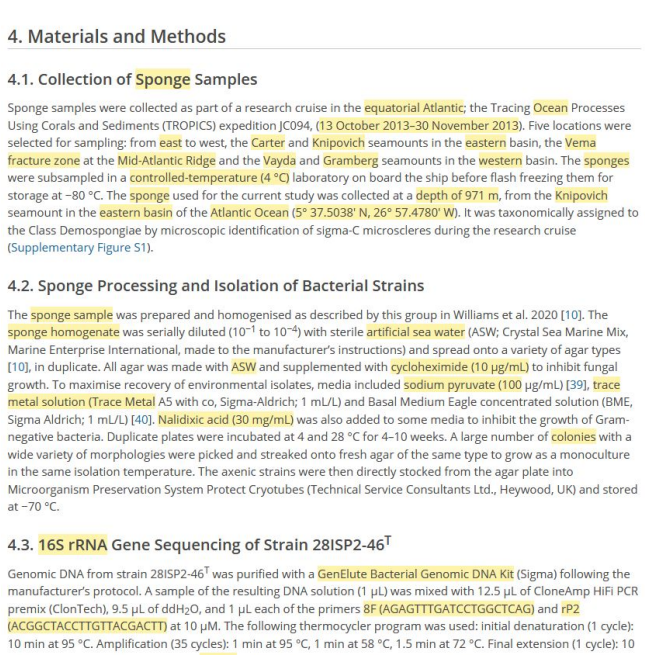
Metagenomics ML/DL Framework: Named Entity Recognition

Entity	Learning Rate	Epoch	Recall	Precision	F1-Score
State	5e-5	110	0.65	0.8	0.71
Treatment	4e-5	30	0.66	0.8	0.73
Kit	2e-5	70	0.94	0.91	0.92
Primer	5e-5	70	0.94	0.97	0.96
Gene	1e-5	10	0.86	0.92	0.89
LS	5e-5	50	0.8	0.95	0.85
LCM	4e-5	50	0.86	1	0.92
Sequencing	5e-5	110	0.84	0.89	0.87

Metagenomics ML/DL Framework: Databases Enrichment

Europe PMC Enrichment

- Trained NER models were used to annotate ~98,213 metagenomics publications in Europe PMC.
- ZOOMA, a semantic ontology mapping tool is used to map predicted annotations to ontologies in OLS.



4. Materials and Methods

4.1. Collection of Sponge Samples

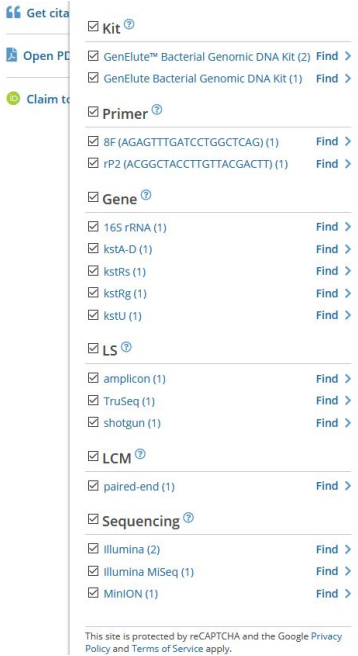
Sponge samples were collected as part of a research cruise in the equatorial Atlantic; the Tracing Ocean Processes Using Corals and Sediments (TROPICS) expedition JC094, (13 October 2013–30 November 2013). Five locations were selected for sampling: from east to west, the Carter and Knipovich seamounts in the eastern basin, the Vema fracture zone at the Mid-Atlantic Ridge and the Vayda and Gramberg seamounts in the western basin. The sponges were subsampled in a controlled-temperature (4 °C) laboratory on board the ship before flash freezing them for storage at −80 °C. The sponge used for the current study was collected at a depth of 971 m, from the Knipovich seamount in the eastern basin of the Atlantic Ocean (5° 37.5038' N, 26° 57.4780' W). It was taxonomically assigned to the Class Demospongiae by microscopic identification of sigma-C microscleres during the research cruise (Supplementary Figure S1).

4.2. Sponge Processing and Isolation of Bacterial Strains

The sponge sample was prepared and homogenised as described by this group in Williams et al. 2020 [10]. The sponge homogenate was serially diluted (10^{-1} to 10^{-4}) with sterile artificial sea water (ASW; Crystal Sea Marine Mix, Marine Enterprise International, made to the manufacturer's instructions) and spread onto a variety of agar types [10], in duplicate. All agar was made with ASW and supplemented with cycloheximide (10 µg/mL) to inhibit fungal growth. To maximise recovery of environmental isolates, media included sodium pyruvate (100 µg/mL) [39], trace metal solution (Trace Metal A5 with co, Sigma-Aldrich; 1 mL/L) and Basal Medium Eagle concentrated solution (BME, Sigma Aldrich; 1 mL/L) [40]. Nalidixic acid (30 mg/mL) was also added to some media to inhibit the growth of Gram-negative bacteria. Duplicate plates were incubated at 4 and 28 °C for 4–10 weeks. A large number of colonies with a wide variety of morphologies were picked and streaked onto fresh agar of the same type to grow as a monoculture in the same isolation temperature. The axenic strains were then directly stocked from the agar plate into Microorganism Preservation System Protect Cryotubes (Technical Service Consultants Ltd., Heywood, UK) and stored at −70 °C.

4.3. 16S rRNA Gene Sequencing of Strain 28ISP2-46^T

Genomic DNA from strain 28ISP2-46^T was purified with a GenElute Bacterial Genomic DNA Kit (Sigma) following the manufacturer's protocol. A sample of the resulting DNA solution (1 µL) was mixed with 12.5 µL of CloneAmp HiFi PCR premix (ClonTech), 9.5 µL of ddH_2O, and 1 µL each of the primers 8F (AGAGTTTGATCCTGGCTCAG) and rP2 (ACGGCTACCTTGTTACGACTT) at 10 µM. The following thermocycler program was used: initial denaturation (1 cycle): 10 min at 95 °C. Amplification (35 cycles): 1 min at 95 °C, 1 min at 58 °C, 1.5 min at 72 °C. Final extension (1 cycle): 10



Get cita
Open PD
Claim to

Kit
- GenElute™ Bacterial Genomic DNA Kit (2) Find
- GenElute Bacterial Genomic DNA Kit (1) Find

Primer
- 8F (AGAGTTTGATCCTGGCTCAG) (1) Find
- rP2 (ACGGCTACCTTGTTACGACTT) (1) Find

Gene
- 16S rRNA (1) Find
- kstA-D (1) Find
- kstRs (1) Find
- kstRg (1) Find
- kstU (1) Find

LS
- amplicon (1) Find
- TruSeq (1) Find
- shotgun (1) Find

LCM
- paired-end (1) Find

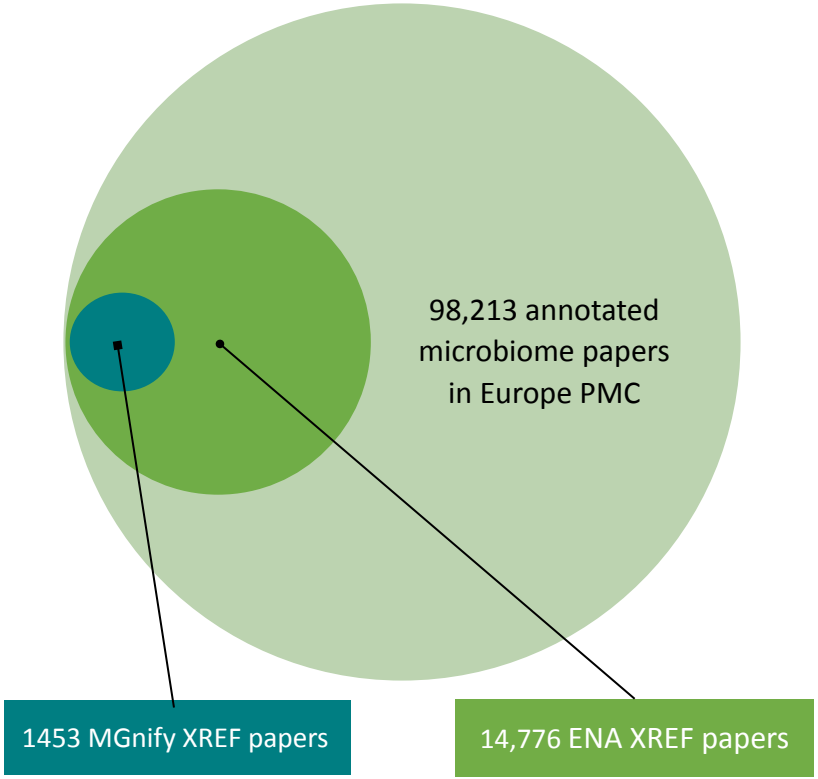
Sequencing
- Illumina (2) Find
- Illumina MiSeq (1) Find
- MinION (1) Find

This site is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply.

Europe PMC search query: ANNOTATION_PROVIDER:"Metagenomics"

Europe PMC annotation API: https://europepmc.org/AnnotationsApi

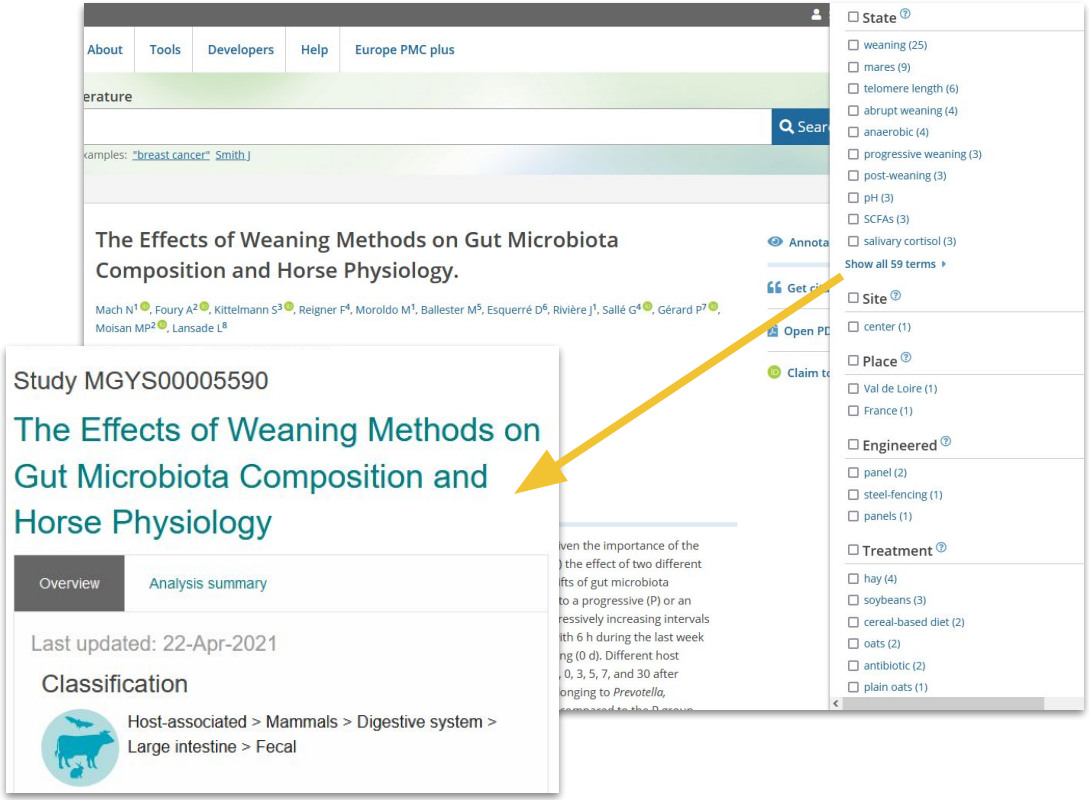
Metagenomics ML/DL Framework: Databases Enrichment



MGnify Enrichment

- Using our machine learning framework, ~ 1453 MGnify and ~ 13,323 new ENA metagenomics studies were annotated for MGnify to use in metagenomics analysis.
- Given their broad range of metadata fields, Europe PMC metagenomics annotations are being included as an enrichment layer in MGnify.

Metagenomics ML/DL Framework: Databases Enrichment



MGnify Enrichment

Examples for the annotated publications:

1,2,3,4,5,6,7,8

Metagenomics ML/DL Framework: Databases Enrichment


```
{
  "curations": [
    {
      "id": "701f872e-960b-4167-b83b-7c3f7fa29ed4",
      "recordType": "sample",
      "recordId": "SAMN14408423",
      "attributePre": "string",
      "valuePre": "string",
      "attributePost": "Place",
      "valuePost": "Alaminos Canyon",
      "attributeDelete": false,
      "assertionMethod": "automatic assertion",
      "assertionEvidences": [
        {
          "identifier": "GAZ:00000004",
          "shortForm": "GAZ_00000004",
          "label": "Alaminos Canyon"
        }
      ],
      "assertionSource": "string",
      "assertionAdditionalInfo": "string",
      "providerName": "EMERALD",
      "providerUrl": "https://gtr.ukri.org/projects?ref=BB/S009043/1",
      "submittedTimestamp": "2020-12-03T12:34:12.865+0000",
      "updatedTimestamp": "2020-12-03T14:04:27.325+0000",
      "suppressed": true
    }
  ],
  "total": 1
}
```


BioSample Enrichment

- Metagenomics annotations are being also integrated into [ELIXIR Contextual Data Clearinghouse] for extending, correcting and improving publicly available annotations on records in sample and sequencing data resources.

Conclusion and Next Steps

- To our knowledge, this work represents the **first large-scale automated enrichment of metagenomics studies** with metadata derived from open access articles.
- Machine and deep learning models were trained, calibrated and validated on curated datasets and applied to annotate metagenomics publications in Europe PMC and enrich MGnify databases.
- Currently, a second machine learning framework is being developed for discovering novel biosynthetic gene clusters from Europe PMC papers.

Acknowledgement

We would like to thank the curators (Molecular Connections) and Santiago Fragoso (MGnify) for their biocuration efforts, and Europe PMC colleagues (Wellcome 108758) for their efforts in annotations integration and visualisation. This work was funded by (Biotechnology and Biological Sciences Research Council (BB/S009043/1).