



EXCELERATE Deliverable D6.5

| | | |
|---|--|------------------------|
| Project Title: | ELIXIR-EXCELERATE: Fast-track ELIXIR implementation and drive early user exploitation across the life sciences | |
| Project Acronym: | ELIXIR-EXCELERATE | |
| Grant agreement no.: | 676559 | |
| | H2020-INFRADEV-2014-2015/H2020-INFRADEV-1-2015-1 | |
| Deliverable title: | Improvement and Application of Eukaryotic Gene catalogue | |
| WP No. | 6 | |
| Lead Beneficiary: | 27, CNRS | |
| WP Title | Marine metagenomic infrastructure as a driver for research and industrial innovation | |
| Contractual delivery date: | 31 May 2019 | |
| Actual delivery date: | 29 May 2019 | |
| WP leader: | Rob Finn Nils Peder Willassen | 1, EMBL-EBI 24, UiT |
| Partner(s) contributing to this deliverable: | CNRS, EMBL-EBI | |

Authors and Contributors:

Eric Pelletier (CNRS), Erwan Corre (CNRS), Guita Niang (CNRS), Arnaud Meng (CNRS), Mark Hoebeke (CNRS) and Rob Finn (EMBL-EBI)

Reviewers:

None

1. Table of contents

| | |
|--|----|
| Table of contents | 2 |
| 2. Executive Summary | 2 |
| 3. Impact | 3 |
| 4. Project objectives | 3 |
| 5. Delivery and schedule | 4 |
| 6. Adjustments made | 4 |
| 7. Background information | 4 |
| 8. Appendix 1: Report on Improvement and Application of Eukaryote Gene Catalog | 8 |
| 8.1. Introduction | 8 |
| 8.2. Data sources | 9 |
| 8.3. Pipelines | 10 |
| Transcriptome data assembly pipeline | 10 |
| Assembly annotation pipeline | 11 |
| CWL description | 11 |
| 8.4. Reference Resource | 12 |
| 8.5. Applications | 13 |
| 8.6. METdb website | 13 |
| Search module | 16 |
| Description of the readset | 19 |
| Description of the assembly | 19 |
| 8.6. Summary and Future plans | 20 |

2. Executive Summary

- A new resource, METdb, has been established to house the data produced from the outputs of the assembly and annotation workflow runs on 489 transcriptomes. This new resource dramatically increases the representation of micro-eukaryotic organisms, both in this new database, as well as being propagated to core data resources such as ENA and UniProtKB. METdb provides a unique collection of eukaryotic gene annotations, and is expected to become an important reference collection for the interpretation of marine metagenomics datasets.
- For this deliverable, we developed two new pipelines for assembly and annotation of marine micro-eukaryotic transcriptomes. These have been converted to CWL (other work) leveraging many of the tool descriptions produced as part of Deliverable D6.3. Their application highlights the re-use of CWL tool descriptions and the outputs of the Compute (WP4) and Interoperability platforms (WP5), and demonstrates how workflows can be used to make new data resources.

- A web interface has been developed to provide users access to the data contained within METdb, and importantly exposed data that have previously languished in undiscoverable laboratory websites, increasing discoverability of this data.
- While this represents an important new development, the marine micro-eukaryotic kingdom remains massively undersampled despite the sequence diversity. The annotations in METdb provide another resource for the discovery of novel enzymes of the biotechnology sector (as well as the new Microbial Biotechnology Community).

3. Impact

A scientific manuscript describing the METdb resource, the underlying bioinformatic pipelines, database design, access web server and sequence resources is in preparation, with a submission date planned by the end of June 2019. The METdb database will be presented at the All-hands meeting in June.

As this is a new resource, there are no usage statistics. However, we expect the results to be disseminated broadly, as the data will be made available via not only METdb, but the also the ELIXIR core data resources the European Nucleotide Archive and UniProt, both of which received millions of hits per month. Thereafter, we anticipate other resources, such as Pfam, will construct new protein families based on these new sequences.

The CWL descriptions of pipelines developed here will be available for others in the community to use. This will ensure consistency of annotations, as well as negating the need of independently recreated the transcriptomics pipeline.

The data contained within METdb has been used as the main reference resource for taxonomic assignment in the next version of the Tara Oceans derived eukaryotes Unigenes catalog (Carradec et al. 2018).

4. Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives:

| No. | Objective | Yes | No |
|-----|--|-----|----|
| 1 | Development and implementation of selected standards for the marine domain. (Task 6.1) | x | |

- | | | |
|---|---|---|
| 2 | Development and implementation of databases specific for the marine metagenomics. (Task 6.2) | x |
| 3 | Evaluation and implementation of tools and pipelines for metagenomics analysis. (Task 6.3) | x |
| 4 | Development of a search engine for interrogation of marine metagenomics datasets and establish training workshops for end users. (Task 6.4) | x |

5. Delivery and schedule

The delivery is delayed: Yes • No ☒

6. Adjustments made

Not applicable

7. Background information

Background information on this WP as originally indicated in the description of action (DoA) is included here for reference.

| | | | |
|---------------------|--|-------------------------------|---------|
| Work package number | 6 | Start date or starting event: | month 1 |
| Work package title | Use Case A: Marine metagenomic infrastructure as driver for research and industrial innovation | | |
| Lead | Nils Peder Willassen (NO) and Rob Finn (EMBL-EBI) | | |

Participant number and person months per participant

P1: EMBL-EBI (28PM) - P17: FCG (2PM) - P20: CCMAR (11PM) – P24 UiT (36PM) – P27: CNRS (10PM) - P31: CNR (10 PM)

Objectives

The main objective for this Use Case is to develop a sustainable metagenomics infrastructure to enhance research and industrial innovation within the marine domain before M36 of the ELIXIR-EXCELERATE project. The main objective will be achieved by the following specific objectives:

- Development and implementation of selected standards for the marine domain. (Task 6.1)
- Development and implementation of databases specific for the marine metagenomics. (Task 6.2)
- Evaluation and implementation of tools and pipelines for metagenomics analysis. (Task 6.3)
- Development of a search engine for interrogation of marine metagenomics datasets and establish training workshops for end users. (Task 6.4)

Description of work

Metagenomics has the potential to provide unprecedented insight into the structure and function of heterogeneous communities of microorganisms and their vast biodiversity. Microbial communities affect human and animal health and are critical components of all terrestrial and aquatic ecosystems. They can be exploited e.g. to identify novel biocatalysts for production of fuels or chemicals (bioprospecting), make functional feed for aquaculture species, and for environmental monitoring. However, in order to expand the potential further for the research community and biotech industry, especially within the marine domain, the metagenomics methodologies need to overcome a number of challenges related to standardization, development of relevant databases and bioinformatics tools. New and emerging sequencing technologies, integration of metadata gives an extra burden to the development of future databases and tools. The Use Case “Marine metagenomic infrastructure as driver for research and industrial innovation” will contribute to the overall objectives of the ELIXIR-EXCELERATE project by developing research infrastructure and service provision specific for the marine domain in order to enable metagenomic approaches responding to societal and industrial needs. The outcome of the proposed Use Case will meet the major needs expressed by the marine domain (e.g. ESF Marine board Position Paper 17 “Marine Microbial Diversity and its role in Ecosystem Functioning and Environmental Change” and Position Paper 15 “Marine Biotechnology: A New Vision and Strategy for Europe”).

Task 6.1: Development and implementation of a comprehensive metagenomics data standards environment for the marine domain (12 PM)

To maximise the impact and long term utility and discoverability of metagenomics datasets, it is essential the experimental methods and data acquisition/storage protocols be established. In Task 6.1, we will bring together a comprehensive metagenomics data standards environment in collaboration with marine experimental scientists, data providers, end users and the existing communities involved in marine standards development. The environment will bring together three components:

- Data format conventions and standards will address the various data types for which sharing is required, that will include contextual data (e.g. sample information, expedition-related data), metadata (e.g. provenance and tracking information, descriptions of experimental configurations and bioinformatics tools in use) and data (e.g. raw sequence data, aligned reads, taxonomic identifications, gene calls).
- Reporting standards will address community-accepted thresholds for richness/precision that are required to make data useful, including depth of raw machine data, such as resolution of sequence quality scoring,

conventions for references to reference assemblies and minimal reporting requirements for contextual data.

- Validation tools will address the automated validation of compliance with conventions and standards and the meeting of minimal reporting expectations for given datasets in preparation by the marine research community. In this task, we will bring together components that exist already – in particular the contextual data and metadata reporting standards we have developed under the Micro B3 project (EU FP7), data standards and conventions developed around our European Nucleotide Archive (ENA) programme, such as CRAM, FASTQ conventions, work existing in the biodiversity and molecular ecology domains (such as tabular data conventions and BIOM matrices) – and construct new components as required. The major output of this work will be a set of well described and navigable elements to aid the marine community in the preparation, sharing, dissemination and publication of highly interoperable and comprehensive metagenomics datasets.

Partners: EMBL-EBI, NO

Task 6.2. Establishment of marine specific data resources (20PM)

Due to the data biases of existing reference databases, only about one quarter of sequences are annotated, and this fraction diminishes further when more diverse samples such as soil and marine are analyzed. To improve the characterization of marine metagenomic samples, this task involves the construction of sustainable public data resources for the marine microbial domain. Task 6.2 will be achieved by establishing marine microbial databases including reference genomes, nucleotide and protein databases. The established databases, based on the standards developed in Task 6.1, will enhance the precision and accuracy of biodiversity and function analysis. The reference databases will be non-redundant datasets generated from sequences acquired from ENA (as part of the International Nucleotide Sequence Database Collaboration), UniProt and other publicly available datasets. In particular, we will use some of the higher-coverage and higher quality sequence outputs from the Tara Oceans and Ocean Sampling Day metagenomic projects, to build high quality marine specific reference databases. All datasets will be checked with respect to quality, consistency, and interoperability, and in compliance with standards developed in

Task 6.1. The respective knowledge-enhanced databases will be the cornerstone for sustainable analysis of marine metagenomics sequence data. The databases will be developed in collaboration with members of the ESFRI infrastructures EMBRC and MIRRI and made publicly available through ELIXIR.

Partners: NO, EMBL-EBI, IT

Task 6.3: Gold-standards for metagenomics analysis (58PM)

The majority of existing metagenomics analysis platforms, while providing insights into the prokaryotic taxonomic diversity and functional potential for individual samples, but lack the tools that enable discoverability across samples and industrial innovation. This task will focus on the evaluation and implementation of new tools and pipelines in order to accelerate research, discoverability and innovation, reducing

time to market for new products. In combination with new standards and databases developed in Task 6.1 and Task 6.2, respectively, new tools for community structure (microbial biodiversity), genetic and functional potential will be evaluated and implemented for environmental applications. For industrial application tools and pipelines for the identification of gene products (e.g. enzymes and drug targets) and pathways will be implemented and made publicly available.

The evaluation and implementation will be performed in near collaboration with end-users (research groups, environmental centers, biotech companies) to ensure usability for the end user community in order to improve [ELIXIR-EXCELERATE] quality, productivity and functionality, as well as reduction of costs for the end-users. New tools and pipelines will be made publicly available through the e.g. META-pipe (ELIXIR-NO), EBI Metagenomics Portal (EMBL EBI) and/or EMBL Embassy cloud technology. Technical requirements will be mapped by WP3 and implemented to meet the requirements of the ELIXIR community. The continued advancement of sequencing technologies and the growing number of public marine metagenomics projects means that it is becoming increasingly difficult to mine these vast datasets. In this task, initially a web-based search engine will be developed for the interrogation of marine metagenomics results available from the EBI Metagenomics Portal, based on combinations of queries to our web services (already in existence, or to be built as part of existing projects outside ELIXIR-EXCELERATE) for the discovery of data through metadata, taxonomic and functional fields. This will extend the back-end search functionality that is to be developed as part of on-going efforts. In addition to being downloadable, we will enable search results to flow into an expanded comparison tool (currently limited to gene ontology terms from samples in the same project), to allow more in-depth analysis of a user selected datasets, allowing functional and taxonomic comparisons. In the second phase of this task, the search engine will build upon the data exchange formats in Task 6.1, and federate the search across different pipeline results sets (e.g. META-pipe), so that different results based on the same underlying dataset, can be amalgamated into a single search. This will dramatically enhance the discoverability across different marine datasets, allowing the identification of common trends and/or differences. These tools will be developed using user-experience testing and in collaboration with end users to ensure they are fit for purpose.

Partners: NO, EMBL-EBI, IT, FR, PT

Task 6.4: Training workshops for end users (7PM)

In this task training workshops will be established, in collaboration with WP11 "ELIXIR Training Programme", for end-users with the aim to facilitate accessibility, by training European researchers and industry to more effectively exploit the data, tools and pipelines, and compute infrastructure provided by the ELIXIR marine metagenomics infrastructure. These training workshops and materials will be converted to online training resources, extending the reach of the workshop.

Partners: EMBL-EBI, NO

8. Appendix 1: Report on Improvement and Application of Eukaryote Gene Catalog

8.1. Introduction

The scarcity of genomic references for micro-eukaryotic marine species limits our understanding of the largest and most diversified biotope on Earth. About 90% of the complete eukaryotic genome-sequencing projects from the Genome Online Database¹ are from animals, fungi and plants. Thus, these reference genomes only encapsulate two of the major groups of eukaryotes, while the rest remains poorly investigated². Deliverable 6.5 has involved a close collaboration between Genoscope, the Roscoff bioinformatics platform and EMBL-EBI to develop a genomic reference database dedicated to micro-eukaryotic marine species transcriptomes, namely METdb³. This resource integrates and harmonises sequence data from multiple independent sources:

- Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) data^{4 5}
- Roscoff Culture Collection⁶
- Roscoff marine station research teams
- Tara Oceans consortium research projects

A total of 489 RNA-seq datasets, encompassing 233 different genera and 27 phyla were gathered. All datasets were assembled and analyzed using two standardised workflows dedicated to (i) the *de novo* transcriptome assembly and (ii) the functional annotation of the assembly. Both workflows were described using the Common Workflow Language (CWL) to ensure the portability and reproducibility (funded by an ELIXIR Implementation Study as CWL was not in EXCELERATE work plan). One of the goals for this resource was to harmonise the transcriptome assembling and annotation workflows, using standardised and up-to-date methods. In addition to the workflow outputs (assemblies and sequence annotations), for each transcriptome we provide a manually curated taxonomic lineage and the corresponding sample environmental metadata. Features of the METdb web server include the possibility to download either subsets or the complete dataset (assembled transcriptomes, predicted proteomes and functional annotations) as well as associated metadata (taxonomy, sampling conditions or quality metrics).

¹ <https://gold.jgi.doe.gov/>

² Sibbald et al. *Nature Ecology & Evolution*1 (April): 0145. <https://doi.org/10.1038/s41559-017-0145>.

³ <http://metdb.sb-roscoff.fr/metdb/>

⁴ Keeling et al. *PLOS Biology*12 (6): e1001889. <https://doi.org/10.1371/journal.pbio.1001889>.

⁵ Johnson et al. *GigaScience*8 (4). <https://doi.org/10.1093/gigascience/giy158>.

⁶ <http://roscoff-culture-collection.org/>

Finally, every transcriptome assembly and associated annotation will be deposited in the European Nucleotide Archive (ENA), where they will then flow to other resources such as UniProtKB.

8.2. Data sources

The primary (raw) data used for this project are sourced from the publicly available sequence archive databases (namely the European Nucleotide Archive at EMBL-EBI). The raw sequences were generated by the MMETSP collaborative project⁷ and Roscoff Marine Station research projects initiated by D. Vault, C. De Vargas, F. Not and L. Guillou, and sequenced by Genoscope in the context of the *Tara* Oceans consortium.

The majority of datasets are dominated by 4 taxonomic supergroups (Stramenopiles, Alveolata, Haptophyta and Viridiplantae), known to be abundant in the global oceans⁸ (Figure 1).

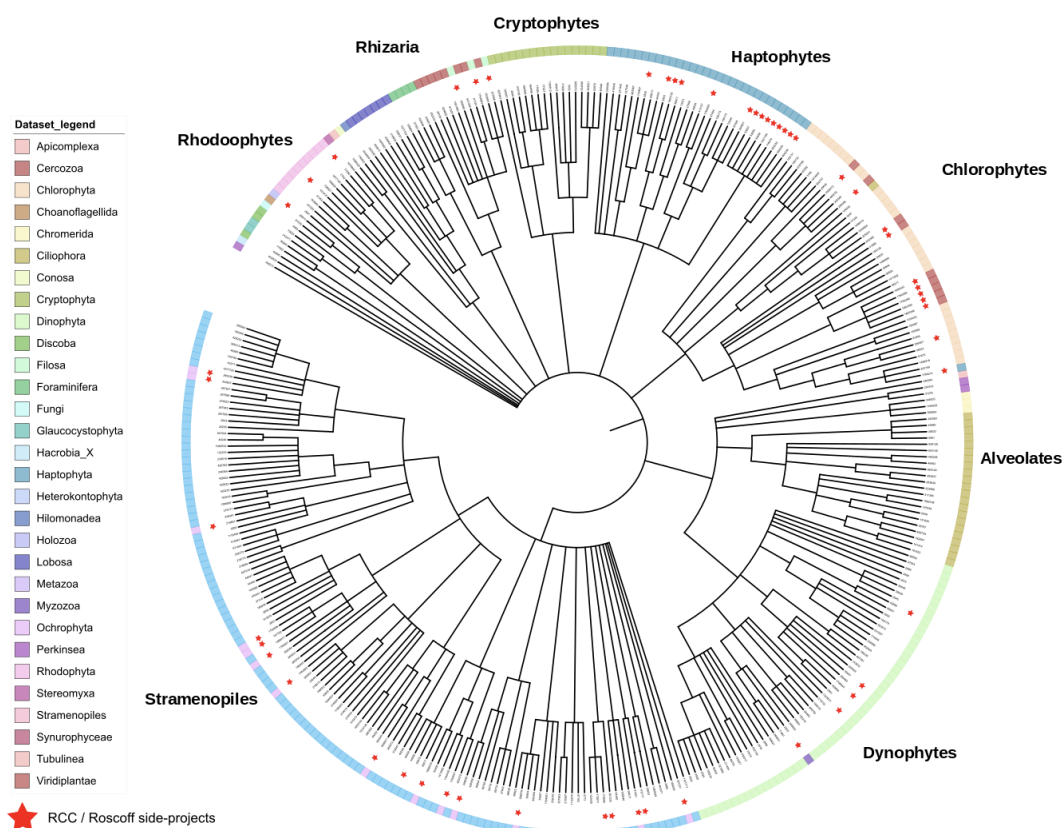


Figure 1 - Phylum covered by the METdb resource (RCC additional species: Haptophyta, Dinophyta, Viridiplantae, Filosa)

⁷ Keeling et al. *PLOS Biology* 12 (6): e1001889. <https://doi.org/10.1371/journal.pbio.1001889>.

⁸ Vargas et al. *Science* 348 (6237): 1261605. <https://doi.org/10.1126/science.1261605>.

8.3. Pipelines

The data processing (transcriptome assembly, functional annotation and quality evaluation) requires stable, reproducible and scalable tools (and workflows) to ensure consistency of results and efficient processing of the data. These were devised as part of the EXCELERATE program, but the CWL implementation funded by ELIXIR. The CWL work build upon work carried out in deliverable D6.3, and have been briefly described for completeness.

Transcriptome data assembly pipeline

We then used the transcriptome analysis workflow originally described in Meng et al. 2018⁹ that has been used in a marine transcriptomic context^{10 11}. It has been designed for *de novo* assembly analysis to promote the discovery of potential new genes in non-model organisms, typified by marine protists. The workflow includes 4 distinct steps: i) raw data processing with Trimmomatic¹² to filter and trim reads according to their sequence quality; ii) readset comparison using Simka to detect possible cross libraries contaminations¹³; iii) *de novo* assembly step using Trinity¹⁴ iv) quality evaluation of the assembled transcripts using Transrate¹⁵ (Figure 2).

Assembly annotation pipeline

Downstream analyses of assemblies includes transcriptome completion evaluation using Busco, coding regions prediction using TransDecoder, and functional annotation of predicted proteins using the InterProScan pipeline from EMBL-EBI.

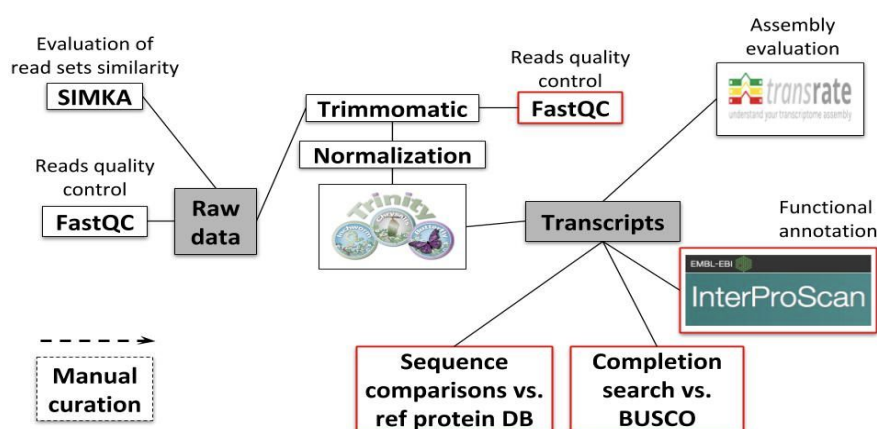


Figure 2 - schematic of assembly and annotation workflows

⁹ Meng et al. 2018. *Molecular Ecology*27 (10): 2365–80.<https://doi.org/10.1111/mec.14579>.

¹⁰ Meng et al. 2018. *Molecular Ecology*27 (10): 2365–80.<https://doi.org/10.1111/mec.14579>.

¹¹ Meng et al. 2018. *Microbiome*6 (1): 105.<https://doi.org/10.1186/s40168-018-0481-9>.

¹² Bolger et al. 2014. *Bioinformatics*30 (15): 2114–20.<https://doi.org/10.1093/bioinformatics/btu170>.

¹³ Bnoit et al. 2016. *PeerJ Computer Science* <https://peerj.com/articles/cs-94/>

¹⁴ Grabherr et al. 2011. *Nature Biotechnology*29 (7): 644–52.<https://doi.org/10.1038/nbt.1883>.

¹⁵ Smith-Unna et al. 2016. *Genome Research*26 (8): 1134–44.<https://doi.org/10.1101/gr.196469.115>.

CWL description

Both assembly and annotation pipelines have been described using the CWL language (<https://www.commonwl.org/v1.0/>) in the context of the aforementioned implementation study: Enabling the reuse, extension, scaling, and reproducibility of scientific workflows (Figures 3 and 4) .

Both pipelines are publicly available on GitHub:

Assembly pipelines:

- paired-end¹⁶
- single¹⁷

Alternative view of the paired-end version¹⁸

Annotation pipeline:

CWL version¹⁹

Alternative view²⁰

These pipelines have been implemented both on high performance/throughput compute platforms and Cloud infrastructures and have been used to performed the METdb assemblies annotation on the EMBL-EBI Embassy cloud infrastructure. Both of the pipelines are currently implemented on the French Bioinformatic Cloud (GenOuest instance) in the context of the Data services and reporting standards work package (WP4) of the EMBRIC project²¹.

8.4. Reference Resource

The complete METdb comprises 489 transcriptome assemblies with associated quality reports and annotations. Some of these assemblies originate from the work of Johnson *et al.*²², corresponding to the MMETSP data, and were downloaded from the public deposition associated with the aforementioned publication²³. The average number of assembled contigs per transcriptome is 59,379, which is characteristic of the heterogeneous genome structures observed among marine microbial eukaryotes²⁴. The transcriptome assembly of *Debaryomyces hansenii* shows the minimum size with 3514 contigs while *Phaeocystis globosa* strain PCC64 has the largest assembly composed of 413,385 contigs. Small and large transcriptome assemblies may result from the lack of sequence data, thereby limiting assembly (small) or causing fragmentation (large). For

¹⁶<https://github.com/EBI-Metagenomics/workflow-is-cwl/blob/26dad276bac124f89086268bcbca962a5c0caca6/workflows/TranscriptomeAssembly-wf.paired-end.cwl>

¹⁷<https://github.com/EBI-Metagenomics/workflow-is-cwl/blob/26dad276bac124f89086268bcbca962a5c0caca6/workflows/TranscriptomeAssembly-wf.single-end.cwl>

¹⁸<https://view.commonwl.org/workflows/github.com/EBI-Metagenomics/workflow-is-cwl/blob/26dad276bac124f89086268bcbca962a5c0caca6/workflows/TranscriptomeAssembly-wf.paired-end.cwl>

¹⁹<https://github.com/EBI-Metagenomics/workflow-is-cwl/blob/26dad276bac124f89086268bcbca962a5c0caca6/workflows/TranscriptsAnnotation-i5only-wf.cwl>

²⁰<https://view.commonwl.org/workflows/github.com/EBI-Metagenomics/workflow-is-cwl/blob/26dad276bac124f89086268bcbca962a5c0caca6/workflows/TranscriptsAnnotation-i5only-wf.cwl>

²¹ <http://www.embric.eu/>

²² Johnson et al. *GigaScience* 8 (4). <https://doi.org/10.1093/gigascience/giy158>.

²³ <https://zenodo.org/record/257410#.WthptohuZaQ>

²⁴ Carradec et al. 2018. *Nature Communications* 9 (1): 373. <https://doi.org/10.1038/s41467-017-02342-1>.

example, the *Phaeocystis globosa* strain PCC64 transcriptome assembly does not fit in the expected transcriptome size known from literature (i.e. ~50,000 contigs)²⁵. The complete set of transcriptome assemblies in METdb shows a mean N50 value of 1,548 bp and a mean remapping rate of 84%. This highlights the overall high quality of the assembled transcriptomes that are included in the resource.

METdb includes data for 479 distinct marine eukaryotic strains that originate from 3 large-scale sequencing projects. The majority of the data (81%, 406 transcriptome assemblies) are representative for 4 taxonomic supergroups: Stramenopiles (35%), Alveolata (20%), Haptophyta (13%) and Viridiplantae (13%). Rare strains correspond to 5, 2 and 2 transcriptome assemblies from high taxonomic levels Euglenozoa, Heterolobosea and Glaucocystophyceae respectively in METdb. In addition, METdb includes 1 transcriptome assembly for an unclassified eukaryote.

The geographic distribution of the data shows that METdb have representative strains that originate from all 5 oceans. 103 samples were collected in the Pacific ocean while only one (*Ankylochrysis* sp., (strain MALI FT191.5 PG1) originated from the Arctic ocean. The inequalities of geographical sampling origins underline a lack of data for several world regions and may explain the poor knowledge that currently exists on the marine organisms as well as their ecology. Thus, we suggest that additional efforts are absolutely required to increase sampling experiments, thereby expanding our knowledge for the “unsampled and undersampled branches on the eukaryotic tree of life”^{26 27 28}.

8.5. Applications

Even though the resource is not yet publicly available, it is already used at Genoscope as main reference for a better taxonomic assignation of the next version of the *Tara* oceans derived eukaryotes unigenes catalog (Carradec et al. 2018). Moreover, as the data are submitted to ENA, our annotations will pass to other important databases such UniProtKB, via established data flows. Once there, we anticipate recourse such as Pfam to construct new entries to represent the novel functionality found in these sets. This will ensure the longevity and broad propagation of the data generated in this deliverable. This will be the first massive contribution of reference marine eukaryote data to the public protein sequence repositories, and as such it will be beneficial to the whole marine plankton biologist community.

8.6. METdb website

The METdb website²⁹ has been created using Django, a Python web application framework. It follows the model-view-controller (MVC) architectural pattern (see Figure 3). The model has been designed as an object oriented class hierarchy and mapped to a

²⁵ Koid et al. 2014.. *PLOS ONE*9 (6): e97801.<https://doi.org/10.1371/journal.pone.0097801>.

²⁶ Sibbald et al. *Nature Ecology & Evolution*1 (April): 0145.<https://doi.org/10.1038/s41559-017-0145>.

²⁷ Martin, C. 2015. *Current Biology*25 (8): R301–7.<https://doi.org/10.1016/j.cub.2015.03.010>.

²⁸ Hug et al. 2016. *Nature Microbiology*1 (5): 16048.<https://doi.org/10.1038/nmicrobiol.2016.48>.

²⁹ <http://metdb.sb-roscoff.fr/metdb/>

relational database (PostgreSQL) schema. The view layer relies on dynamic HTML pages generated with the Django templating engine as well as client-side technologies (JavaScript/AJAX, Bootstrap, DataTables, Highcharts). The controller layer, selecting the appropriate view based on user interaction and extracting view specific data from the database to inject it in the templates, is provided by Django.

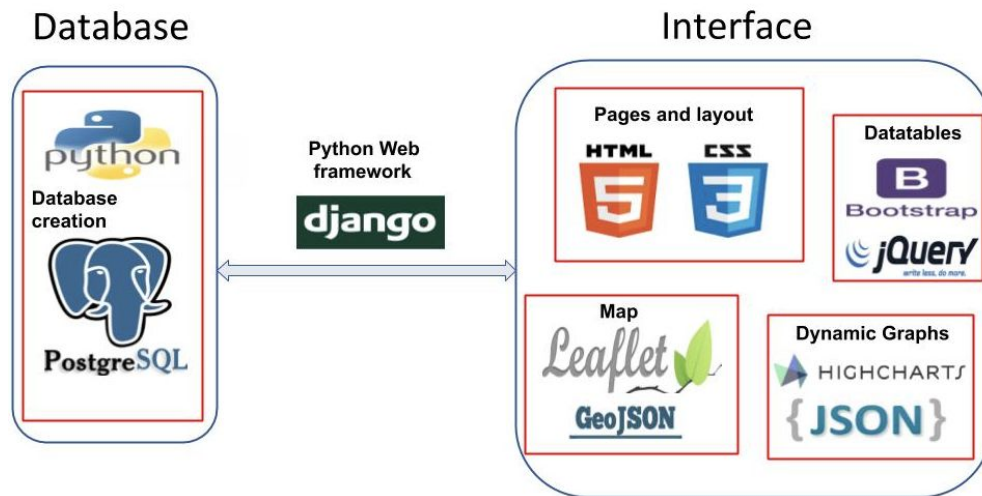


Figure 3 - Overview of the METdb architecture.

The METdb website offers the possibility for the user to explore the database by using a “simple” or “advanced” search function and select a subset of interest, based on any of the associated data be it taxonomy, function or geographic locations of a project. For a selected dataset, the user can access the summary reports for either the sequence reads or the assemblies, with cross-references to sequence data in external databases (e.g. ENA, NCBI taxID, WoRMS). The selected subset can then be downloaded with associated data (QC stats, metadata and annotation) and quality reports as a compressed archive. The global organization of the website in pages and components is shown in Figure 4.

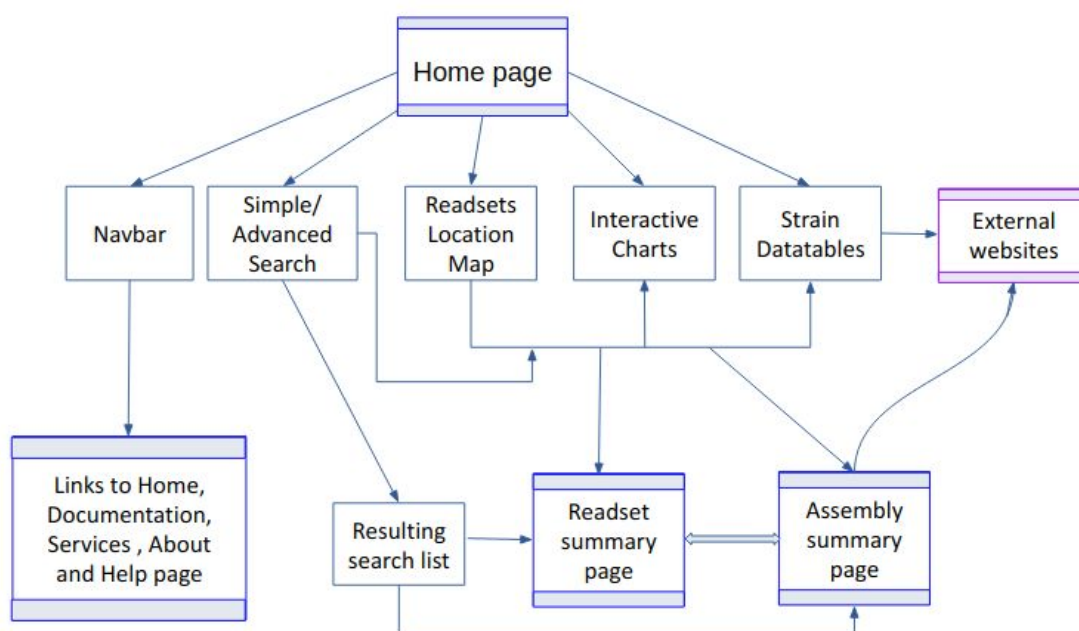




Figure 4 -Global organization of the METdb website. Key:  : website page,  home page main component.

The METdb web site home page provides a dynamic interactive overview of the taxonomic composition of the resource, a geographic representation of the isolation sites, as well as query services to access details about the current resource. Thus, the main user's entry points (geographic, taxonomic, functional) are covered. The final pieces of information the user can have access to are assembled transcriptomes, annotation results and metadata associated with the organism isolation and library preparation (Figure 5).

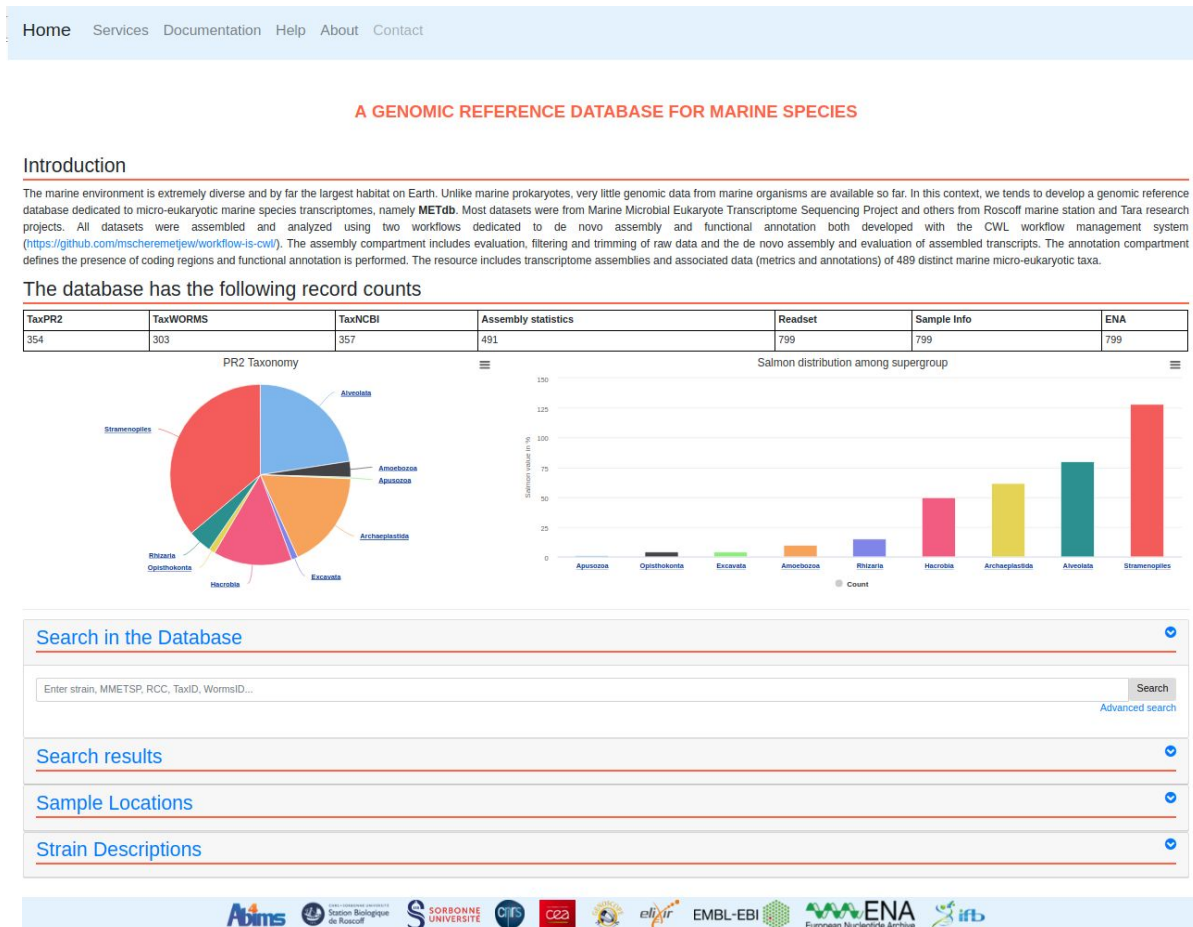


Figure 5 - Home page of the METdb website.

In the following sections, we outline some of the METdb website functionality:

Search module

Statistical interactive charts (Figure 6a), readsets geographic location sampling maps (Figure 6b), strain description tables and resulting datasets list are updated according to the search functions value (in the example below “Alveolata”). They also lead to raw sequence reads and an assembly summary page (Figure 6c).

Advanced search

| | | | |
|--|---|--|--|
| Supergroup: <input type="text" value="Alveolata"/> | Genus: <input type="text"/> | Species: <input type="text"/> | Strain: <input type="text"/> |
| NCBI taxonomy ID: <input type="text"/> | Worms AlphaId: <input type="text"/> | Sample alias: <input type="text"/> | Source: <input type="text"/> |
| Owner: <input type="text"/> | Worker: <input type="text"/> | | |

[Search](#)

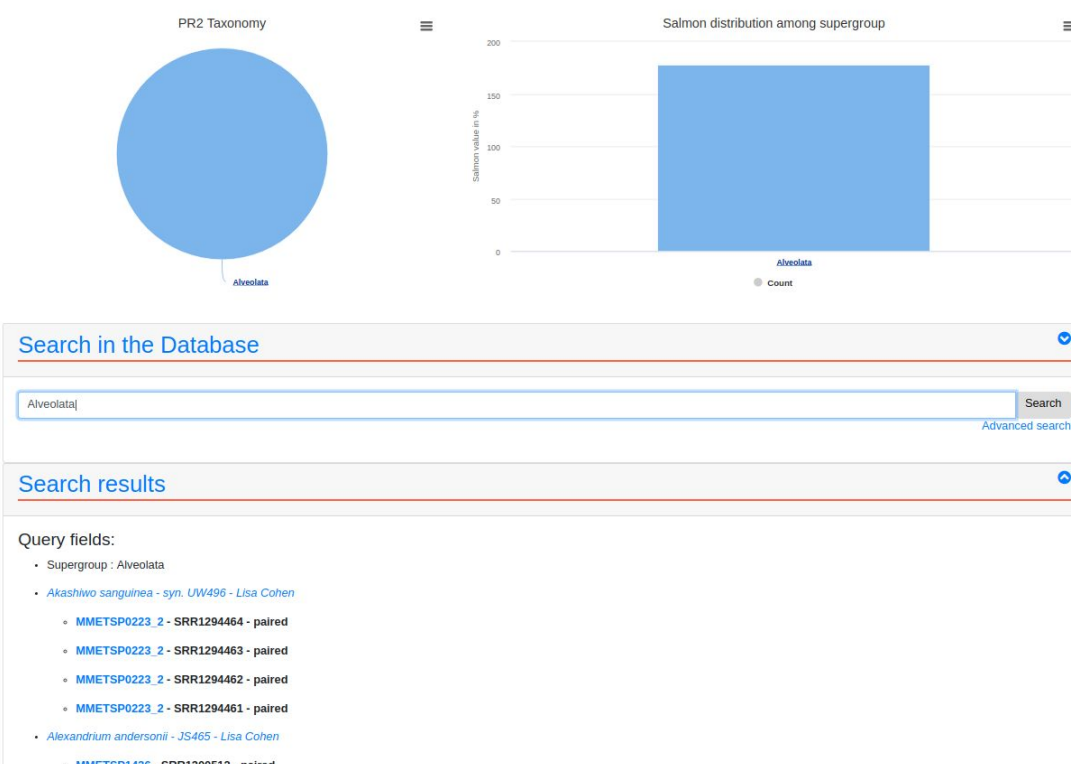


Figure 6a - Search results datasets list and statistical interactive charts.

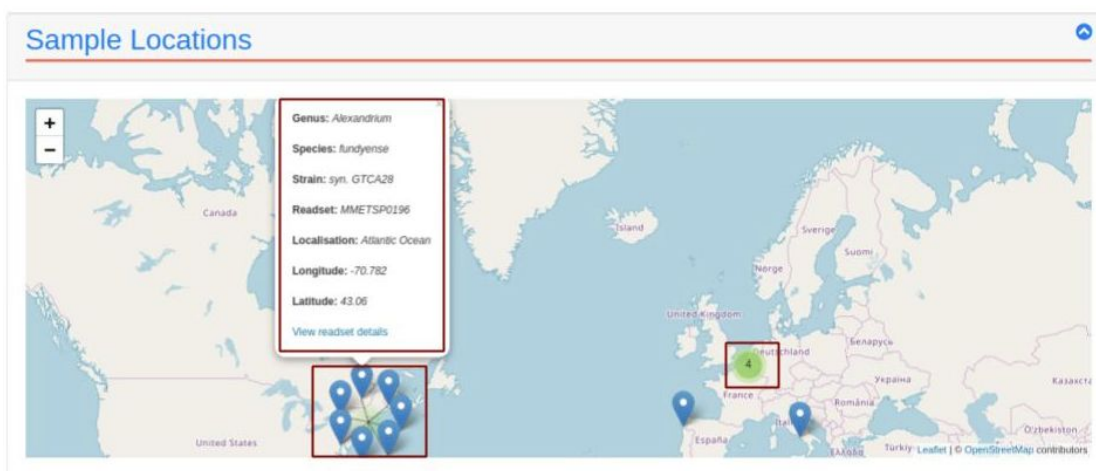


Figure 6b - Samples geographical localization.

| Strain Descriptions | | | | | | | | | |
|-------------------------------|----------------|------------------------------|---------------|---------------|-------------|------------|--|------------------------|------------------------|
| Show 10 entries | | Search: <input type="text"/> | | | | | | | |
| Details | Phylum | Class | Order | Family | Genus | Species | Strain | NCBI ID | WoRMS ID |
| Q | Dinoflagellata | Dinophyceae | Dinophyceae_X | Dinophyceae_X | Akashiwo | sanguinea | syn. UW496 | 143672 | 232546 |
| Q | Dinoflagellata | Dinophyceae | Dinophyceae_X | Dinophyceae_X | Alexandrium | andersonii | JS465 | 327968 | 246835 |
| Q | Dinoflagellata | Dinophyceae | Dinophyceae_X | Dinophyceae_X | Alexandrium | catenella | OF101 | 2925 | 231873 |
| Q | Dinoflagellata | Dinophyceae | Dinophyceae_X | Dinophyceae_X | Alexandrium | fundyense | syn. GTCA28 | 2932 | 231873 |
| Q | Dinoflagellata | Dinophyceae | Dinophyceae_X | Dinophyceae_X | Alexandrium | margalefi | AMGDE01CS-322 | 109239 | 233447 |
| Q | Dinoflagellata | Dinophyceae | Dinophyceae_X | Dinophyceae_X | Alexandrium | minutum | syn. AL | 39455 | 109711 |
| Q | Dinoflagellata | Dinophyceae | Dinophyceae_X | Dinophyceae_X | Alexandrium | monilatum | syn. AM01 | 311494 | 231875 |
| Q | Dinoflagellata | Dinophyceae | Dinophyceae_X | Dinophyceae_X | Alexandrium | temarensis | syn. CCMP115 | 2926 | 109714 |
| Q | Dinoflagellata | Syndiniales | Dino-Group-li | Dino-Group-li | Amoebophrya | sp | Ameob2 | 88552 | 109448 |
| Q | Dinoflagellata | Dinophyceae | Dinophyceae_X | Dinophyceae_X | Amphidinium | carterae | syn. AMPHI | 2961 | 109723 |
| Showing 1 to 10 of 98 entries | | | | | | | Previous 1 2 3 4 5 ... 10 Next | | |

Figure 6c - Strains description.

This part lists all available organisms matching the query, with links to external taxonomic references, and to the details of each assembly.

Description of the readset

This page (figure 7) displays taxonomy attributes corresponding to a readset, quality values, metrics and also associated result files related to each analysis performed on the readset.

| Readset Summary | |
|--|---|
| Taxonomy | |
| Metdbid | METDB_00008 |
| Genus | <i>Alexandrium</i> |
| Species | <i>monilatum</i> |
| Strain | syn. AM01 |
| Culture | CCMP3105 |
| Assembly | Link to associated assembly |
| Read Metrics | |
| MMETSP ID | MMETSP0097 |
| ENA BioProject accession identifier | PRJNA248394 |
| Analysis project type | Illumina HiSeq 2000 |
| ENA Run accession identifier | SRR1296898 |
| ENA BioSample accession identifier | SAMN02740408 |
| Number of bases | 3,357,981,400 |
| Average lenght | 50 |
| Maximum base quality | 38.5 |
| Minimum base quality | 31.7 |
| Percentage of remapping of library to assembly | 90.1 |
| Layout | paired |
| Status | combined |
| Associated result files | |
| Quality file | |
| Read quality file (fastqc) | SRR1296898_1_fastqc.zip |
| Readset remapping against assembly | |
| Quantification file (sf) | quant.sf |
| Metadata info (json) | meta_info.json |
| Command info (json) | cmd_info.json |

Figure 7- Readset description.

Description of the assembly

This page (Figure 8) summarizes information about an assembly. Each assembly page corresponds to a transcriptome of a studied strain. Taxonomy information with links to external databases (WoRMS, NCBI, ENA), assembly metrics (number of contigs, percentage of good remapping to reads, etc.), associated readsets used to perform the assembly with links to each readset summary page, resulting files for each analysis performed on the assembly available for download and a map of readset sampling locations are displayed in the page.

Assembly Summary


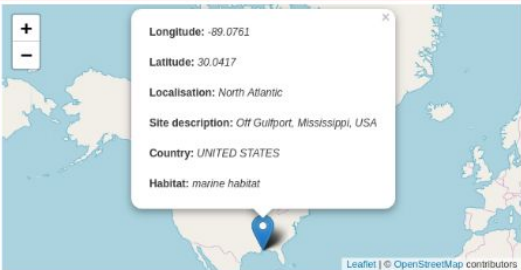
| Taxonomy | | Assembly metrics | |
|--|---|---|--------------------------------|
| Metdbid | METDB_00008 | Percentage of remapping of all readsets on transcriptom | 89.7 |
| Kingdom | Eukaryota | Number of contigs | 128,815 |
| Supergroup | Alveolata | Largest contig | 11,495 |
| Phylum | Dinoflagellata | Number of bases | 130,346,526 |
| Class | Dinophyceae | Contigs average lenght | 1011.9 |
| Order | Dinophyceae_X | Number of contigs with orf | 81,467 |
| Family | Dinophyceae_Xx | n50 | 1,523 |
| Genus | Alexandrium | GC | 0.66541 |
| Species | monilatum | Percentage of good mapping | 0.79056 |
| Strain | syn. AM01 | Transcriptome analysis worker | Arnaud Meng |
| TaxNCBI | 311494 | Associated readsets | |
| TaxWoRMS | 231875 | | |
| Associated result files | | Associated readsets | |
| Quality file | | Readset (paired) | MMETSP0097 - SRR1296898 |
| Read quality file | SRR1296895_1_fastqc.zip | Readset (paired) | MMETSP0096 - SRR1296897 |
| Reads dissimilarity evaluation | | Krona | |
| Abundance-based Bray-Curtis distance matrix | mat_abundance_braycurtis.csv.gz |  Click to open the krona chart. | |
| Presence-absence-based Bray-Curtis distance matrix | mat_presenceAbsence_braycurtis.csv.gz | Environmental parameters | |
| Reference Assembly | |  | |
| Contigs fasta file | MMETSP_alexandrium-monilatum-ccmp3105_trinity_v2.6.5_elixir.fasta | Longitude | -89.0761 |
| Assembly metrics | | Latitude | 30.0417 |
| Assembly metrics file (csv) | MMETSP_alexandrium-monilatum-ccmp3105_transrate-v1.0.2_elixir.csv | Site | North Atlantic |
| Contigs with good mapping (fasta) | good.trinity_Alexandrium-monilatum-CCMP3105.fasta.gz | Site description | Off Gulfport, Mississippi, USA |
| Contigs metrics (csv) | contigs.csv | Country | UNITED STATES |
| Contigs with bad mapping (fasta) | bad.trinity_Alexandrium-monilatum-CCMP3105.fasta.gz | Habitat | marine habitat |
| Reads remapping to assembly | | | |
| Metadata info (json) | MMETSP_alexandrium-monilatum-ccmp3105_salmon-v0.9.1_elixir.json | | |

Figure 8 - Assembly description.

8.6. Summary and Future plans

During the final few months of this work, we will finalize the assembly and annotation submissions to ENA, thereby providing the archival record of the data that can be found within METdb. This will allow the data to be propagated to other archive databases and

knowledge bases, using pre-established workflows, rather than requiring the import from METdb. We will finalise the manuscript describing the resource and make METdb publicly accessible (currently the site is password protected while the final data uploads are verified and we complete the website testing in different browsers).

METdb houses 489 RNA-seq datasets, representing 27 different phyla and 233 different genera. While this represents a major influx of data, only a small fraction of the micro-eukaryotes are captured in this resources. Thus, we recommend to the community to continue sampling this important group of microbes. As and when new dataset arise, our open-access workflows, which can be applied in a range of cloud environments, can be used by the community with the results potentially being important in METdb and submitted to ENA, without the data producers needing to develop their own pipelines. This work has also provided a consistent and reproducible dataset, which can be used as an important reference database for the interpretation of Marine Metagenomics data.