EXCELERATE Deliverable D6.3

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:	Report describing a set of tools, pipelines and search engine for interrogation of marine metagenomic data	
WP No.	6	
Lead Beneficiary:	1 - EMBL-EBI	
WP Title	Marine metagenomic infrastructure as a driver for research and industrial innovation	
Contractual delivery date:	31 August 2018	
Actual delivery date:	17 May 2019	
WP leader:	Rob Finn, EMBL-EBI Nils Peder Willassen, UiT	1 - EMBL-EBI; 23 - UiT
Partner(s) contributing to this deliverable:	1 - EMBL-EBI, 23 - UiT, 30 - CNR, 26 - CNRS	

Authors and Contributors:

Alex Mitchell, Maxim Scheremetjew, Gianluca De Moro, Bruno Fosso, Monica Santamaria, Graziano Pesole, Shriya Raj, Lars Ailo Bongo, Nils Peder Willassen, Robert D. Finn.

Reviewers:

None

1. Table of contents

Table of contents 2

Executive Summary 3

Impact 3

Project objectives 4

Delivery and schedule 4

Adjustments made 5

Background information 5

Appendix 1: Report describing a set of tools, pipelines and search engine for interrogation of marine metagenomic data. 9

8.1 Background 9

8.2 Overview and Status 10

8.2.1 Benchmarking Metagenomics Tools and Workflows 10

8.2.1.1 Amplicon benchmarking 10

8.2.1.2 Improving the detection of SSU rRNAs prior to classification 12

8.2.1.3 Extension of taxonomic profiling to eukaryotic organisms 12

8.2.1.4 Datasets for benchmarking shotgun pipelines 13

8.2.1.5 Semi-synthetic datasets for marine metagenomics pipeline assessment 13

8.2.1.6 Shotgun benchmarking results 13

8.2.2 Pipeline Reproducibility and Portability 15

8.2.2.1 Pipeline reproducibility 15

8.2.2.2 Pipeline cloud deployment 16

8.2.2.3 Cloud deployment of META-pipe 16

8.2.2.4 Cloud deployment of MGnify 16

8.2.2.5 Other Analysis Pipelines 18

8.2.3 Towards converging on a gold standard pipeline 19

8.2.3.1 Metagenome Assembled Genomes (MAGs) 20

8.2.3.2 Proteins for industrial biotechnology discovery 21

8.2.4 Training on resources for the Marine Community 22

8.2.4.1 Peer Review Articles 22

8.2.4.2 Webinars and videos 22

8.2.4.3 Online documentation and training guides 22

8.2.4.4 In person training, seminars and conferences 23

8.2.5 Increasing the discoverability of data 24

8.2.5.1 Faceted search 24

8.2.5.2 Federated searches across microservices 25

Horizon 2020 grant n. 676559

8.3 Summary & Future plans 27
8.4 References 29

2. Executive Summary

- Pipelines expanded to increase scope of analysis, tools updated and tailoring of analysis to meet the demands of the Marine Metagenomics community.
- Analysis pipelines described in the Common Workflow Language (CWL) to enable reproducibility, facilitate comparisons and enhance development agility.
- Benchmark datasets generated for evaluating amplicon analyses across a range of biomes and 6 different marine shotgun metagenomics datasets developed.
- Cloud deployment of pipelines to increase compute capacity.
- New data analysis outputs that will feed the databases developed as part of this work package.
- Examples generated for federating searches across resource programmatic interfaces.
- Wide range of different training activities and material developed.
- Major interaction with Compute (WP4) and Interoperability (WP5) platforms in sharing problems and driving solutions (e.g. AAI and CWL, respectively)

3. Impact

12 different training workshops in 7 different countries delivered to over 300 participants from both academic and industrial backgrounds

Pipelines enhanced and evaluated, that have shown improvements in recall of >70% on some datasets between version 3.0 and 4.1 of the MGnify pipeline.

New analysis datasets generated through increased scope which have provided: 46,000 marine and associated biome amplicon analyses, 3,931 assemblies from aquatic environments, 2,073 non-redundant metagenome assembled genomes, protein reference databases that exceed 600 million sequences. Much of this increase capacity has been achieved through collaborative work with the compute platform, which has enabled the deployment of workflows within different cloud environments to increase capacity for the community.

Example software for performing federated searches and expanded faceted searches.

8 publications have directly come from the activities described here (3 in Nucleic Acids Research, 2 in F1000 and one in each of Bioinformatics, GigaScience and Nature).

META-Pipe: We provide a set of tools, pipelines and search engine for interrogation of marine metagenomic data. Although we have not opened the service for all ELIXIR users due to computation resource restraints, META-pipe has been used to process 2616

Horizon 2020 grant n. 676559

datasets, with a total input size of 92GB and output 209GB, for 72 users. The MMP portal that provides the META-pipe service has been accessed by 2141 unique users.

Working with the Interoperability platform (WP5), we have provided extensive feedback based on the utility of CWL, which has improved definition framework and enhanced execution engines. Collectively between the two WP, our collaborative has driven the uptake of CWL as community standard for the reproducibility of scientific workflows.

The MGnify resource (previously called EBI Metagenomics) is a functionally rich portal encompassing metagenomics data archiving, standards compliance, functional and taxonomic analysis, facilitating data assembly, analysis, exploration and interpretation. MGnify provides cross-biome analysis, and includes over 30,000 publicly available datasets specifically from the marine environment. More broadly, we have over 200,000 datasets, representing 100s TB of input data. Our toolkit for accessing our programmatic interface has been downloaded 28,000 times since its release in April 2018. We have 2111 registered users (only required for pre-publication analysis), and approximately 1,000 unique IPs visit MGnify per month.

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

Horizon 2020 grant n. 676559

6. Adjustments made

The deliverable was completed on time, but the report slightly delayed because of major conflicting reports and the addition of two new deliverables that build upon some of the changing needs of the metagenomics community.

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:	1
Work package title	**Use Case A: Marine metagenomic infrastructure as driver for research and industrial innovation**		
Lead	Nils P Willassen (NO) and Rob Finn (EMBL-EBI)		
Participant number and person months per participant 1 – EMBL 28.00, 16 –FCG 2.00, 19 – CCMAR 11.00, 23 – UiT 36.00, 26 – CNRS 10.00, 30 – CNR 21.31			
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) Work Package Leads: Nils P Willassen (NO) and Rob Finn (EMBL-EBI)			
Description of work and role of partners			

Horizon 2020 grant n. 676559

WP6 - Use Case A: Marine metagenomic infrastructure as driver for research and industrial innovation

[Months: 1-48]

UiT, EMBL, FCG, CCMAR, CNRS, CNR

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

Horizon 2020 grant n. 676559

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 analysed. To improve the characterisation

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 particularly, we will use some of the higher-coverage and higher quality sequence outputs from the TaraOceans 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 centres, biotech

companies) to ensure usability for the end user community in order to improve 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- ELIXIR) 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: PT, NO

Horizon 2020 grant n. 676559

8. Appendix 1: Report describing a set of tools, pipelines and search engine for interrogation of marine metagenomic data.

8.1 Background

The World's oceans and seas represent the largest single biome on earth, comprising >97% of the world's total biosphere, from the tropics to the polar waters and from well-lit surface waters to the deep abyss. The health of this biosphere is essential to the future welfare and prosperity of humankind. This immense size is often underappreciated, as not only is 71% of the World's surface covered in water, but also the average depth of the ocean is over 2 km, with parts of the ocean extending to over 10 km in some regions (e.g. Mariana Trench). Approximately 50% of the atmospheric oxygen is produced from the microbes and macroalgae found in the oceans, with 50% of all biomass found in the oceans.

The inhabitants of marine ecosystems harvest and transduce solar energy, with an estimated contribution towards global primary productivity between 50% to 90% (1). They catalyse the key biogeochemical transformations of all nutrients and trace elements that sustain the organic productivity of the ocean. Marine microbes also play critical roles (both positively and negatively) in the aquaculture industry, such as causing susceptibility to pathogens in salmon farming which impacts yields. In 2012, 66.6 million tonnes of foodstuffs were produced from aquaculture sources, nearly equivalent to that arising from naturally occuring sources and is expected to supersede in the coming years. Marine microbes also produce and consume most greenhouse gases (carbon dioxide, nitrous oxide, and methane), which is of particular importance in the context of anthropogenic disturbance of marine ecosystems. Finally, they also represent a vast and dynamic reservoir of genetic variability that is yet to be exploited. Despite this, the rich marine microbial biodiversity is significantly underrepresented in biological databases and relatively few products derived from the sea have made it through to industrial production compared to those arising from terrestrial environments.

The field of metagenomics, whereby scientists analyse the sum of the genetic material found within a particular environment, has expanded substantially over the past decade. This growth reflects the scope of environments (e.g. marine, soil and human gut representing the most common biomes but expanded to include diverse environments, such as cow rumen, oil pipes and food production, to name a few), and organisms studied (e.g. predominantly bacteria but now viruses and microeukaryotes) and the range of experimental techniques applied (e.g. amplification of a single marker gene to deep shotgun sequencing using different DNA sequencing platforms). Due to the underlying experimental variations and the fact that there are hundreds of different tools that could be potentially used, there is an almost infinite number of permutations and combinations to produce analysis workflows. This is further complicated by the use of different reference

databases through which, many tools draw upon known data to make assertions about function or taxonomy.

Nevertheless, this increase in popularity stems not only from the fact that diminished sequencing costs means that these experiments are far more tractable, but also because the approach alleviates the need for culturing and is starting to provide access to the 99% of organisms that are yet to be isolated, cultured and sequenced. The use of metagenomic methods have been widely applied to the study of marine microbes, but the lack of a good reference database (and tooling) for the marine metagenomics community has restricted our understanding and their subsequent exploitation.

Within this deliverable D6.3, we focused on benchmarking tools and reference databases for the analysis of marine metagenomic datasets. At the outset of this project, the two main pipelines for analysing shotgun metagenomic datasets under consideration were META-pipe and MGnify (formerly called EBI Metagenomics), produced by UiT and EMBL-EBI, respectively. From this work, we developed benchmark datasets, evaluated new tools and applied iterative improvements to the pipelines that have been employed in this study. Overall, these pipelines have converged towards a general consensus, despite them fulfilling different roles within the research community. We also investigated ways to increase marine metagenomics analysis capacity through the use of cloud infrastructures. Finally, we generated an extensive training portfolio to help train new and update existing users on our metagenomics analysis pipelines.

8.2 Overview and Status

8.2.1 Benchmarking Metagenomics Tools and Workflows

8.2.1.1 Amplicon benchmarking

One of the crucial steps in almost all microbiome-based analyses is inference of community composition through taxonomic classification. For decades now, the most common approach for taxonomic assignment of microbial species has been through classification of ribosomal RNA (rRNA) sequences (2). Despite both the META-pipe and MGnify pipelines utilising the small subunit ribosomal RNA (SSU rRNA) for taxonomic assignment, they employ very different approaches. From our pilot work comparing these pipelines on a limited number of datasets, the performance of the different approaches compared to each other was inconclusive. For example, META-pipe with the LCAClassifier tool (3) and the SilvaMod reference databases generally identified more unique taxa for the marine sediment datasets, while MGnify, using the QIIME tool (4) (version 1.0) and Greengenes (5) as a reference database, identified more taxa for the faecal datasets. Furthermore, as the LCAClassifier generally offers resolution up to genus rank, we also observed that META-pipe was less likely to classify at species level compared to MGnify.

At the start of the project, the most widely used tools for this purpose were the mothur (6) and QIIME software packages (3, 4). Both tools take individual genetic markers (e.g. the 16S rRNA gene, conserved across the prokaryotic domains) and compare them to a reference database, assigning a taxonomic lineage to each of the queried sequences. Greengenes (5), NCBI (7), RDP (8) and SILVA (9) are some of the most widely used rRNA sequence databases. Similar to our pilot study, others have demonstrated that

classification using SILVA-based reference databases performs better than GreenGenes, particularly when applied to environmental sequences (3). However, the MGnify pipeline needs to accommodate analysis of datasets from any source biome. Ultimately, the success of these analyses is not only dependent on the breadth and diversity of annotated sequences available in public repositories, but also on the accuracy of the classification algorithms used by each of the tools. By default, QIIME makes use of the UCLUST clustering method (10) to assign biological sequences to a reference database, while mothur wraps the naïve Bayesian RDP classifier, developed by Wang, et al. (11), for sequence classification. Two other tools — MAPseq (12) and QIIME 2[1] have recently been released, providing alternative assignment methods. Within the scientific literature, tools are rarely published without demonstrating an improvement over pre-existing tools. However, the benchmark datasets and evaluation criteria are also not consistently applied.

For the purpose of this work, we wanted to independently determine which tool and reference database combination provided the best results against a range of biomes, while also factoring in the computational overhead. At the time of conducting this benchmark, neither the infrastructure nor the scope of our assessment (large reference databases, long calculation times) could not be accommodated as part of WP2.

In contrast to the use of mock communities, *in silico* benchmarking approaches provide an agnostic view on the efficiency of the computational pipelines, independent of experimental variation and technical biases. Therefore, in this work (13), we constructed a set of simulated 16S rRNA gene sequences representative of genera commonly found in the human gut, ocean and soil environments (using data already present within MGnify), to evaluate the accuracy of MAPseq, mothur, QIIME and QIIME 2 with different reference databases, namely Greengenes, NCBI, RDP, and SILVA. We also evaluated the use of the most commonly targeted subregions of the 16S rRNA gene to determine the relative impact on the performance of taxonomic assignment. Using this benchmarking dataset, we demonstrated that regardless of the database used, QIIME 2 outperformed all other tools in terms of overall recall at both genus and family levels, with QIIME 2 in conjunction with the SILVA database providing the superior tool and reference database combination. However, QIIME 2 was also the most computationally expensive tool, with CPU time and memory usage almost two and 30 times higher than MAPseq, respectively. Due to these overheads, QIIME 2 is less favourable to use in a production service when resources are limited. Since MAPseq showed the highest precision with miscall rates consistently below 2%, we recommended the use of SILVA and MAPseq for use in production (a recommendation that has since been adopted by both the MGnify and META-pipe pipelines).

The publication describing the details of this benchmark (13) provides a comprehensive guide for the marine metagenomics community when designing metagenomic experiments and analyses. Over and above the computational recommendations listed previously, one notable outcome was that the variable region 1 of the SSU rRNA is frequently truncated in reference databases severely limiting it use, while variable regions 3 and 4 typically performed the best.

[1] https://qiime2.org/

Horizon 2020 grant n. 676559

8.2.1.2 Improving the detection of SSU rRNAs prior to classification

While investigating the taxonomic assignment as described in the previous section, it remains essential to select the SSU rRNA sequences found within the larger corpus of sequences found in shotgun metagenomic datasets. Prior to this work, the tool rRNAselector (14) was used for the identification of bacterial SSU rRNAs. This tool bundled both the HMMER software tool (14–16) and profile HMMs used to represent bacterial SSU rRNAs in the forward and reverse direction, as the version of HMMER then being used only searched in the forward direction. During this work, we reworked the rRNAselector with *nhmmer* (17), the appropriate tool in the HMMER suite for searching nucleotides and SSU rRNAs from Rfam (18, 19), the non-coding RNA families database. During our benchmarking of these tools, we also used the corresponding co-variance models from Rfam and *Infernal* to validate the sequences that were being identified as having secondary structures in accordance with the models. This new SSU rRNA method improved the rates of detection by over 25% in a handful of extreme cases[2].

Overall, this section of work culminated in the MGnify analysis pipeline being updated to incorporate more sensitive models for the detection of the SSU rRNA marker gene. We also adopted MAPseq as our tool for comparing the identified SSU rRNAs to the reference database (which was also updated to use the latest version of the SILVA database). These modifications were optimised and integrated within our production pipelines (deployed in versions 4.0 and 4.1 of the MGnify pipeline) and published.

8.2.1.3 Extension of taxonomic profiling to eukaryotic organisms

While META-pipe offers eukaryotic classifications based on mitochondrial and plastid SSU rRNA sequences, the MGnify pipeline lacked the capacity to perform such eukaryotic taxonomic profiling at the outset of this project. Although the expanded SSU rRNA analysis described in the previous section provided the capability of identifying eukaryotic SSU rRNAs and their taxonomic classification, it is widely recognised that the SSU rRNA only provides a limited resolution as a taxonomic marker for eukaryotic sequences, with either the internal transcribed spacers (ITS) or large subunit (LSU) rRNA (also known as the 28S rRNA) being widely used alternatives. To address the need of accommodating these additional taxonomic marker genes/regions, the MGnify team introduced the capability of LSU rRNA taxonomic profiling in a manner similar to the SSU rRNA approach. This involved expanding the Rfam library to include the LSU rRNA models, the inclusion of the SILVA LSU rRNA reference database, as well as the pipeline software necessary to enable the profiling using the MAPseq tool. The inclusion of ITS analysis presents a somewhat complicated proposition. While amplification of either ITS1 or ITS2 does provide a sequence that can undergo classification, ITS detection is more complicated in shotgun metagenomic datasets: the flanking SSU, LSU and/or 5/5.8S rRNAs need to be identified and the rRNA gene(s) removed before the remaining sequence can undergo taxonomic classification. Within shotgun metagenomics, even merged paired end sequences are unlikely to provide much ITS sequence. Thus, further evaluation of different approaches for the integration of ITS is required, such as using fast read map approaches, but this is likely to suffer from a high false negative rate. As a compromise, the MGnify team have been investigating the development of a pipeline component for just the analysis of ITS1/2 amplicon data. In deliverable D6.2, we reported the improvements to ITSoneDB, a reference database for ITS1 (20, 21). This reference database, as well as the ITS1 and ITS2 databases provided by UNITE, are undergoing

[2] https://www.ebi.ac.uk/metagenomics/assemblies/ERZ376968

evaluation. We are also surveying which tools may be suitable for taxonomic inference using approaches beyond simple read mapping.

8.2.1.4 Datasets for benchmarking shotgun pipelines

In a complementary, parallel benchmarking effort, UAlg (ELIXIR-PT) developed a range of marine shotgun metagenomics benchmark datasets. These were submitted to MGnify and analysed with pipeline versions 3.0 and 4.1 representing those before and after the taxonomic assignment improvements. These datasets were not analysed by META-pipe but were also submitted to MG-RAST, another widely used analysis platform (22).

8.2.1.5 Semi-synthetic datasets for marine metagenomics pipeline assessment

To discover a “gold-standard” for metagenomic analysis, these tools and pipelines need to be evaluated against a set of “knowns” to determine their accuracy and performance. As marine metagenomic samples provide access to numerous novel and diverse microbial sequences, constructing realistic benchmarks, specifically for the marine sector, is difficult and time consuming, but nonetheless an essential component of the work presented. A set of six semi-synthetic marine metagenomes were created with a high number of taxa (representing diversity), sequencing errors, and unknown reads. Each of the metagenomes were constructed using genomics data from marine organisms with a full genome published in ENA. A significant proportion of the work involved the selection of organisms that can be considered “marine", which required manual verification and filtering of automatically selected genomes.

The metagenomes thus created contained genomic data from 82 eukaryotic (corresponding to 18 phyla, 74 genera), 365 prokaryotic (19 phyla, 217 genera), and viral organisms selected and mixed to simulate a real marine metagenome. For each prokaryotic organism no more than five were selected from a total of 541 strains. For every read used, the available functional and taxonomic information was captured - but equalling this annotation represents the ultimate benchmark test, as many of the functional and taxonomic assertions are made on assembled sequences. To simulate the complexity of a real dataset, an error profile was generated from a real marine metagenome and was used in the simulation of the reads. Additionally, a large number of shuffled reads were simulated and included in the metagenomes. All metagenomes were constructed using randomly selected sequences from the selected genomes, so the percentage of 16S rRNA and 18S rRNA could vary from organism to organism, but we ensured that all metagenomes contained some sequences from all the organisms. These datasets have been submitted to the ENA, and will be made publically available upon publication.

8.2.1.6 Shotgun benchmarking results

The six synthetic metagenomes created have different numbers of reads and different compositions. As such, only four of the six metagenome datasets have been submitted for analysis using two different versions of the MGnify pipeline (v. 3.0 and 4.1), and MG-RAST (V. 4.0.2). The datasets were also submitted to META-pipe but the metagenomes were not analysed. The two metagenomes that have not been analysed represent different benchmarks currently out of scope: (1) a smaller test metagenome designed to evaluate the run time of the single tools (so not appropriate here) and (2) a viral metagenome that is currently out of scope for all of the aforementioned pipelines.

Despite capturing the functional information for every sequence used in each metagenome, it was not possible to evaluate the precision of the functional annotation due to the difference in functional annotation methods. Furthermore, MGnify and MG-RAST use highly different tools and databases for the functional analysis, so their outputs cannot be directly compared. While all the four metagenomes were successfully analysed by MG-RAST, two of them gave unusable results from MGnify. In one metagenome, for example, the pipeline recognised only one phylum/genus in version 3.0 and four genera in version 4.1, while MG-RAST correctly identified more than 90% of the phyla/genera. The underlying reason appears to be the extremely low number of 16S/18S rRNA gene sequences in these datasets, highlighting the need for MGnify to consider inclusion of other approaches for taxonomic profiling.

The two metagenomes used for the comparison are each comprised of 25 million paired reads (80% sequences simulated from prokaryotic organisms, 4% sequences from eukaryotic and 16% from “shuffle” sequences). One of the metagenomes was generated using 100 bp long reads and the other one using 250 bp long reads. In these datasets, the same amount of sequences has been randomly selected from each prokaryotic and eukaryotic organism. To evaluate the taxonomic analysis between MGnify and MG-RAST, we calculated the percentage of phylum/genus recognised, percentage of false positives (reads classified in genera / phyla not present in the original data) and the accuracy. The accuracy was calculated as the average percentage error between the percentage of sequences observed (VO, percentage of sequences of a group on the total number of sequences with a result) and expected (VA , percentage of sequences of a group on the total number of sequences submitted) for a certain phylum or genus (|(VA- VO)/VA| * 100).

Overall, the results clearly demonstrate that MGnify version 4.0 recognised a bigger fraction of phyla than the previous one (100% of phyla recognised in the datasets, 14% more the previous version), but the number of false positive sequences did increase slightly (from 2.0% to 2.6%).

The average error (percentage of sequences per phylum on the total) is comparable between the two versions with an average value of 100%. MG-RAST recognised 100% of the phyla with a 0.55% false positive and a 30.48 error percentage . When we considered the identified genera we saw a 16.75% of improvement in correctly identified genera from version 3.0 to 4.1 of MGnify (89.75% correctly identified) and in this case, the number of false positives was considerably lower in version 4.1 (17% in 4.1, compared to ~120% in version 3.0 [caused by observed assignments far exceeding the number of actual observations]). Overall, the number of false positives decreased by about 60% in MGnify v4.1 and the percentage of error decreased from 92% to 42%. The results (and accuracy) were highly comparable to MG-RAST, with 87.6% of genera correctly assigned (2.1% less compared to MGnify), but the number of sequences assigned to an incorrect genus was slightly lower (8.24%).

Given the outcomes presented above, the best approach for the optimal shotgun analysis currently may be to use both pipelines and compare the results. While this is a suboptimal solution, in the longer term, the inclusion in MGnify of either a K-mer based mechanism for taxonomic profiling, or the use of protein matches against a reference database as is performed by MG-RAST, will alleviate this duplication of effort. Nevertheless, this comparison independently confirmed the work by EMBL-EBI and demonstrated that there

had been considerable improvement in the taxonomic profiling between versions 3.0 to 4.1 of the MGnify pipeline.

8.2.2 Pipeline Reproducibility and Portability

8.2.2.1 Pipeline reproducibility

From our work in deliverable D6.1, the broader metagenomics community highlighted the lack of standards surrounding the informatics[3]. To address this, and through collaboration with the Interoperability Platform (WP5), we extended deliverable D6.3 to ensure that our pipelines are described in a systematic fashion using the Common Workflow Language (CWL) standard, which increases the reproducibility and portability of the pipelines. CWL provides a mechanism for describing each tool in a workflow, including input files and parameters, software version and source and outputs files. Two or more tools descriptions can then be combined into a workflow, which in themselves can be combined to generate more complex workflows (see Figure 1). Having generated a CWL description of a workflow, this needs to be executed via a workflow execution engine.

The MGnify pipeline versions 3.0, 4.0 and 4.1 have each been described using the CWL, and these workflows were verified to perfectly replicate the corresponding bespoke workflow Python code[4]. While the CWL version of MGnify is close to production ready, it has not yet entered full production. The failure to enter full production with CWL is not an issue with CWL, but rather the lack of workflow execution engines that are sufficiently mature to encapsulate all areas of the CWL specification used in our workflows. We have worked with authors of community workflow execution engines such as Toil, to report issues and evaluate new versions. Recently, the MGnify/EMBL-EBI teams tested the latest versions of IBM Spectrum LSF (formerly IBM Platform LSF), which includes CWLEXEC that supports running CWL workflows on LSF. This involved working with IBM engineers, who fixed problems or added missing features. This period of testing has just completed and the MGnify team are now preparing to enter production concomitant with the release of version 5.0 of the MGnify pipeline.

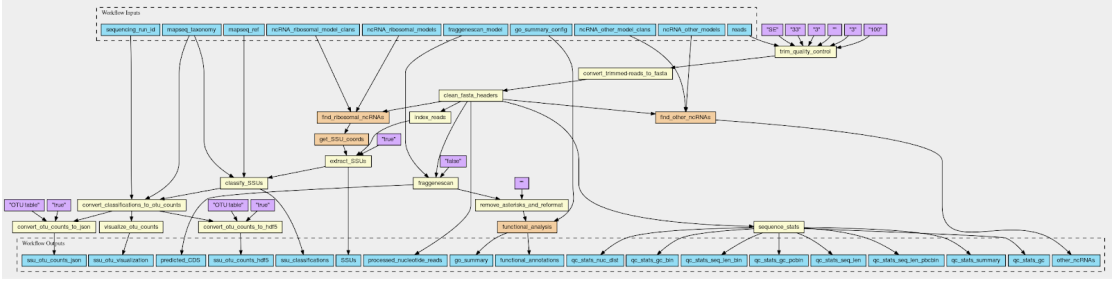

Figure 1 - Example of MGnify core pipeline rendered using the CWL viewer[5]. The boxes coloured blue represent inputs/outputs, purple parameters encoded in the CWL, orange sub-workflows and yellow individual tools within this workflow.

8.2.2.2 Pipeline cloud deployment

One of the typical bottlenecks of metagenomics analysis faced by the marine community is gaining access to sufficient compute resources. While sequence facilities are commonly

[3] https://paperpile.com/c/nBv5Ob/zazL
[4] see https://github.com/EBI-Metagenomics and various “CWL” label repositories therein
[5] https://view.commonwl.org/

Horizon 2020 grant n. 676559

accessible within departments at European research institutes, access to high performance compute (HPC) clusters within those institutes is more restricted. At the project outset, neither the META-pipe nor the MGnify pipelines had been deployed as production pipelines beyond the compute infrastructure that they were initially developed and deployed on. Consequently, there is always a tendency for developers to accidentally assume or incorporate certain characteristics of that local HPC cluster. Furthermore, and as indicated in the section describing the CWL associated activities above, pipelining software, third-party tools and reference databases are typically installed in central locations. In an effort to democratise access to metagenomics pipelines, we, together with members in WP4, have expanded the technical capacity surrounding the deployment of our pipelines to evaluate the ease with which pipelines can be “picked-up” and deployed in different cloud environments.

8.2.2.3 Cloud deployment of META-pipe

To improve analysis throughput on a distributed HPC architecture, the META-pipe analysis backend has been re-implemented using the Spark scalable data analysis software stack. Since this development, and with the aid of WP4, META-pipe has been deployed using a distributed architecture, with three central servers and geographically distributed execution managers. Currently, there are four META-pipe execution managers: (i) the Sigma2 Stallo supercomputer in Tromsø, a Norwegian academic HPC; (ii) the CSC cPouta OpenStack-based Infrastructure-as-a-Service cloud, Finland; (iii) the CESNET-MetaCloud OpenNebula cloud that supports the open cloud computing interface, Czech Republic; and (iv) the commercial Amazon EMR cloud service. cPouta is an ELIXIR compute service while CESNET-MetaCloud is part of the EGI Federated Cloud. The META-pipe Authorization service, which integrates the ELIXIR Authorization and Authentication Infrastructure (AAI), allows single sign-on to services across the ELIXIR infrastructure. We use the Authorization service to authorise access to data on the META-pipe storage system and jobs in the META-pipe job queue. The META-pipe Authorization server was among the first SAML2 service providers that integrated with ELIXIR AAI.

META-pipe will continue as a Galaxy service in the Norwegian e-Infrastructure for Life Sciences (NeLS). It was developed by ELIXIR Norway to provide its users with a system enabling data storage, sharing, and analysis in a project-oriented fashion. The system is available through easy-to-use web interfaces, including the Galaxy workbench for data analysis and workflow execution. Users confident with a command-line interface and programming may also access it through Secure Shell (SSH) and application programming interfaces (APIs). Although we have not opened the service for all ELIXIR users due to computation resource restraints, META-pipe has been used to process 2,616 datasets, with a total input size of 92 GB and output 209 GB, for 72 users.

8.2.2.4 Cloud deployment of MGnify

Again in collaboration with WP4 (specifically members of ELIXIR-EMBL), the MGnify team enabled the deployment of the corresponding analysis pipeline on the Embassy Cloud platform. This required extensive refactoring of the pipeline code to make it completely generic and agnostic to the installation environment, as well as the development of an Ansible playbook, which encapsulated all of the installation dependencies (software and databases) of the MGnify pipeline. Access to this deployment is mediated via the Cloud

Portal[6] using the ELIXIR AAI system. Once logged in, the portal enables a user to instantiate the MGnify analysis pipeline via a simple configuration interface and triggers the analysis of a particular dataset (see Figure 2). In the bottom panel of Figure 2, the user is requesting the analysis of the project ERP111373, and for the analysis to be run over 10 compute nodes. Once deployed, the first component fetches the sequence data from the European Nucleotide Archive (ENA). When executed within EMBL-EBI, the MGnify pipeline can fetch data via a filesystem copy; within the cloud setting this was not possible however. Thus, this initial data fetching module had to be rewritten to enable an alternative mechanism for data retrieval.

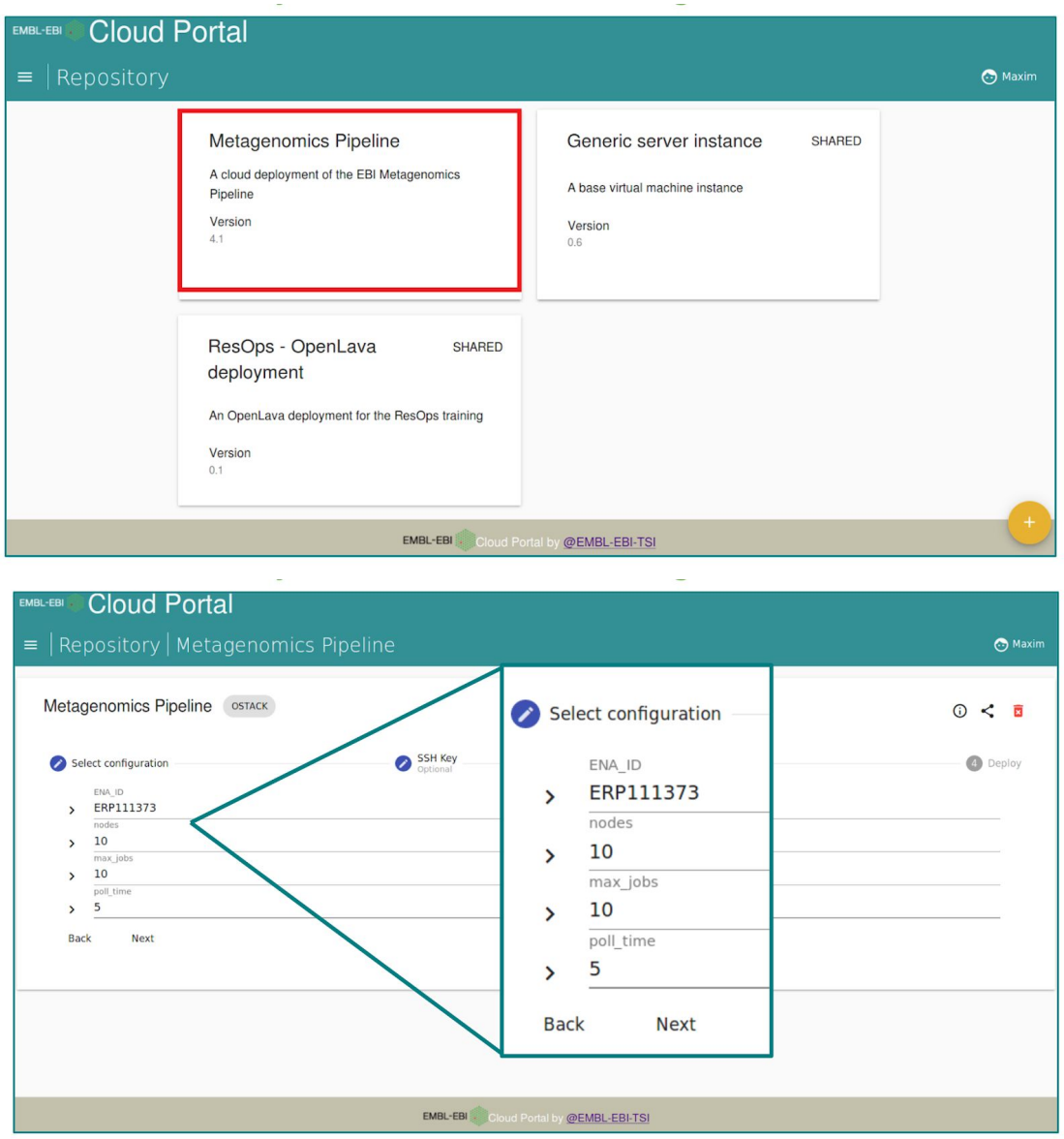

Figure 2 - Top panel shows the repository of applications available to the authenticated user ("Maxim"). The bottom configuration panel is for the metagenomics application, which facilitates virtual machine configuration, the analysis that needs to be performed, and SSH key hosting.

We achieved a major milestone through the cloud deployment and the execution of the pipeline on cloud infrastructure . Furthermore, we optimised the compute resource (CPU + RAM) assignments for individual analysis tasks (e.g. amplicon or shotgun) so as to avoid

[6] https://cloud-portal.ebi.ac.uk/

Horizon 2020 grant n. 676559

wasting compute resource as well as increase the analysis throughput. As cloud compute resources were limited within the ELIXIR Excelerate project, we decided to focus primarily on amplicon analysis of datasets associated with the marine environment that were yet to be analysed by MGnify. This represented a backlog of ~46,000 datasets. Despite the conservative and generic configuration of the virtual machines executing the annotation pipeline, with the initial deployment, we managed to attain an initial throughput of just over 1,000 amplicon datasets per day by running the procedure in an automated manner. Through our optimisations, we increased this throughout to over 4-fold using the same volume of compute resources, which enabled us to rapidly clear the current backlog of public marine datasets.

This has clearly demonstrated the potential of using Cloud resources to rapidly process metagenomic amplicon datasets held within ENA. We also evaluated the annotation of other metagenomic datasets (such as shotgun and metatranscriptomic data) on the same Cloud Portal, obtaining similarly encouraging results, albeit it at lower rates of throughput due to their significantly greater computational overhead.

Despite the success of scaling up the throughput on metagenomics analysis of Cloud compute, it nevertheless introduces new challenges. Of particular note are data ingress and egress from the Cloud compute infrastructures, as well as the lack of connection to the MGnify pipeline monitoring database (essential when dealing with such large numbers of dataset). We overcame the data transfer issues by developing a new component for data ingress using an FTP pull first, while the bulk transfer of results back to the EMBL-EBI filesystem was achieved via a direct secure copy to the EMBL-EBI filesystem, followed by unpackaging and upload to the MGnify database.

Other aspects of Cloud infrastructure also need to be enhanced, including autoscaling of virtual machines horizontally based on the number of submitted jobs and requesting appropriate CPU and memory resources to ensure successful execution. The latter was crudely tackled by having specific resource requirements for different experiment types (e.g. shotgun or amplicon analysis); however, even within an analysis pipeline different stages would have different resource requirements. This would require greater orchestration of virtual machines based on requirements, a feature that is too advanced for our current implementation or resources. Other areas of resource optimisation could be achieved through the application of machine learning methods to predict job requirements based on input file sizes and associated metadata, as has been prototyped for the assembly pipeline.

While we accomplished the processing of over 46,000 marine amplicon datasets on Embassy Cloud, it took nearly six months to complete the upload to MGnify (manual curation of biome for the upload process being the major bottleneck). Now, MGnify is approaching a near comprehensive set of analysed Marine amplicons, facilitating the assessment of the World's oceans' current biodiversity.

8.2.2.5 Other Analysis Pipelines

CNR worked to upgrade the reference databases available in the BioMaS (Bioinformatic analysis of Metagenomic ampliconS) pipeline (24). BioMaS encapsulates all the required steps for the analysis of metabarcoding data, from evaluation of next generation sequencing (NGS) raw data to sequence taxonomic classification. This platform already incorporated the reference databases RDP II and ITSoneDB, which are upgraded to the

last available releases (11.5 and 1.138, respectively). The Greengenes reference database has not been updated since 2013 and the last available version is already deployed in BioMaS. Based on the benchmarking results and to ensure harmonisation across the pipelines produced as part of WP6, the CNR team are working towards the inclusion of the SILVA database (version 132). BioMaS has been deployed in the ITSoneDB workbench (ITSoneWB) based on Galaxy[7] leveraging on ITSoneDB as reference database (see Task 6.2). Some of the described activities concerning the ITSoneWB are included in the Elixir 2017 Implementation Study for Integration of Italian Node entitled "A web service supporting ITS1 based survey of marine communities". Furthermore, it is possible to directly query ITSoneDB and import the resulting data in Galaxy through the specific tool integrated in the Galaxy workbench. Also QIIME (4, 23) and mothur (6) pipelines have been integrated within the ITSoneWB. Moreover, the CNR team has developed MetaShot (25) , a workflow for the taxonomic profiling of host-associated microbiomes. It implements a multistep procedure including all the required steps to manage and analyze Illumina shotgun paired end (PE) reads. MetaShot was demonstrated to outperform both Kraken (26) and MetaPhlAn2 (27) by using an in-silico generated- and a mock- microbial community. CNR team is working on a new MetaShot release designed for the analysis of data produced by IonTorrent/IonProton platforms too.

8.2.3 Towards converging on a gold standard pipeline

As highlighted in the introduction the field of metagenomics is changing very rapidly, both in terms of the range of biomes sampled and the type of experimental analyses. Nevertheless, both the META-Pipe and MGnify pipelines have many commonalities. For example, the META-pipe pipeline was upgraded to include a new 16S rRNA taxonomic classification component based on MAPseq and SIIVA (*i.e.* the same as MGnify). As was highlighted by the shotgun benchmarking, pipelines should also consider incorporating other forms of taxonomic profiling. Such a profiling approach was added to META-pipe using a peptide-based classification. Unlike MGnify (which is restricted by the need to keep broad appeal), the META-pipe workflow has also incorporated reference databases based on the MAR databases described in deliverable D6.2 e.g. SILVAmar, a 16S rRNA database based on the manually curated MarRef database, KajuMar, a resource for taxonomic classification using sequencing reads. Functional assignment of marine samples has been improved by implementing MarRef. Use of the MarRef sequence database, which consists of about 35 million marine CDSs, gives better functional assignment compared to existing databases e.g. RefSeq.

The tools and approaches to metagenomics have changed significantly since the EXCELERATE project was formulated. For instance, one emerging pipeline has been the assembly and binning of datasets to produce metagenome assembled genomes (MAGs). While assembly was always offered as part of META-pipe, this was not offered in MGnify, and neither resource provided the capacity for generating metagenomics bins nor MAGs.

The MGnify range of services has now been extended to include assembly of shotgun metagenomic samples using MetaSPAdes as the standard assembly tool. This is in contrast to META-Pipe which utilises MEGAHIT. Both have their advantages and disadvantages, with MGnify using MEGAHIT as an alternative assembly method for very diverse environments and/or extremely large datasets. For example, 204 Tara Oceans

[7] http://itsonewb.cloud.ba.infn.it/galaxy

datasets were assembled with MetaSPAdes, while 43 required more than 2 TB of memory using MetaSPAdes (the largest machine accessible to the MGnify team), thereby assembled with MEGAHIT instead. Using this strategy for assembly, the MGnify team has now applied this pipeline to 3,931 shotgun metagenomics datasets from aquatic environments, of which 2,505 have passed our assembly quality control criteria and 90% have now been uploaded to the ENA and are publicly available. Once uploaded, these assemblies were analysed by MGnify, e.g. EMOSE[8] and Tara Oceans[9].

8.2.3.1 Metagenome Assembled Genomes (MAGs)

While not a community service, the MGnify team has investigated the use of various binning tools for the grouping of related contigs into sets that are believed to have originated from a single genome, guided by the CAMI benchmarking results (28). For the following work, we used MetaBAT 2.0 (29, 30) for binning, followed by an estimation of completeness and contamination by CheckM (31). This nascent pipeline was also used to identify novel bacteria in the human gut (29).

Having applied the binning and quality control estimations to the set of assemblies from aquatic biomes, we have been able to bin each sample and dereplicate the dataset to produce a set of 2,073 MAGs (Figure 3). Note, 1,983 of these MAGs are novel (i.e. no equivalent species found in public databases, regardless of sample source), with 60% of them estimated to be high-quality, near complete genome assemblies. Only 10% of these genomes can be placed taxonomically within a known genus, while only half can be placed taxonomically in a known family, highlighting the huge wealth of unknown genetic diversity found in marine environments.

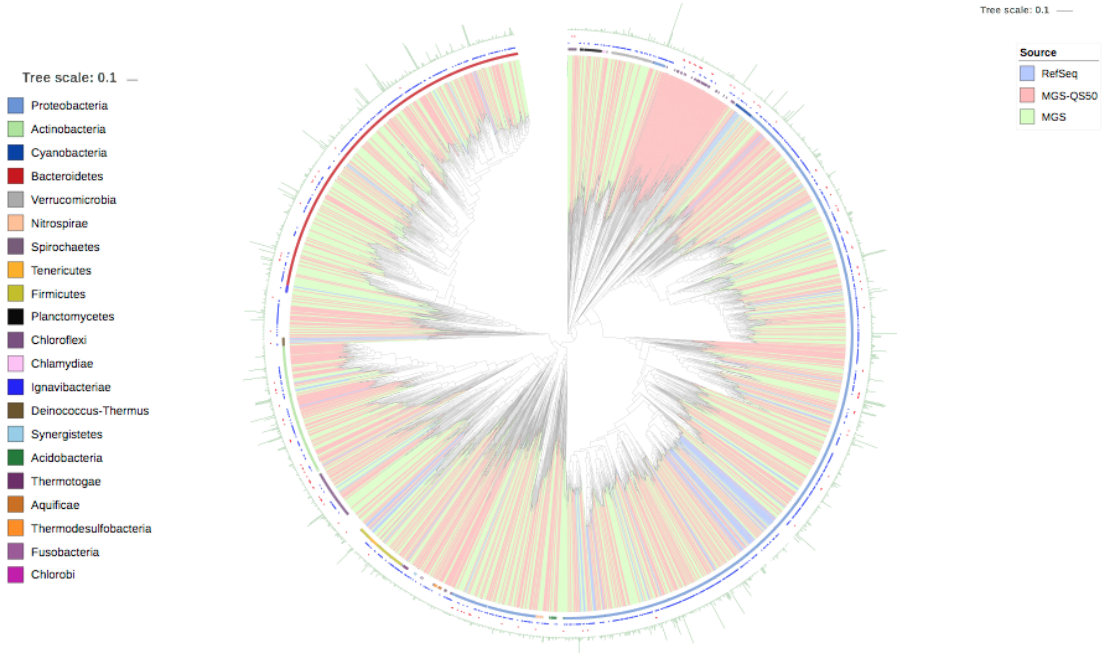

Figure 3 - The distribution of the MAGs generated by the new MGnify assembly pipeline. The phylogenetic tree is based on the SpecI marker genes. The colours radiating from the tree indicate

[8] https://www.ebi.ac.uk/metagenomics/studies/ERP112966
[9] https://www.ebi.ac.uk/metagenomics/studies/ERP104174

Horizon 2020 grant n. 676559

isolate genomes (blue, e.g. RefSeq/MarRef/MarDB), high-quality near complete (green, >90% completeness <5% contamination) and medium quality (red, QS50 = % complete - 5x contamination). The next two rings indicate whether the MAG was found in a marine sample (dark blue) and/or fresh water sample(s) (red). The outer ring, (green columns), indicates the number of samples where that MAG was found.

These MAGs are yet to be made publicly available as the ENA needed to develop new sets of categories to allow the deposition of metagenomic assemblies, binned assemblies and representative sets. This infrastructural work, which is not part of EXCELERATE, has now been completed with the MGnify team embarking on their submission of ~3,400 assembled datasets from aquatic environments to the ENA. Once deposited, they will be functionally analysed within MGnify to provide an important Marine reference catalogue. This dataset can be subsequently merged into the MAR databases, reported as part of D6.2.

8.2.3.2 Proteins for industrial biotechnology discovery

One key advantage of metagenomic assemblies is that it provides access to longer contigs, which in turn present two important features: (1) the ability to identify full length proteins; (2) the genomic context of those proteins, e.g. single operon, which may allow identification of associated functions or the inference of a complete pathway. Based on our interactions with the biotechnology sector, we know that having access to complete protein sequences is crucial for commercial exploitation.

The MGnify team has been developing mechanisms of generating non-redundant sets of proteins from their assemblies. In the last build of the MGnify protein database, version 2018_12, there were 1.2 billion non-redundant proteins[10]. Of these, 134 million originated from a marine environment;, 37 million of these were full length. At the time of building this protein database, some of the large marine projects, such as EMOSE and Tara Oceans had not been incorporated. We expect the number of marine proteins to increase dramatically once these are incorporated into that protein dataset. The Tara Oceans assemblies alone encapsulate 563 million unique sequences, of which there are 114 million that are full length. Currently, <1% of these sequences is represented in typical reference databases, such as UniProt (which contains ~150 million sequences) or RefSeq. Previous studies have produced a marine gene catalogue of about 40 million bacterial sequences (32). Other projects, such as OSD, have around 7 million proteins predicted from their assemblies. Amalgamating all of the marine assemblies, we now estimate that previous gene catalogues may have significantly underestimated the bacterial diversity, as clustering our unified marine dataset using LinClust (33) has yielded *256 million clusters*.

We anticipate that this new marine protein dataset will provide an important new reference set for Marine metagenomics analysis, especially for those interested in mapping functions to different environments and researchers undertaking novel enzyme discoveries.

8.2.4 Training on resources for the Marine Community

In collaboration with various members of the the Training Platform (WP11), we have undertaken a wide range of training activities, and developed a broad collection of

[10] ftp://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/

materials using a variety of media. In the following section we provide a range of examples, rather than an exhaustive list.

8.2.4.1 Peer Review Articles

Via articles published in F1000, we described how the META-pipe analysis service is integrated with ELIXIR AAI to provide user-friendly single sign-on for ELIXIR users[11], and how the analysis are run on an ELIXIR cloud platform chosen by the user[12].

8.2.4.2 Webinars and videos

We provided webinars on how to use META-pipe[13] and MGnify[14], and presented workshops on how to use META-pipe[15]. Table 1 provides an extensive list of the online videos/materials produced from the workshops.

8.2.4.3 Online documentation and training guides

Additionally, the Marine Metagenomics Portal has a documentation[16] and a community page[17] that provides further information to end-users.

In an effort to improve the understanding of our pipelines, we migrated all of the MGnify documentation to ReadTheDocs[18]. This provides a wide range of media to be incorporated and presented. It also makes the materials more readily searchable, is able to provide versions of the documentation that match the different pipeline versions, and can export the documentation in formats such as PDF, enabling users to download locally. These user guides complement our existing EMBL-EBI Train Online tutorials and webinars that cover the website[19], data submission process[20] and analyses, thus adding greater contextual background and theory.

8.2.4.4 In person training, seminars and conferences

To help provide training to the scientific community, we held 12 different training workshops in seven different countries (Table 1). To broaden the reach of these workshops, all of them were associated with training materials of some kind, providing an opportunity to those unable to attend the course to still benefit from it. For example, the four-day Metagenomics Bioinformatics hands-on workshop at EMBL-EBI only reaches a limited audience of 30 attendees. However, in 2018, 13 lectures from the course were videoed and developed into a virtual training course. These are currently being reviewed prior to public release at the end of May. All of the hands-on practicals were captured and a virtual machine image generated that captured the software necessary to carry out the practicals.

MGnify has also featured in a number of other training courses and marine-related outreach activities, including the Microbial Metagenomics: A 360º Approach workshop in Heidelberg, (Germany) the Marine Microbes Gordon Research Conference in Lucca

[11] https://paperpile.com/c/nBv5Ob/J2aI
[12] https://paperpile.com/c/nBv5Ob/kmei
[13] https://www.youtube.com/watch?v=uSsvIZhY8Hs&feature=youtu.be; https://www.youtube.com/watch?v=42cNWSmle4E
[14] https://www.ebi.ac.uk/training/online/course/ebi-metagenomics-analysing-and-exploring-metagenomics-data
[15] Videos are available at: https://www.youtube.com/playlist?list=PLjiXAZO27elBa5zGKCpwvRXxx-kF52Iuf
[16] https://mmp.sfb.uit.no/documentation/#/
[17] https://mmp.sfb.uit.no/community/
[18] https://emg-docs.readthedocs.io/en/latest/
[19] https://www.ebi.ac.uk/training/online/course/ebi-metagenomics-portal-quick-tour
[20] https://www.ebi.ac.uk/training/online/course/ebi-metagenomics-portal-submitting-metagenomics-da

(Italy) and the Ramon Margalef Summer Colloquia at the Institute of Marine Sciences, Barcelona (Spain).

Table 1 - list of training workshops, hosting city/country, dates and associated materials.

Workshop	Location	Date	Link
GOBLET/ELIXIR workshop for metagenomics training materials re-use.	Hinxton, UK	07 - 08 April 2016	https://www.elixir-europe.org/events/gobletelixir-workshop-metagenomics-training-materials-re-use
Summer School 2016 in Metagenomics	Paris, France	12 - 16 Sep 2016	http://www.france-bioinformatique.fr/en/evenements/summer_school_metagenomics
Metagenomics Bioinformatics	Hinxton, UK	13 - 20 Oct 2016	https://www.ebi.ac.uk/training/events/2016/metagenomics-bioinformatics-1
Metagenomics data analysis	Helsinki, Finland	03 - 06 April 2017	https://www.csc.fi/web/training/-/metagenomics
Workshop on Computational Metagenomics: Methods, Standards and Experimental Procedures	Bari, Italy	19 - 20 June 2017	https://elixir-iib-training.github.io/website/2017/06/19/metagenomics-workshop-and-school-bari.html
Summer School in Advanced Computational Metagenomics	Bari, Italy	21 - 23 June 2017	https://elixir-iib-training.github.io/website/2017/06/19/metagenomics-workshop-and-school-bari.html
Metagenomics Bioinformatics	Hinxton, UK	2-5 Oct 2017	https://www.ebi.ac.uk/training/events/2017/metagenomics-bioinformatics-2
Euromarine Open Science Exploration (EMOSE)	Porto, Portugal	11-15th Sep 2017	https://www.euromarinenetwork.eu/EMOSE
First Marine Microbiome workshop - Metagenomics and Bioinformatics for Biodiversity	Nador, Morocco	05 - 09 Feb 2018	http://medicalintelligence.org/marmicrobiome2018/
Elixir-Excelerate Workshop on Marine Metagenomics	Oreiras, Portugal	7 - 11 May 2018	http://elixir-portugal.org/event/elixir-excelerate-workshop-marine-metagenomics
Metagenomics Bioinformatics	Hinxton, UK	17 20 July, 2018	https://www.ebi.ac.uk/training/events/2018/metagenomics-bioinformatics-3
Hands-on workshop in Marine Metagenomics	Tromsø, Norway	26-30 Nov, 2018	https://elixir.mf.uni-lj.si/enrol/index.php?id=43

Horizon 2020 grant n. 676559

8.2.5 Increasing the discoverability of data

Another key task in this deliverable was to increase the discovery of the Marine metagenomic data and their associated analyses. Below we describe the improvements to the websites for discovering datasets, and how the different services provided by Marine metagenomics may be combined to address a range of research questions.

8.2.5.1 Faceted search

With the rapid expansion in the number of datasets (from a few 1000s to over 200,000), it has become increasingly important to improve access to the data contained within MGnify to facilitate exploration and discovery. To this end, we have exposed all of the sample metadata and analysis summaries within the EBI search (36), which provides a search infrastructure for enabling simple faceted searches, and implemented within the context of the MGnify resource. A search input box is present on all pages, allowing entry of free text (e.g. 'human') or colon-separated fields and values (e.g. 'experiment_type:amplicon'). Searches are subdivided into three levels: projects, samples and runs, as each level has different metadata and analysis results available. The results are displayed in separate tabs and can be filtered by facets and/or numerical search controls, as appropriate for the data type. For example, run-level has the richest set of indexed facets that can be used for filtering, with Organism, GO-terms and InterPro annotations. The latter two can also be used as search terms, and the results can then be filtered by fields such as temperature or depth. Using this search interface, it is possible to narrow down datasets rapidly and easily (for example, to discover all runs that contain antibiotic biosynthesis monooxygenase sequences in soil, where Actinobacteria are found, determined using metatranscriptomics). While we anticipate most users to access this via the website, the search is actually implemented as an API (application programmatic interface), a service which is called by the website. However, this API is exposed publicly, allowing users to access the service via software.

8.2.5.2 Federated searches across microservices

In addition to the search described above, MGnify has developed an entirely new website (not part of the Work Package 6 efforts). This website is backed by a comprehensive API using OpenAPI, a recognised standard for producing RESTful compliant APIs. This exposes all of the data that has been generated in MGnify, namely the assembled Marine metagenomics datasets, taxonomic assignments and functional analyses. Querying against the MGnify API allows users to start investigating functional properties alongside sample metadata such as temperature and depth. We have also deployed a sequence similarity search[21], which allows access to the proteins identified within the metagenomic assemblies. This search just provides access to the 280 million cluster representatives from the near 1 billion, non-redundant sequence set housed in MGnify. Again, this services has an associated API, and both are more extensively described here (37).

Additional APIs are also provided by other resources, such as ENA, which provides access to more sequence metadata that may be available in MGnify and datasets that are publicly available but are yet to be processed by MGnify. Making a single resource that amalgamates all of the content would be impractical due to the volumes of the data, the

[21] not funded as part of this project, https://www.ebi.ac.uk/metagenomics/sequence-search/search/phmmer

various data release procedures and the lack of a single unifying output format. However, it is far better for users to query across multiple APIs in order to retrieve the specific data items they are interested in. While all of the aforementioned APIs are well documented and describe their uses and the endpoints, it can be often be difficult for users to identify the end-point crosstalk between two different APIs.

To help users illustrate the different ways that our APIs could be combined or federated, we asked the wider Marine Community to pose some questions that they would like to ask of the data. From these responses, we have developed a portfolio of “use cases” that demonstrate how one would access the API, including a summary of the services used, and code examples of how this could be achieved and the expected outputs. For example, “*Retrieve all related sulfatase sequences sequences in all assembled marine samples stored in EMBL-EBI MGnify platform”*. In this scenario, the user wishes to slice across many datasets looking for a specific enzyme - a typical query from an industrial biotechnology company. The outputs requested were (i) a file containing all of the sequences in FASTA format and a (ii) zoomable map of occurence. In this example, this required using the MGnify API and the Google map view. Figure 4 provides a condensed version of the code snippet. In another example, we looked at retrieving functional assertions and sample metadata to look for correlations (see Figure 5).

Horizon 2020 grant n. 676559

```perl
#!/usr/bin/env perl
#
# An example of how to access the MGnify API and use it to ask the following question
#
# Retrieve all the protein sequences of a specific gene (in this example sulfatases)
# in all marine samples stored in EBI MGnify platform. This will just look at assembled
# sequences. All sequences will be stored in an FASTA file and the locations will be
# stored in a CSV file, to allow them to be easily uploaded to the Google API.

use strict;
use warnings;
use JSON::API;
use IO::Zlib;
use DDP;
use Getopt::Long;

#First step is to get a list of all projects/samples for the biome of interest.
#Filtered according to coastal marine and being metagenomics/metatranscriptomic
my $biome = "root:Environmental:Aquatic:Marine";
my $expt = "assembly";
my $iprOfInterest = "IPR006124";
my $hmm = "PF01676.hmm"; #The hmm corresponding to the IPR

#User can overrule the above defaults.
GetOptions( "biome=s" => \$biome,
   "expt=s" => \$expt,
   "ipr=s" => \$iprOfInterest,
   "hmm=s" => \$hmm ) or die "Could not get options\n";

#Set up the MGnify JSON API connection
my $api = JSON::API->new( 'https://www.ebi.ac.uk/metagenomics/api/v1',
      agent => "mgnifyscan/v0.1",
      protocols_allowed => [ qw/https/ ],
      env_proxy => 1,
      ssl_opts => { verify_hostname => 0 });

#Samples - this slightly cheats and gets all samples by upping the page size. Should check links
my $samples = $api->get("biomes/$biome/samples?experiment_type=$expt&page_size=1000");
#We are going to store all of the fasta files in here.
my @downloads;
#Sample metadata of positively matching sequences go in here.
my $storedMetaData;

#Iterate through the response object and find the assemblies that have matches to the IPR of interest.
#We first look in the summary of results before grabbing the potentially large fasta file.
foreach my $s (@{$samples->{data}}){
 ..
}
```

Figure 4 - Top - reduced version of code snippet indicating how to access the different APIs. Bottom - Distribution of marine samples containing sulfutases based on the output of this map, with the intensity of the colour proportional to the number of instances of the sulfatase found at that location[22].

[22] An interactive version of this map can be found here: https://drive.google.com/open?id=19F3xpMMPdjYIYQHKUtMRH7KjvjIXpFJa&usp=sharing

Horizon 2020 grant n. 676559

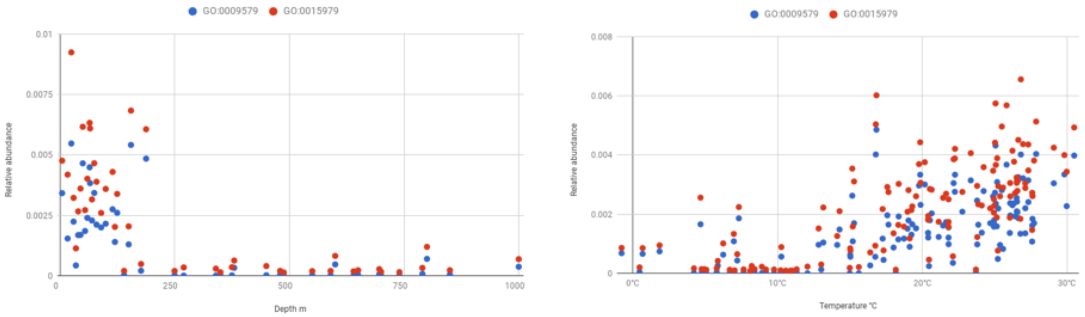

Figure 5 *- Correlation between depth (left) and temperature (right) and photosynthesis-related GO term counts, normalised by number of InterPro annotations for Tara Oceans project PRJEB1787.*

In the final phase of this work package, the various members of the Marine Metagenomics community are finalising these “use cases” examples for federating API searches across different resources. We aim to publish these, both on the ELIXIR marine metagenomics website and on the ELIXIR F1000 track. We anticipate these being key resources for the integration of the various services and resources developed throughout this project.

8.3 Summary & Future plans

It is notable that all of the pipelines that include taxonomic profiling (MGnify, META-pipe and BioMaS) of the SSU rRNA, have converged on the use of SILVA as the comprehensive reference database. Unbeknownst at the time of conducting the benchmarking work, SILVA was selected as an ELIXIR core data resource (WP3). Our increased dependency on SILVA highlights the critical role and need for continued support of such databases.

EMBL-EBI will continue to maintain our interactions with the Compute and Interoperability platforms to develop and enhance the range of microbiome-related analysis services that are provided. We will specifically focus on the scaling and distribution of our pipelines across different European compute infrastructures. In particular, this will involve the unification of the CWL and cloud deployment activities reported herein, which have been identified as objectives in the EOSC-Life project. This will require the refinement of the processes of deployment and the increased integrations of CWL workflows with containers. Unlike the MGnify amplicon analysis described above, where a single machine image (i.e. type) was used to perform the entire analysis, more careful alignment between a jobs resource request and actual utilisation needs to be made, leading to more cost-effective analysis and improved throughput.

We also expect to release another version of the MGnify pipeline (v5.0), which will integrate the ITS analysis and additional functional analyses, especially those that are suited to the analysis of larger contigs arising from assemblies. It will also be fully described using CWL, with the CWL workflow being executed in production. We will also continue to analyse new marine amplicon and whole metagenome shotgun datasets as they become available (e.g. Tara Ocean’s polar circle expedition). We will make our first wave of MAGs available for public download via MGnify/ENA, and ensure that these are accessible for inclusion into the MAR database.

Horizon 2020 grant n. 676559

While the work described here has focused on the Marine Metagenomics community, the technical developments in particular, namely the Cloud deployment and description of workflows in CWL, are broadly applicable to the wider metagenomics community and beyond. The use of CWL enables the pipelines to be packaged and readily compared. They also offer significantly greater flexibility, as the introduction of a new tool within a workflow becomes a matter of writing a small text (YAML) file that describes the tool, and the updating of the workflow CWL to call this file. This enables more agile development, with staff able to spend more time on developing things that matter (e.g. new analysis tools or data visualisation), as opposed to dealing with large, cumbersome pipelining code that always has a limited shelf life.

Horizon 2020 grant n. 676559

8.4 References

1. Falkowski,P.G., Barber,R.T. and Smetacek,V.,V. (1998) Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science*, **281**, 200–207.

2. Pace,N.R., Stahl,D.A., Lane,D.J. and Olsen,G.J. (1986) The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences. *Advances in Microbial Ecology*, **9**, 1–55.

3. Lanzén,A., Jørgensen,S.L., Huson,D.H., Gorfer,M., Grindhaug,S.H., Jonassen,I., Øvreås,L. and Urich,T. (2012) CREST--classification resources for environmental sequence tags. *PLoS One*, **7**, e49334.

4. Caporaso,J.G., Kuczynski,J., Stombaugh,J., Bittinger,K., Bushman,F.D., Costello,E.K., Fierer,N., Peña,A.G., Goodrich,J.K., Gordon,J.I., *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

5. DeSantis,T.Z., Hugenholtz,P., Larsen,N., Rojas,M., Brodie,E.L., Keller,K., Huber,T., Dalevi,D., Hu,P. and Andersen,G.L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069–5072.

6. Schloss,P.D., Westcott,S.L., Ryabin,T., Hall,J.R., Hartmann,M., Hollister,E.B., Lesniewski,R.A., Oakley,B.B., Parks,D.H., Robinson,C.J., *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.

7. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–43.

8. Cole,J.R., Wang,Q., Fish,J.A., Chai,B., McGarrell,D.M., Sun,Y., Brown,C.T., Porras-Alfaro,A., Kuske,C.R. and Tiedje,J.M. (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.*, **42**, D633–42.

9. Quast,C., Pruesse,E., Yilmaz,P., Gerken,J., Schweer,T., Yarza,P., Peplies,J. and Glöckner,F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–6.

10. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

11. Wang,Q., Garrity,G.M., Tiedje,J.M. and Cole,J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, **73**, 5261–5267.

12. Matias Rodrigues,J.F., Schmidt,T.S.B., Tackmann,J. and von Mering,C. (2017) MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis. *Bioinformatics*, **33**, 3808–3810.

13. Almeida,A., Mitchell,A.L., Tarkowska,A. and Finn,R.D. (2018) Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from

commonly sampled environments. *Gigascience*, **7**.

14. Lee,J.-H., Yi,H. and Chun,J. (2011) rRNASelector: a computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries. *J. Microbiol.*, **49**, 689–691.

15. Finn,R.D., Clements,J. and Eddy,S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–37.

16. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.

17. Wheeler,T.J. and Eddy,S.R. (2013) nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, **29**, 2487–2489.

18. Nawrocki,E.P., Burge,S.W., Bateman,A., Daub,J., Eberhardt,R.Y., Eddy,S.R., Floden,E.W., Gardner,P.P., Jones,T.A., Tate,J., *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–7.

19. Kalvari,I., Argasinska,J., Quinones-Olvera,N., Nawrocki,E.P., Rivas,E., Eddy,S.R., Bateman,A., Finn,R.D. and Petrov,A.I. (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.*, **46**, D335–D342.

20. Santamaria,M., Fosso,B., Licciulli,F., Balech,B., Larini,I., Grillo,G., De Caro,G., Liuni,S. and Pesole,G. (2018) ITSoneDB: a comprehensive collection of eukaryotic ribosomal RNA Internal Transcribed Spacer 1 (ITS1) sequences. *Nucleic Acids Res.*, **46**, D127–D132.

21. Santamaria,M., Fosso,B., Consiglio,A., De Caro,G., Grillo,G., Licciulli,F., Liuni,S., Marzano,M., Alonso-Alemany,D., Valiente,G., *et al.* (2012) Reference databases for taxonomic assignment in metagenomics. *Brief. Bioinform.*, **13**, 682–695.

22. Wilke,A., Bischof,J., Gerlach,W., Glass,E., Harrison,T., Keegan,K.P., Paczian,T., Trimble,W.L., Bagchi,S., Grama,A., *et al.* (2016) The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Res.*, **44**, D590–4.

23. Ten Hoopen,P., Finn,R.D., Bongo,L.A., Corre,E., Fosso,B., Meyer,F., Mitchell,A., Pelletier,E., Pesole,G., Santamaria,M., *et al.* (2017) The metagenomic data life-cycle: standards and best practices. *Gigascience*, **6**, 1–11.

24. Fosso,B., Santamaria,M., Marzano,M., Alonso-Alemany,D., Valiente,G., Donvito,G., Monaco,A., Notarangelo,P. and Pesole,G. (2015) BioMaS: a modular pipeline for Bioinformatic analysis of Metagenomic AmpliconS. *BMC Bioinformatics*, **16**, 203.

25. Fosso,B., Santamaria,M., D'Antonio,M., Lovero,D., Corrado,G., Vizza,E., Passaro,N., Garbuglia,A.R., Capobianchi,M.R., Crescenzi,M., *et al.* (2017) MetaShot: an accurate workflow for taxon classification of host-associated microbiome from shotgun metagenomic data. *Bioinformatics*, **33**, 1730–1732.

26. Davis,M.P.A., van Dongen,S., Abreu-Goodger,C., Bartonicek,N. and Enright,A.J. (2013) Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods*, **63**, 41–49.

27. Truong,D.T., Franzosa,E.A., Tickle,T.L., Scholz,M., Weingart,G., Pasolli,E., Tett,A., Huttenhower,C. and Segata,N. (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods*, **12**, 902–903.

28. Sczyrba,A., Hofmann,P., Belmann,P., Koslicki,D., Janssen,S., Dröge,J., Gregor,I., Majda,S., Fiedler,J., Dahms,E., *et al.* (2017) Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. *Nat. Methods*, **14**, 1063–1071.

29. Almeida,A., Mitchell,A.L., Boland,M., Forster,S.C., Gloor,G.B., Tarkowska,A., Lawley,T.D. and Finn,R.D. (2019) A new genomic blueprint of the human gut microbiota. *Nature*, **568**, 499–504.

30. Kang,D.D., Froula,J., Egan,R. and Wang,Z. (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, **3**, e1165.

31. Parks,D.H., Imelfort,M., Skennerton,C.T., Hugenholtz,P. and Tyson,G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, **25**, 1043–1055.

32. Sunagawa,S., Coelho,L.P., Chaffron,S., Kultima,J.R., Labadie,K., Salazar,G., Djahanschiri,B., Zeller,G., Mende,D.R., Alberti,A., *et al.* (2015) Ocean plankton. Structure and function of the global ocean microbiome. *Science*, **348**, 1261359.

33. Steinegger,M. and Söding,J. (2018) Clustering huge protein sequence sets in linear time. *Nat. Commun.*, **9**, 2542.

34. Raknes,I.A. and Bongo,L.A. (2018) META-pipe Authorization service. *F1000Res.*, **7**.

35. Agafonov,A., Mattila,K., Tuan,C.D., Tiede,L., Raknes,I.A. and Bongo,L.A. (2019) META-pipe cloud setup and execution. *F1000Res.*, **6**, 2060.

36. Park,Y.M., Squizzato,S., Buso,N., Gur,T. and Lopez,R. (2017) The EBI search engine: EBI search as a service-making biological data accessible for all. *Nucleic Acids Res.*, **45**, W545–W549.

37. Mitchell,A.L., Scheremetjew,M., Denise,H., Potter,S., Tarkowska,A., Qureshi,M., Salazar,G.A., Pesseat,S., Boland,M.A., Hunter,F.M.I., *et al.* (2018) EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res.*, **46**, D726–D735.