EXCELERATE Deliverable 6.2

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 on comprehensive metagenomic data standards environment	
WP No.	6	
Lead Beneficiary:	1: EMBL-EBI	
WP Title	Marine metagenomic infrastructure as a driver for research and industrial innovation	
Contractual delivery date:	28 February 2018	
Actual delivery date:	28 February 2018	
WP leader:	Rob Finn, EMBL-EBI Nils Peder Willassen, UiT	1: EMBL-EBI; 24: UiT
Partner(s) contributing to this deliverable:	EMBL-EBI, UiT, CNR, CNRS, GENOSCOPE	

Authors and Contributors:

Guy Cochrane (EMBL-EBI; lead author), Rob Finn (EMBL-EBI), Nils Peder Willassen (UiT), Lars Ailo Bongo (UiT), Monica Santamaria (CNR), Bruno Fosso (CNR), Eric Pelletier (Genoscope), Erwan Corre (Roscoff), Alex Mitchell (EMBL-EBI) and Simon Potter (EMBL-EBI)

Table of contents

1. Executive Summary

- We have previously reviewed data standards and best practices in marine metagenomics and have published our review and three core recommendations (ten Hoopen *et al.*, 2017)
- On our first recommendation - that scientists comply with data standards and recommended best practices, we assess in this Annex the status across sampling, sequencing, analysis and results archiving stages of metagenomics studies
- We report for our second recommendation - that computational analysis process be described formally, our progress through collaborative work with the Interoperability platform on adopting the Common Workflow Language (CWL)
- We describe that in response to our third recommendation - that the results of computational metabarcoding and metagenomics analysis processes be archived

Horizon 2020 grant n. 676559

- an extension to the European Nucleotide Archive (ENA) that supports submission, archiving and data access services for these data types

2. Impact

We assess the impact of standards, best practices around data and computational processes using metrics such as adoption by core and other ELIXIR resources, the number of data/workflow records in these resources complying with standards and the proportion of records complying with standards. Specific metrics are explained where provided in Appendix I: (EXCELERATE Deliverable 6.2: Annex I).

3. 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

4. Delivery and schedule

The delivery is delayed: Yes ☑ No

5. Adjustments made

None

Horizon 2020 grant n. 676559

6. Background information

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

Work package number	WP6	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)

Work Package Leads: Nils P Willassen (NO) and Rob Finn (EMBL-EBI)

Description of work and role of partners

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

Horizon 2020 grant n. 676559

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 particularly, 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: PT, NO

Horizon 2020 grant n. 676559

7. Appendix 1: Report on comprehensive metagenomic data standards environment

7.1. Summary

- This Annex relates to EXCELERATE Deliverable 6.2
- We have previously reviewed data standards and best practices in marine metagenomics and have published our review and three core recommendations (ten Hoopen *et al.*, 2017)
- On our first recommendation - that scientists comply with data standards and recommended best practices, we assess in this Annex the status across sampling, sequencing, analysis and results archiving stages of metagenomics studies
- We report here for our second recommendation - that computational analysis processes be described formally - our progress through collaborative work with the ELIXIR Interoperability Platform on adopting the Common Workflow Language (CWL)
- We describe here, in response to our third recommendation - that the results of computational metabarcoding and metagenomics analysis processes be archived - an extension to the European Nucleotide Archive (ENA) that supports submission, archiving and data access services for these data types

7.2. Introduction

As has been reported previously, in the early part of EXCELERATE, partners responsible for Task 6.1 reviewed existing data standards, conventions, and practices around the handling of the data and computational aspects of metagenomics and metabarcoding studies. This process resulted in the publication, in June 2017, of a joint paper in Gigascience (ten Hoopen *et al.*, 2017).

In the publication, we considered metagenomics studies as linear workflows around which data and metadata elements tracked first physical processes and then computational steps: First, environmental context for samples and the sampling method are described ("sampling"); second, sequencing libraries are prepared and sequencing machines configured ("sequencing"); third, resultant data are analysed and processed ("analysis"); and finally, data are archived and published ("archiving"). At each of these stages, we presented existing data standards and conventions and, where these were unavailable, we recommended best practice. In all, in our publication we made three calls to the community: first, we recommended compliance with data standards and best practice; second, we recommended a structured approach to description of computational analytical workflows; and third, we recommended the archiving of the data that are the output of computational analyses.

In this deliverable report we follow the same logical path from samples, through sequencing and analysis, to archiving. At each stage, we report on the status of compliance and relevant developments.

Horizon 2020 grant n. 676559

7.3. Sampling

MIxS continues as the overarching family of standards of use around sampling, from environmental context to sampling methods and early processing. MIxS continues to be maintained both as a standard by members of the Genomics Standards Consortium (in which many Use Case partners participate) and in the relevant data archives that support compliance, namely BioSamples and the European Nucleotide Archive (operated by Use Case and other EXCELERATE partners). In this section, we report on the current status of MIxS compliance in in archival data and outline a number of areas of MIxS-related development.

At the time of writing, in the relevant data repositories for the domain (the European Nucleotide Archive (ENA) and BioSamples Database (BioSamples)[1], 60% of sample records for marine environmental sequencing (metabarcoding or shotgun metagenomics) have been associated with appropriate MIxS checklists[2]. In comparison to rates of selection of appropriate checklists for other domains, this is high; we see, for example a figure of 50% for soil environmental samples and 25% across all of ENA. Given that limited time has passed since the publication of the paper, we do not expect yet to see growth in this figure (and indeed do not see it), but will continue to monitor. Importantly in addition, the impact of community awareness of data standards at point of submission cannot be measured entirely by us of checklists in full; rather, individual attributes that feature in checklists tend to be used significantly outside checklists. For example, while only 60% of EMBL-EBI-submitted samples use full checklists, 83% of these records show full Environment Ontology annotations for biome, feature and material.

Task partners have extended MIxS, albeit through efforts supported from outside EXCELERATE. First, a new checklist has been deployed under the UniEuk project (http://unieuk.org/) that relates to marine (and other) microbial eukaryotes; this checklist, detailed at https://www.ebi.ac.uk/ena/data/view/ERC000040, supports sample reporting for 18S amplicon diversity studies within UniEuk, but also has a use in related studies across marine environmental genomes. Second, we have under the EMBRIC project (http://www.embric.eu/), mapped concepts and attributes from standards of relevance to cultured bacteria, archaea, cyanobacteria, fungi, protozoa, microalgae, yeast, virus and phage in microbiology domain biological resource centres (http://www.embric.eu/sites/default/files/deliverables/D4.1_Data%20standards%20for%20EMBRIC.pdf).

7.4. Sequencing

In the publication, we reported that best practice was to report data to an INSDC database, such as ENA; we continue to see an increase in marine data in INSDC databases, with over 43,000 sequenced metabarcoding and metagenomics libraries at the end of 2017

[1] Submissions of sample data to ENA trigger automatic propagation to EMBL-EBI's BioSamples database; statistics provided here are valid from both the ENA and BioSamples perspectives.
[2] Checklists in the ENA and BioSamples system are sets of structured sample attributes grouped appropriately for given sample and study types.

Horizon 2020 grant n. 676559

(https://www.ebi.ac.uk/ena/data/warehouse/search?query=%22tax_tree(408172)%22&domain=read).

Library descriptors are required for submission that include library source, strategy and layout. SInce these are validated upon submission to INSDC databases, all 43,000 sequenced marine libraries meet this standard. Many library record include further attributes, but indexing system are not yet able to provide statistic on these.

Ongoing work includes the integration of the "EnSeqlopoedia" dictionary of sequencing applications (http://enseqlopedia.com/enseqlopedia/) into the EMBL-EBI's Ontology Lookup Service (https://www.ebi.ac.uk/ols/index) to assist in user selection of structured terms when describing sequencing libraries. Finally, we are planning of the implementation of a checklist system for library ("experiment") records in ENA, equivalent to that used for sample records, that would provide greater support for users during submission and data reuse and offer greater opportunities for tracking of usage of extended library descriptors.

Horizon 2020 grant n. 676559

7.5. Analysis

As reported in the publication, a major deficit in reporting the results from a metagenomics analysis is a complete and accurate description of the computational analysis pipeline or workflow used to interpret the data.

7.5.1. Describing workflows

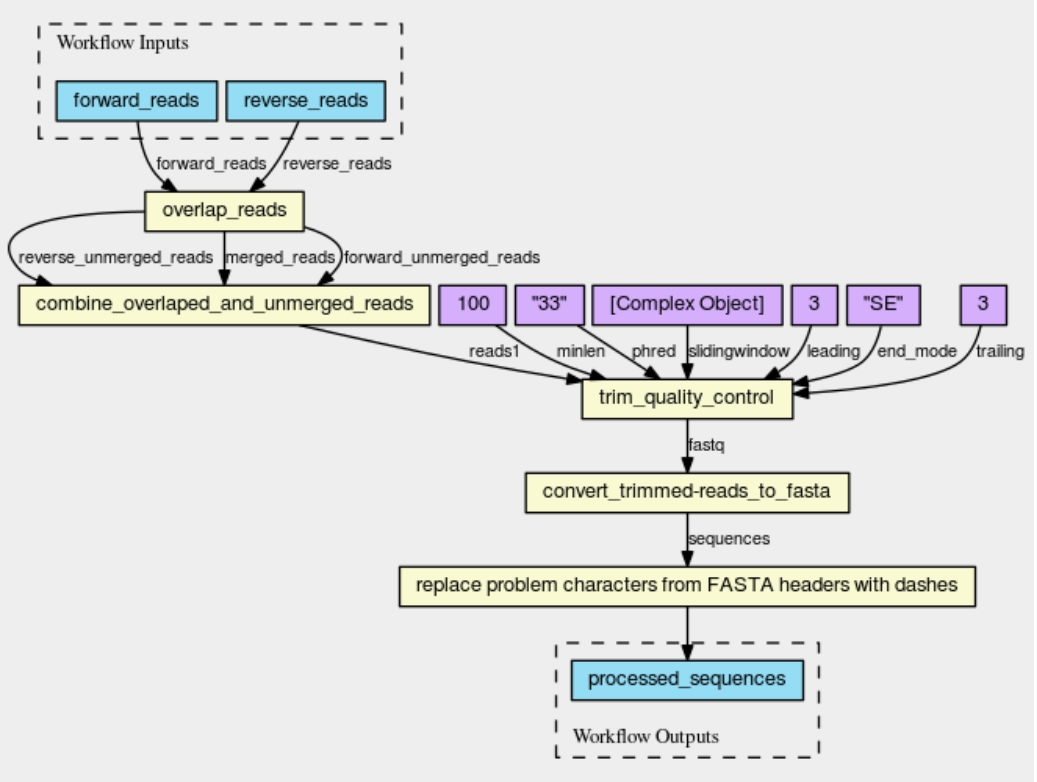
Having established the need for better analysis reproducibility, we discussed our needs at length with the ELIXIR Interoperability Platform. Our work clearly sets out the need for a system that encapsulates tools (versions and parameters), reference databases (versions) and computational resource requirements, which could be flexible enough to change rapidly as new tools are developed, and could scale elastically based on input sizes. Based on the Interoperability Platform recommendations, we evaluated a new, cross-execution workflow description language, the Common Workflow Language (CWL). The CWL specification has matured significantly during the course of this project, proceeding from draft-3 to v1.0, with a v1.0.2 now in preparation. Concomitant with the development of the specification, there has been an increasing number of workflow execution frameworks and tools for producing CWL tool descriptions.

The need to encapsulate data provenance, sample metadata and analysis workflows is not specific to Metagenomics. Indeed, the Research Object initiative has been advocating the need to provide a mechanism to associate together related resources about a scientific investigation so that they can be shared using a single identifier for many years. The use of a workflow language such as CWL (encapsulating the tools, reference datasets and parameters used), together with the storage of both sequence data and analysis results in recognised archival databases brings our particular uses case very close to achieving the gold standard of data science - a research object that allows the complete recreation of an entire digital experiment.

7.5.2. Adoption of CWL: EBI Metagenomics

The EMG became the first adopters of CWL within the Marine Metagenomics Use Case. Using CWL and the reference implementation execution engine cwltool, the analysis results for version 3.0 of the EMG analysis pipeline were demonstrated to be perfectly replicated. This approach replaced 1,000s of lines of Python code responsible for executing the existing in-house pipeline with a considerably simpler few hundred lines of CWL. An example of part of the CWL description of pipeline version 3.0 viewed with the workflow viewer (http://view.commonwl.org) is shown below.

Horizon 2020 grant n. 676559



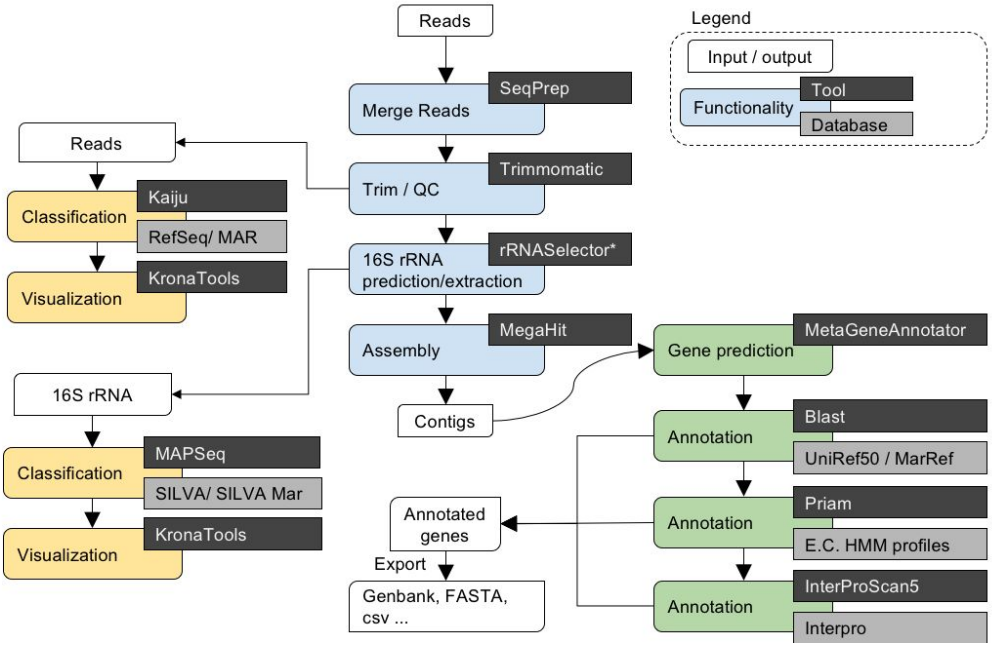
Having overcome the learning curve of producing one CWL description, it has been relatively simple to produce a series of other workflows, such as one encapsulating pipeline version 4.0. All EMG workflow descriptions are currently available on GitHub (https://github.com/EBI-Metagenomics/ebi-metagenomics-cwl). A drawback that we have encountered is that the CWL specification does not perform branching of the workflow, so different CWL descriptions have had to been produced for different input data types, e.g. single and paired-end Illumina sequencing. These require slightly different preparatory steps prior to the taxonomic and functional analysis, although both workflows inherit a common core workflow. Thus, for a single conceptual pipeline version in EMG, there need to be several corresponding CWL descriptions.

The progress described above enabled Michael Crusoe, the CWL Project Lead, to showcase the EMG CWL description at the Genome Standards Consortium meeting in May 2017 in Brisbane, Australia. We have also interacted with pipeline providers outside of this project consortium. In particular, MG-RAST have moved from their own workflow language to CWL, based on our experience/recommendation. We have started trialling the reciprocal execution of workflow elements at each other's sites. This has also instigated discussions on containerisation of the tools used within the workflow and the best practices associated with this (*e.g.* one tool per container or many tools in one container).

Horizon 2020 grant n. 676559

7.5.3. Adoption of CWL: META-pipe

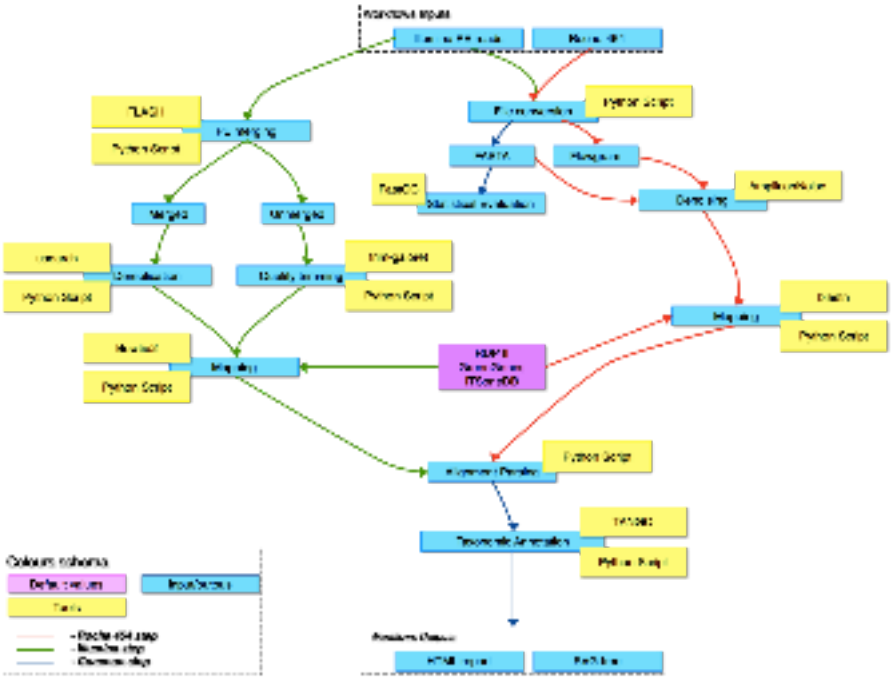
META-pipe is a pipeline for taxonomic classification and functional assignment of environmental samples. META-pipe has been designed and developed to serve the marine domain by implementing marine specific databases for taxonomic classification and functional assignment. META-pipe consists of three modules. Module one performs quality check (QC), trimming, prediction of 16S rRNA (SSU) and assembly of sequence reads. Module 2 perform taxonomic classification, while module 3 performs gene prediction and functional assignment of assembled contigs. The META-pipe team is in progress of implementing CWL based upon the figure below.



7.5.4. Adoption of CWL: BioMaS

BioMaS (Bioinformatic analysis of Metagenomic AmpliconS; Fosso *et al.*, 2015), has been developed as freely available web application aimed at the deep taxonomic profiling of both prokaryotes and eukaryotes microbiomes by means of metabarcoding. Its completely automatic workflow integrates consolidated third-party tools and custom Python and Bash scripts and its process is managed by the Job Submission Tool (JST; Donvito *et al.*, 2012), enabling the execution of multiple independent tasks in a cloud environment. Briefly, BioMaS includes the quality evaluation of input NGS data, their denoising and dereplication, their mapping against selected reference databases and their assignment to taxonomic classes (see workflow below).

Horizon 2020 grant n. 676559



In response to the need to formalise descriptions of computational workflows, and at the same time taking the opportunity to progress BioMaS to improved methods and more flexibility with data formats, the BioMaS workflow is being subjected to a deep re-writing procedure allowing to: (i) effectively process multiple samples in parallel; (ii) include an OTU picking procedure and (iii) produce a biom output. The workflow specification and wrappers are already, or will be, written using the Common Workflow Language (CWL).

7.5.5. Future plans

As indicated above, different pipelines are in a variety of states of transition to the adoption of CWL for describing their pipelines, with EMG currently the furthest along this path. Collectively, our experiences are that it is relatively quick to convert a well understood/documented pipeline to a CWL description. The next challenge is execution of this workflow and optimising this for the compute environment available. This has focused us more on resource requirements and ensuring that these are also captured as part of the CWL description. However, due to the broad range of input sizes and different complexities, a simple resource range is insufficient. In the latest version of the CWL specification, its developers have integrated functionality based on our use case where this needs to be estimated and added dynamically to the workflow description as a set of input parameters.

Over the coming year, we will start working in containerising the tools used within the workflows, to allow more agile deployment. Our efforts with CWL are being assisted by two ELIXIR Implementation Studies, one from each of the Interoperability and Compute Platforms. We will also try to develop a common repository of the CWL tool

Horizon 2020 grant n. 676559

descriptions and workflows, as well as consolidate some of our training materials surrounding the use of CWL.

7.6. Emergence of metagenome assemblies

For industrial applications full length peptides are required, so analysis of shotgun reads to provide the functional signature of the metagenome is inadequate. To meet these demands, and in-keeping with general advances in the field, the EBI metagenomics (EMG) pipeline was extended to enable metagenomics assembly. The process of performing metagenomics assembly generates many contigs that, unlike the typical intermediate outputs from the EMG pipeline, require archiving. Thus, in accordance with best practice, we have developed a process whereby EMG assemblies are submitted back to ENA. Subsequently, these are retrieved and analysed by EMG. These assembles - termed third-party analyses within ENA - are archived and accessioned and linked to the original project, providing provenance of the original data, as well as data about how these were provisioned.

As will be described in more detail in other deliverable reports, the recovery of genomes from metagenomes (so called MAGs) is becoming more routine within the scientific community. In the absence of an absolute ground truth, there is a need to assess their quality. Through our involvement with the wider metagenomics community, we contributed to a community standards paper outlining the basic requirements for assessing the quality of MAGs (Bowers *et al.*, 2017). Within EMG, we have developed a nascent pipeline that adheres to the best practices defined in this article, including coverage metrics of the contigs by reads, and the completeness and contamination. This is currently driving the requirements gathering in ENA, so that appropriate metrics will be captured and become the focus of future work.

7.7. Results archiving

Our third recommendation was that the results of analysis be archived. The ubiquitous data types that represent the outputs of environmental sequencing are identification data, representing sets of identification of biological entities in sequenced samples, with associated annotations relating to confidence and provenance. Identifications take the form of functional calls, such as InterPro hits, GO terms or gene symbols (for shotgun metagenomics) and taxonomic calls such as Operational Taxonomic Units and taxa within established taxonomies (for metabarcoding and shotgun metagenomics).

To support the archiving of these data types, we have extended ENA, using its "analysis" object schema, to support "identification tables". These data structures are now supported through the RESTful submissions interface, in the back-end archiving infrastructure and in web and programmatic search and retrieval services. Currently supported as tabular files, we remain open to considering further formats, such as BIOM, as user demand and resources allow. A next priority is the creation of openly available validation software for identification tables to allow deeper rigour in preparing, archiving and interoperating on these data.

There are as yet few data sets within this extended part of ENA and no published records. Our first major use of identification tables will be for the analysis outputs from EMGs' assembly-based analyses described above.

Horizon 2020 grant n. 676559

7.8. References

Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu WT, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Genome Standards Consortium, Lapidus A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF, Hugenholtz P, Woyke T. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 2017 Aug;35(8) 725-731. doi:10.1038/nbt.3893. PMID: 28787424.

Fosso B, Santamaria M, Marzano M, Alonso-Alemany D, Valiente G, Donvito G, et al. BioMaS: a modular pipeline for Bioinformatic analysis of Metagenomic AmpliconS. BMC Bioinformatics [Internet]. 2015;16:203.

Donvito G, Vicario S, Notarangelo P, Balech B. PoS (EGICF12-EMITC2) 029. Proc. EGI Community Forum 2012/EMI Second Tech. Conf. (EGICF12-EMITC2). 2012. p. 26–30.

ten Hoopen P, Finn RD, Bongo LA, Corre E, Fosso B, Meyer F, Mitchell A, Pelletier E, Pesole G, Santamaria M, Willassen NP, Cochrane G. The metagenomic data life-cycle: standards and best practices. Gigascience. 2017 Aug;6(8) 1-11. doi:10.1093/gigascience/gix047. PMID: 28637310; PMCID: PMC5737865.

Horizon 2020 grant n. 676559