EXCELERATE Deliverable D6.4

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 assessment criteria and standardisation of metagenome assembled genomes (MAGs) from Marine samples	
WP No.	6	
Lead Beneficiary:	EMBL-EBI	
WP Title	Marine metagenomic infrastructure as a driver for research and industrial innovation	
Contractual delivery date:	31 May 2019	
Actual delivery date:	29 May 2019	
WP leader:	Rob Finn Nils peder Willassen	1, EMBL-EBI 24, UiT
Partner(s) contributing to this deliverable:	EMBL-EBI, UiT	

Authors and Contributors:

Rob Finn (EMBL-EBI), Alex Mitchell (EMBL-EBI), Guy Cochrane (EMBL-EBI), Josephine Burgin (EMBL-EBI), Nils Peder Willassen (UiT)

Reviewers:

None

1. Table of contents

2. Executive Summary

- Demonstration of the application of completeness/contamination estimates on metagenome assembled genomes (MAGs) produced as part of WP6. These results of these estimates are discussed in relationship of the MIMAGs and MISAGs standards.
- Survey of the genomes, MAGs and single amplified genomes (SAGs) found in the MAR database. This demonstrates that many of the draft genomes are heavily contaminated and/or incomplete.
- Description of the new infrastructure in the European Nucleotide Archive, and how this has been structured to accommodate the data generated within WP6 (and more broadly).
- We also highlight the current limitations and issues of the tools employed to estimate completeness and contamination.

3. Impact

This deliverable is primarily a report to understand the quality of the genomes found in the MAR databases, and those being recovered from metagenomics datasets by MGnify. This will allow users of these resources to better understand the quality of the genomes that they can download.

The use of CheckM has been reported in at least 4 different workshops, with total audiences exceeding 100. The use of CheckM on the MGnify genomes was presented at the GRC on Marine microbiomes, which had 200 participants.

4. Project objectives

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

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

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

5. Delivery and schedule

The delivery is delayed: Yes • No☑

6. Adjustments made

No adjustments made

7. Background information

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

Work package number	6	Start date or starting event:	month 1
Work package title	Use Case A: Marine metagenomic infrastructure as driver for research and industrial innovation		
Lead	Nils Peder Willassen (NO) and Rob Finn (EMBL-EBI)		
Participant number and person months per participant P1: EMBL-EBI (28PM) - P17: FCG (2PM) - P20: CCMAR (11PM) – P24 UiT (36PM) – P27: CNRS (10PM) - P31: CNR (10 PM)			
Objectives The main objective for this Use Case is to develop a sustainable metagenomics infrastructure to enhance research and industrial innovation within the marine domain before M36 of the ELIXIR-EXCELERATE project. The main objective will be achieved by the following specific objectives:			

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

Description of work

Metagenomics has the potential to provide unprecedented insight into the structure and function of heterogeneous communities of microorganisms and their vast biodiversity. Microbial communities affect human and animal health and are critical components of all terrestrial and aquatic ecosystems. They can be exploited e.g. to identify novel biocatalysts for production of fuels or chemicals (bioprospecting), make functional feed for aquaculture species, and for environmental monitoring. However, in order to expand the potential further for the research community and biotech industry, especially within the marine domain, the metagenomics methodologies need to overcome a number of challenges related to standardization, development of relevant databases and bioinformatics tools. New and emerging sequencing technologies, integration of metadata gives an extra burden to the development of future databases and tools. The Use Case “Marine metagenomic infrastructure as driver for research and industrial innovation” will contribute to the overall objectives of the ELIXIR-EXCELERATE project by developing research infrastructure and service provision specific for the marine domain in order to enable metagenomic approaches responding to societal and industrial needs. The outcome of the proposed Use Case will meet the major needs expresses 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: EMBL.EBI, NO

8. Appendix 1: Report on assessment criteria and standardisation of metagenome assembled genomes (MAGs) from Marine samples

8.1. Background

Whole-Genome Shotgun (WGS) sequencing has been the technology of choice for genome sequencing of cultivable organisms. Genome assemblers are unified by the assumption of sequence overlap among read sequences in the dataset, thereby enabling the progressive extension of sequences into contigs and reconstructing the original genomic DNA sequence. However, the presence of repetitive regions, and errors introduced by the sequencing process can make the approach unfeasible or computationally challenging, leading to genome misassemblies that warrant additional experimental and informatics analyses to identify and correct. As such, only ~13% of the prokaryotic genome sequencing projects in public databases are considered completely finished and the remaining 87% are deposited as draft genome sequences (have an average of 190 contigs) (1).

Recent technological developments have facilitated unprecedented access to the uncultured genomes or “the microbial dark matter”, using either single-cell or metagenomic sequencing technologies. Although both Single Amplification Genome (SAG) and the Metagenome Assembled Genome (MAG) approaches have proven powerful, there are a number of challenges associated with each of these approaches. Starting from one genome sequence, SAG sequencing , is demanding due to PCR artifacts, such as uneven coverage depth, missing regions, chimeric molecules, providing incomplete genomes of short length. It is further complicated by contamination of free DNA originating from reagents, kits or even within the samples. Generation of MAGs from environmental samples requires high sequencing depth and ideally, a large number of samples with the same richness but different relative species abundance in order to identify and assemble identical bins. In addition, the quality of MAGs is highly dependent on the quality of the metagenome assembly and each bin (or MAG) often represents a population of closely related organisms (i.e. species or strains) rather than a single organism.

While the quality of isolate WGS genomes has traditionally been evaluated using assembly statistics, such as total assembly size, number of contigs, contig N50/L50, and maximum contig length, where N50 is defined as the sequence length of the shortest contig at 50% of the total genome length and L50 as the number of contigs whose summed length is N50. However, these statistics are less meaningful in the case of MAGs and SAGs. Thus, in the absence of a close reference genome how can the quality of a genome, MAG or SAG be assessed?

8.1.1. Single Copy Genes (SGCs)

Assessing quality of SAGs and MAGs is usually performed by identifying and counting universal Single Copy Genes (SCGs). These SCGs are genes which are found ubiquitously across bacterial and archaeal lineages, and are only found once within a genome. Several lists of such SCGs exist and consist mainly of genes encoding for ribosomal proteins and other housekeeping genes (2, 3). By using such lists, one can estimate the completeness and contamination of SAGs or MAGs. In short, completeness is the number of unique SCGs present in the genome divided by the number of unique SCGs in the list. Contamination is estimated by counting the number of SCGs present in multiple copies, as only one copy of each SCG should be present per genome.

CheckM (3), the most used software for assessing assembly completeness and contamination, use ubiquitous and single-copy marker genes that are specific to a genomic lineage within a reference tree (2). The lineage-specific marker sets were determined for all nodes within the reference genome tree by identifying single-copy genes present in ≥97% of all descendant genomes. The quality of a genome, in terms of completeness and contamination, can be estimated using the presence/absence of these genes defined at any parental node between the genome's position in the reference tree and the root.

8.1.2. MIMAGs and MISAGs

In 2017, the Genomic Standards Consortium (GSC) published two standards for reporting on the quality of MAGs and SAGs (2, 4): Minimum Information about a Single Amplified Genome (MISAG) and; the Minimum Information about a Metagenome-Assembled Genome (MIMAG). These standards are primarily aimed at improving the reporting of assembly quality, and estimates of genome completeness and contamination, and provide criteria for describing the quality of the genomes, summarised in Table 1.

Table 1. Classification of assembly quality. Definitions used in the table are as follows: [1]Q50 = Phred quality score of 50: probability of one incorrect base call in 100,000 (99.999% base call accuracy). [2]Assembly statistics, including to total assembly length, number of chromosomes and plasmids, number of scaffolds and contigs, contig and scaffold N50, and maximum contig length. [3] Completeness score - the ratio of observed single-copy marker genes to total single-copy marker genes in chosen marker gene set (%). [4] Contamination score - the ratio of observed single-copy marker genes in ≥2 copies to total single-copy marker genes in chosen marker gene set (%).

Quality	***Description***
Finished	*Single, validated, contiguous sequence per replicon without gaps or ambiguities with a consensus error rate equivalent to Q50[1] or better. Assembly statistics[2] report.*
High Quality Draft	*Multiple fragments where gaps span repetitive regions. Assembly statistics report. Presence of the 23S, 16S and 5S rRNA genes and at least 18 tRNAs.* *Completeness score[3] > 90%* *Contamination score[4] < 5%*

Medium Quality Draft	*Many fragments with little to no review of assembly other than reporting of standard assembly statistics.* *Completeness score* ≥ 50% *Contamination score < 10%*
Low Quality Draft	*Many fragments with little to no review of assembly other than reporting of standard assembly statistics.* *Completeness score < 50%* *Contamination score < 10%*

As indicated in Table 1, the MIMAGs and MISAGs use fixed rates of completeness and contamination. We compare these standards to another commonly used metric, the quality score (QS), first introduced by Parks et al (2, 4, 5) and defined as:

$$\text{completeness} - 5 \times \text{contamination}$$

With only genomes with a quality of ≥50 considered as being of acceptable quality, termed QS50. The multiplication factor of the contamination means that there is a trade-off, ensuring that partial genomes can only contain minimal contamination.

In this deliverable, we describe the application of the various completeness and contamination estimates as applied to the MAGs generated as part of WP6, the genomes contained in the MAR reference databases, and the emerging infrastructure for the capturing of MAGs within the European Nucleotide Archive (ENA).

8.2. Overview and Status

8.2.1. Quality assessment of MAGs in MGnify

As described in deliverable D6.3, the MGnify team (EMBL-EBI) has performed large scale assembly and binning on shotgun metagenomic datasets from aquatic samples. Below we demonstrate the result of running CheckM on the 17,830 bins from 1,419 assemblies of aquatic samples (Figure 1). From these 17,830 bins, 3,069 high quality MAGs were obtained (i.e. >90% completeness, <5% contamination), while 4,663 were deemed to be of medium quality (i.e. >50% completeness, <10% contamination), and the remaining 10,098 bins classified as low quality. These low quality bins show a huge variation in completeness and contamination, but with the majority still tending to have low levels of contamination (<5%).

Notably, only a tiny fraction (<1%) of the MAGs deemed as high-quality MAGs by CheckM analysis actually reached the MIMAGs high quality standard, due a a lack of ribosomal RNA sequences: while tRNAs were frequently identified (mean number of different tRNAs was 15), the ribosomal RNAs are rarely found within the bins. This is a well known feature of metagenomic assemblies using tools such as MetaSPAdes (5–7), as the conserved regions of the ribosomal rRNA form a convergence point on the de Bruijn graph, a feature that is removed during the assembly refinement stage of MetaSPAdes algorithm. Nevertheless, in related work (8) performing *de novo* assemblies on the human gut, where isolate genomes could be used as a reference, the estimated completeness and genome alignments (isolate vs MAG) were highly correlated.

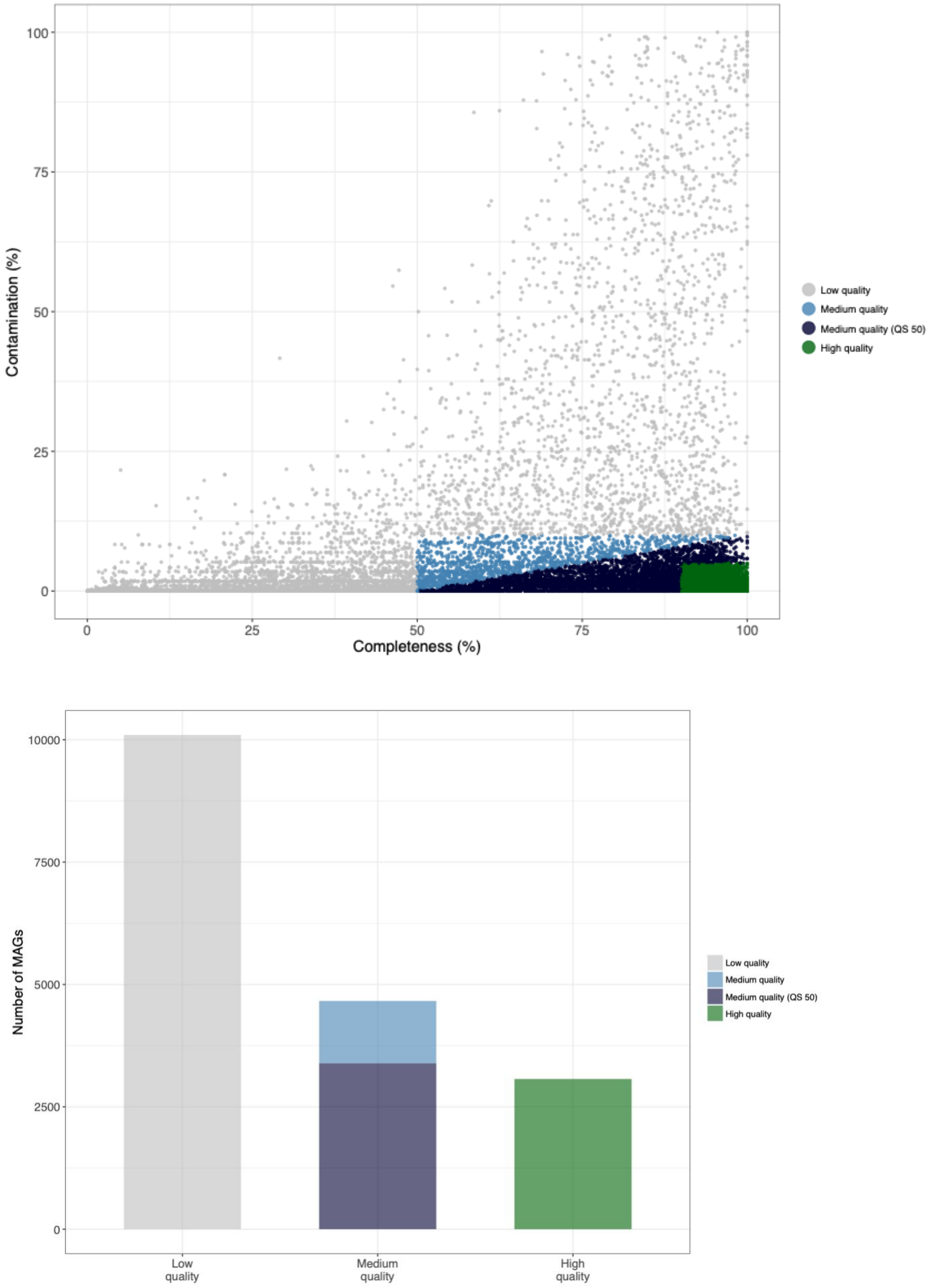


Figure 1 - Top is a scatter plot of the completeness vs contamination of the 17,830 bins. Each point is colour-coded according to the classification of the bin as being low, medium or high quality.

The medium quality bins are subdivided into those exceeding QS score of 50 (dark blue) and lower quality bins (light blue). The bottom chart shows the number of bins belonging to each category.

In figure 1, the medium quality MAGs are divided into two, with the darker blue corresponding to those MAGs that pass the QS50, criteria. In our experience of comparing to isolate genomes and inspecting the bin quality, using tools such as Anvi'o (8, 9), this QS50 criteria provides a better criteria, than the >50% completeness and <10% contamination. Furthermore, as illustrated in figure 1 (bottom), this stricter criteria does not lead to a significant reduction of MAGs, with 1,273 medium quality MAG failing the QS50, while 3,385 pass.

8.2.2. Assessing the assembly quality of published microbial genomes recovered from cultivated, metagenomes and single cells included in MarRef and MarDB

To date, no comprehensive analysis of the quality of the microbial genomes deposited to the INSDC databases[1] (ENA, DDBJ, and NCBI) has been performed. To provide an insight into the quality of deposited genomes we chose to use the manually curated MAR databases (10), MarRef and MarDB, which contain a marine subsection of the INSDC microbial genomes. While MarRef contains only complete and gapless marine genomes, MarDB contains genomes regardless of the level of completeness.

Methods

All genome sequence datasets were retrieved from either ENA (European Nucleotide Archive[2], or NCBI[3]. The archaeal and bacterial genomes included in the MAR databases were annotated either using the NCBI's Prokaryotic Genome Automatic Annotation Pipeline (PGAP) (11) . However, approx. 35% of the downloaded genomes lack PGAP annotation and was annotated using the Prokka software (12).

[1] http://www.insdc.org
[2] http://www.ebi.ac.uk/ena
[3] https://www.ncbi.nlm.nih.gov/

The completeness and contamination of each genome assembly in MarRef and MarDB was estimated using CheckM v.1.0.13 using the lineage_wf workflow and the QS for each genome assembly in the MAR databases was calculated. These metrics, together with the genome annotations where then compared to the MIMAGs and MISAGs quality definitions, as summarised in Table 2.

Using the completeness and contamination system for assessing the quality WGSs, MAGs and SAGs, the QS values for high quality draft should be ≥ 65 , medium quality draft ≥ 0 to < 65, while low quality draft < 0 as shown in Table 2. Finished genomes should in principle have QS >> 65, were all replicons only consists of one verified continuous sequence.

Table 2. Estimation of quality score based upton MIMAG and MISAG standards. [1]Quality score (QS) = completeness – 5x contamination. [2]A finished assembly is defined as a single verified contiguous sequence per replicon without gaps and have approximately 100% completeness and 0% contamination.

Classification	QS[1]	Completeness	Contamination
High	≥ 65	≥ 90%	< 5%
Medium	≥ 0 to < 65	≥ 50% to 90%	< 10%
Low	< 0	< 50%	< 10%

MarRef and MarDB genome statistics

The version of MarRef and MarDB used in this study contains 735 and 10767 entries, respectively. While more than 98 % of the genomes in MarRef are generated using WGS sequencing, the most commonly used technology to access complete genomes from cultivated isolates, the rest of the entries are MAGs. No SAGs have been identified as finished and included in MarRef. MarDB, on the other hand, is a mix of WGS, MAGs and SAGs. WGSs and MAGs account for 93.1% of the entries, with 45,3% and 47,8% of the entries, respectively. So far only 6.9% of the entries in MarDB are SAGs. The statistics are summarized in Table 3.

Table 3. Overview of the content in the MAR databases divided into different technologies for recovering genomes.

			Assembly length (bp)		
Database	**Technology**	**Number of entries**	**Average**	**Min length**	**Max length**
MarRef	WGS	727	3 829 776	490 885	9 708 656

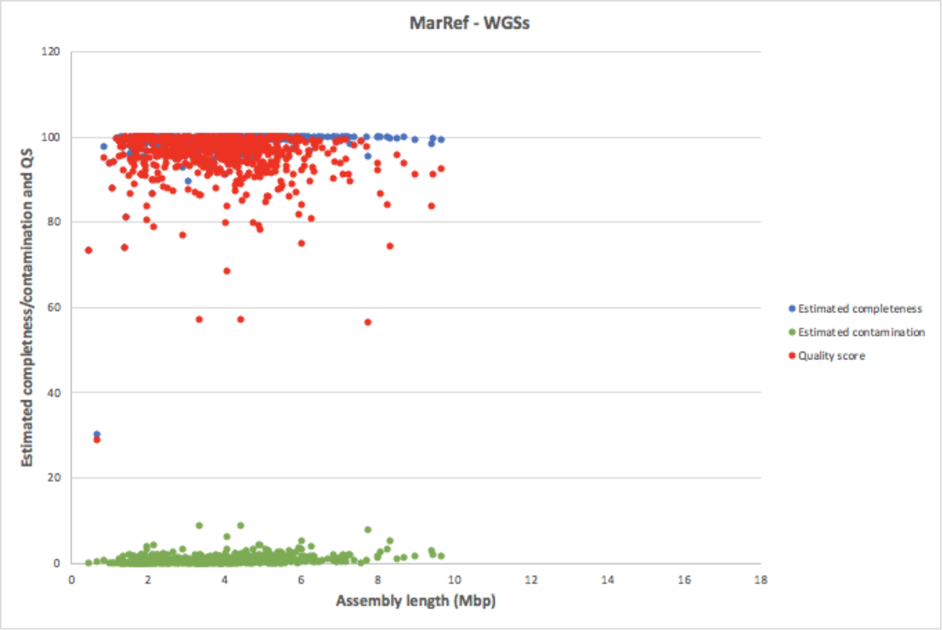
	MAG	11	2 233 426	593 366	9 384 763
MarDB	WGS	4879	4 167 928	238 717	16 377 176
	MAG	5143	2 337 402	103 922	10 752 934
	SAG	739	1 082 947	134 516	3 579 979

The average genome assemblies length varies significantly between the different technologies used to generate the assemblies. The average length for the WGSs in MarRef and MarDB is approx. 3,829 and 4,167 Mbp, respectively, ranging from 0,409 to 9,708 Mbp in MarRef and 0,238 to 16,337 Mbp in MarDB. The MAGs in MarRef and MarDB have an average length of 2,233 and 2,337 Mbp respectively, ranging from 0,593 to 9,385 in MarRef and 0,104 to 10,753 Mbp in MarDB. The average assembly length of SAGs in MarDB is 1,082 and the length varies from 0,135 to 3,580 Mbp.

The completeness, contamination and quality score the entries in MarRef and MarDB are presented in Fig. 2, Fig. 3 and summarized in Table 3.

MarRef Quality Scores

For MarRef, which contains only finished genomes, 99.3% of the entries have a quality score of 65 or higher, were 91.7% of the entries have a QS higher than 90. Only five entries have a QS less than 65, and accounts for 0.7 % of all entries. For the WGS, with an average completeness of 99,02% and contamination of 0.63%, only 0,6% of the entries have a QS < 65 and 99,4 % ≥ 65. For the MAGs in MarRef, which accounts for only 11 entries, have an average completeness of 87,06% and contamination of 0,26%. In MarRef 91% of the entries have a QS > 65, while only one entry has a QS < 65 as shown in Fig. 2.



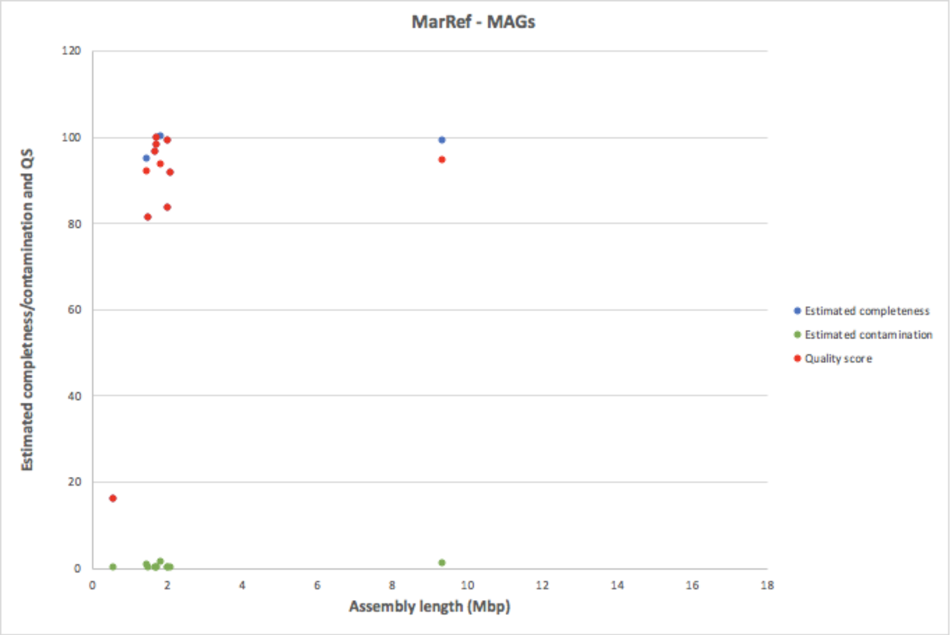


Figure 2. Distribution of WGS and MAG assembly lengths (Mbp) in MarRef as a function of completeness, contamination and quality score.

MarDB Quality Scores

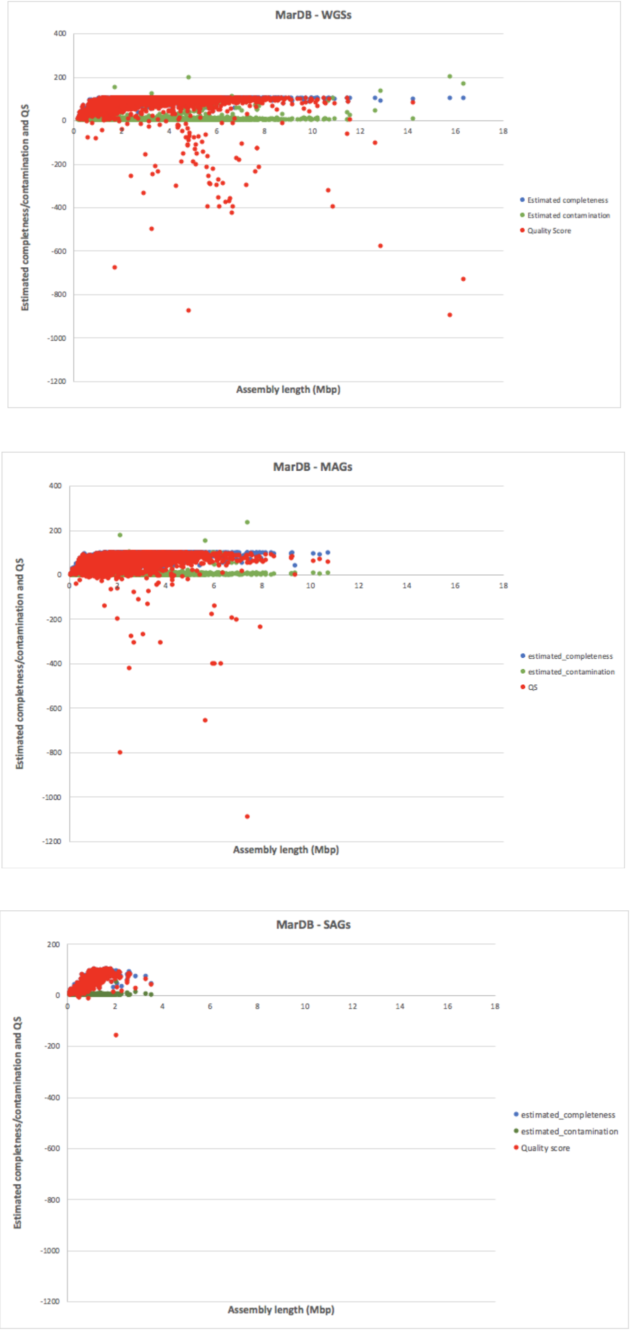


Figure 3. Distribution of WGS, MAG and SAGs assembly lengths (Mbp) in MarDB as a function of % completeness, contamination and quality score.

For MarDB, 7337 entries out of the total 10767 entries, 69,1%, have a QS ≥ 65, which can be regarded as high-quality assembly drafts, while 3213, which accounts for 29,8% of all entries, can be classified as medium-quality drafts. The rest, 1,6%, can be considered as low-quality drafts.

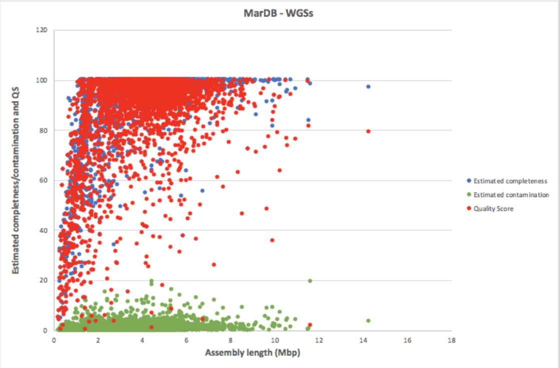
The estimated average completeness and contamination for WGS in MarDB are 94,45% and 2,08%, respectively. For the MAGs 75, 28% and 2,65% and SAGs, 58,78% and

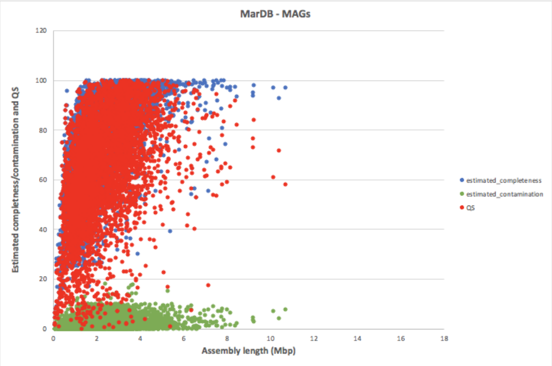
0,37%. For the WGSs, MAGs and SAGs in MarDB, 88, 77 and 6 entries, respectively, have QS < 0%, which represents approximately 1,8, 1,5 and 0,81% of the entries as shown in Figure 3 and Table 4, which indicate low completeness and/or high level of contamination of theses entries. For the MAGs, 20 entries have a QS < -100%, with a contamination ranging from 37 to 237%. For WGS 52 entries have a QS < -100%, were the contamination range from 40 to 200%. Only one SAG entry has a QS less than -100%.

Table 4. Quality Score (QS) for the genome assemblies in MarRef and MarDB. *QS = completeness – 5 x contamination

		Quality score (QS) *		
Database	**Technology**	**< 0**	**0 to 65**	**≥ 65**
MarRef	WGS	0	4	723
	MAG	0	1	10
MarDB	WGS	88	374	4417
	MAG	77	2450	2616
	SAG	6	389	344

For better visualization of the differences between the three technologies used to generate genome assemblies in MarDB, all entries with QS < 0 (low quality draft assemblies) was removed as shown in Fig. 4. For the WGS entries, 90,5% have a QS > 65, while for the MAGs and SAGs only 50,9% and 46,4%, respectively, have a QS higher than 65, which can be classified as high-quality draft assemblies. The number of entries which can be classified as medium quality draft assemblies (QS between 0 and 65), are 7,7%, 47,6% and 52,5% for the WGSs, MAGs and SAGs, respectively.





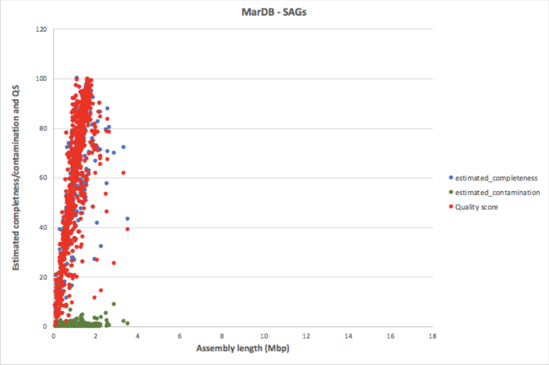


Figure 4. Distribution of WGS, MAG and SAG assembly lengths (Mbp) in MarDB as a function of % completeness, contamination and quality score.

8.3. Discussion

All entries in MarRef are closed and should be classified as finished draft assemblies or finished genomes. The average assembly lengths of WGSs and MAGs in MarRef are 3,829 Mbp and 2,233 Mbp, respectively, ranging from 0,409 to 9,708 Mbp. No SAGs have been identified in MarRef, indicating that this technology may not be useful for generating finished genomes. The number of contigs in the MarRef entries are below < 10, which show that some of the genome assemblies also include closed plasmids in addition to one or several chromosomes.

For the WGSs in MarRef, the average completeness is 99,02% and varies from 30% to 100%, with seven entries < 90%, while the average contamination is only 0,63 and varies from 0% to 8,62%, with only six entries with > 5%. Of the five entries with QS < 65, which indicates medium-quality draft assemblies, two entries have very low completeness, 15,72% and 30,13%, but also low contamination, 0% and 0,29%, respectively. For the MAGs in MarRef the average completeness and contamination are 87,06% and 0,26%, respectively.

MMP03766451, the entry with the lowest QS value (15,72), is the Bacterium AB1 strain AB1-8, a closed MAG with an assembly length of 593,366 bp and a GC content of 20,9%. This rare bacterium was recovered from *de novo* assembled metagenomics reads from the marine bryozoan *Bugula neritina*. The WGS with the lowest QS value (28,68), with the completeness of 30,13% and contamination of 0,29%, is *Salinicoccus sp*. BAB 3246 (MMP06324084). The bacterium was isolated at the coastal region of Gujarat, India consists of one closed chromosome of 713,204 bp. Both these bacteria are examples of closed genomes, with only one replicon/contig, which seems to lack some of the lineage specific marker genes, which give rise to the low completeness and QS.

The average assembly lengths of WGSs, MAGs, and SAGs in MarDB are 4,167, 2,233 and 1,082 Mbp, respectively, ranging from 0,134 to 16,377 Mbp. The number of contigs in MarDB entries varies from 1 to 8951, where approx. 12,6% have 200 or more contigs.

The estimated average completeness for the WGS in MarDB is 94,45% and much higher than for the MAGs and SAGs, 75,28% and 58,78%, respectively, which indicate that the WAGs give the most complete genome assemblies and SAGs the least complete. The average contamination for WGSs, MAGs and SAGs are 2,08%, 2,65% and 0,37%, respectively. The low contamination found in the SAGs are as expected due to the single-cell sequencing technology. The contamination of the WGSs and MAGs are approx. on the same level, with some more contamination in MAGs than WGSs.

There are 171 entries with QS < 0, which indicates a high level of contamination. The MAG entry MMP08159310, is annotated as a Euryarchaeota archaeon strain Lau_6, have a QS of -1089, indicating a very high level of contamination. The 7,41 Mbp genome consists of 298 contigs, with a completeness of 94,07% and 236,57%. According to the published paper (13) (Supplementary Table 3), the contigs contains 4 genomes which explain the high level of contamination. Another example is *Alcanivorax* HI0035, MMP04580768 a WGS, with estimated completeness of 100% and contamination of 105,97 % giving a QS value of -429. The genome length is 6,69 Mbp, with coverage of 25x, and consists of 3128 contigs. The high degree of contamination is probably due to contamination of similar species during the cultivation of the bacteria.

8.3.1. MAGs in the archives

At the outset of the EXCELERATE project, there was no systematic representation of MAGs within ENA (or INSDC), nor the support for the capture and presentation of MIMAGs metadata. To deal with this, new data structures and supporting submission and data presentation systems have been developed in ENA and are under ongoing implementation across INSDC. Work within ENA has been informed and integrated by the ELIXIR Marine Metagenomics Community with practical software development work covered under BBSRC funding.

Data structures

We have developed data structures that allow metagenome assembly data to be captured appropriately into ENA (figure 5). Assembly data related to MAGs have been classed into three tiers: (i) primary assemblies (per-sample assemblies of all contigs covering all species); (ii) binned metagenome assemblies (partitioned contigs relating to what is asserted to belong to a particular taxonomic group) and (iii) Metagenome-Assembled Genomes (MAGs; those assemblies asserted to belong to a particular taxonomic group that are informative as reference points alongside isolate assemblies, as judged by the scientists who generate them).

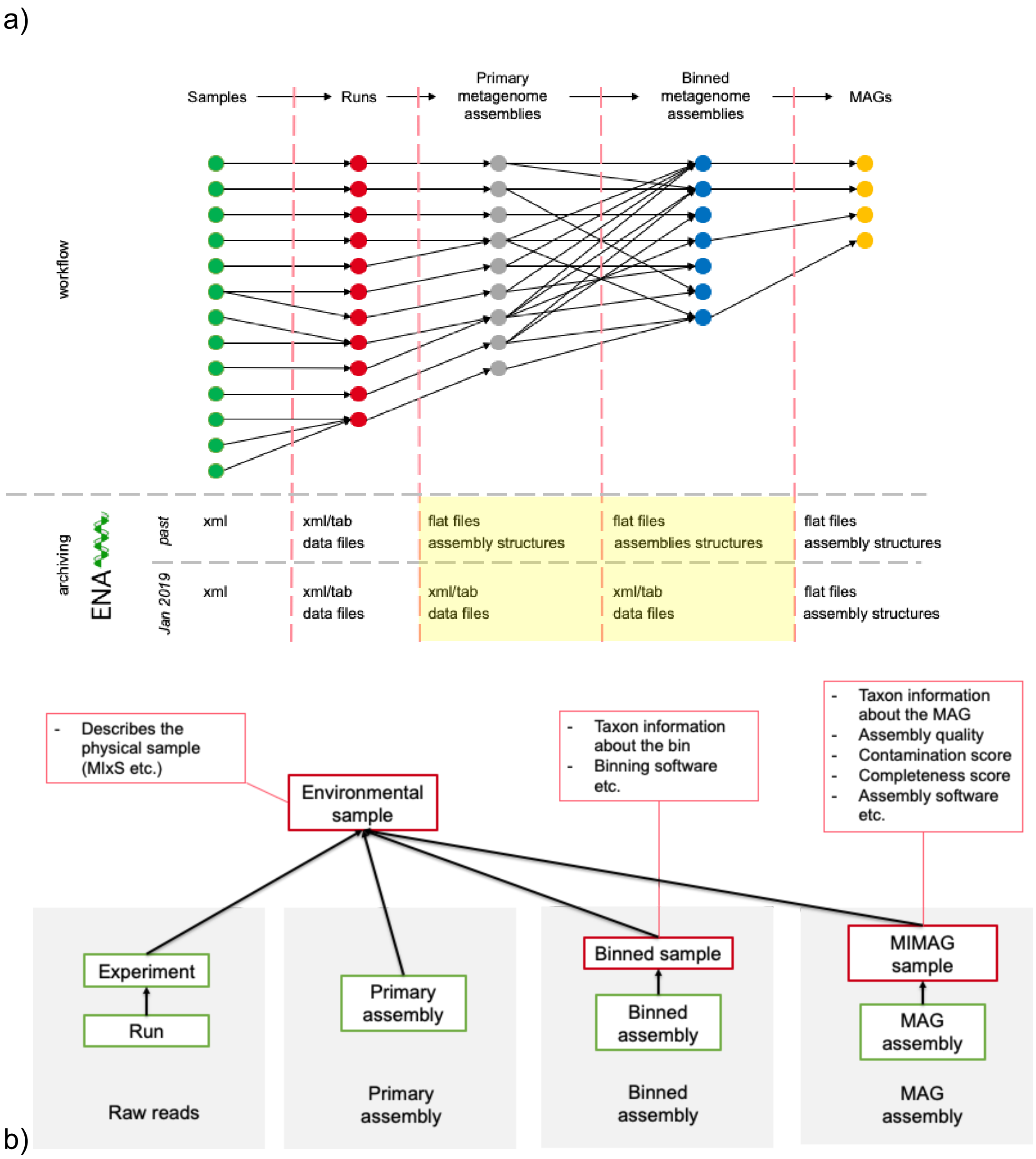


Figure 5: ENA data model for assemblies showing a) analytical workflow-based relationships between objects and data formats used in ENA submission and b) metadata fields across different tiers.

Data are stored and presented from ENA differently according to tier. Primary metagenome assemblies and binned metagenome assemblies are handled in a new data structure, specifically a class of the ENA Analysis object, that provides a simple metadata record with pointer to fasta data file. Records in the MAG tier are presented using conventional sequence flat files.

Data submissions

The ENA data submissions systems have been updated to accommodate the three tiers. This has involved work on two of the ENA data submission applications, command-line interface (Webin-CLI[4]) and RESTful interface (Webin-REST[5]). Work carried out on these systems includes integration of new MIMAGS-compliant sample checklists specifically built for metagenome assembly data and support for analysis object-based submission across all three tiers with indicators to show tier assignment. A user workflow is shown in Figure 6 and documented on Webin-CLI[6].

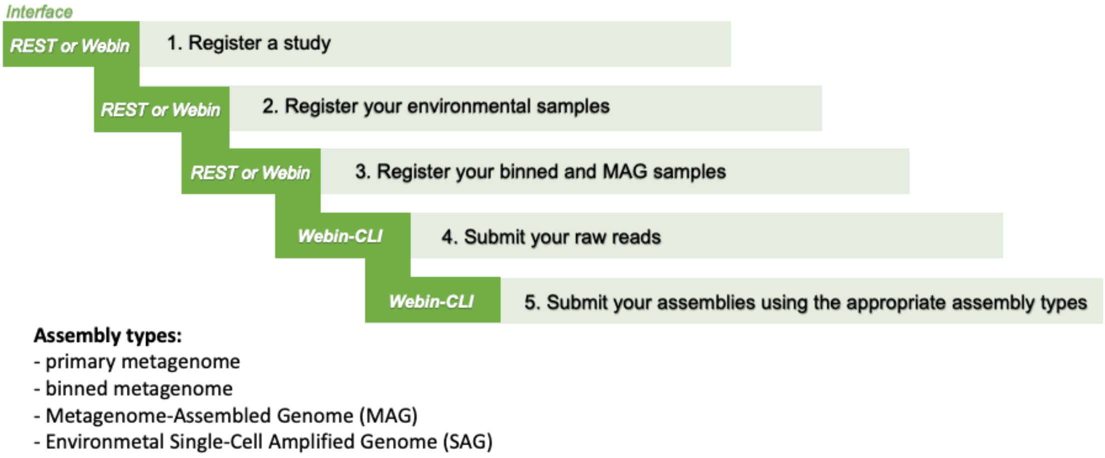

Interface
REST or Webin
1. Register a study
REST or Webin
2. Register your environmental samples
REST or Webin
3. Register your binned and MAG samples
Webin-CLI
4. Submit your raw reads
Webin-CLI
5. Submit your assemblies using the appropriate assembly types
Assembly types:
- primary metagenome
- binned metagenome
- Metagenome-Assembled Genome (MAG)
- Environmetal Single-Cell Amplified Genome (SAG)

Figure 6: Workflow for submission of metagenome assembly data to ENA.

Data access

Data in the primary metagenome assembly and binned metagenome assembly tiers are presented as analysis objects (available by accession, e.g. https://www.ebi.ac.uk/ena/data/view/ERZ829069 and searchable under the “Analysis” domain of the ENA Advanced Search[7] and the analysis-related results from the ENA Discovery API[8]. Data in the MAG tier are presented as assemblies and sets of contigs (available by accession, e.g. https://www.ebi.ac.uk/ena/data/view/GCA_900538285, and under the “Assembly” and “Contig set” domains of ENA Advanced Search and assembly and wgs_set results from the ENA Discovery API).

Content

Current content spans 15,079 data sets in the primary metagenome assembly tier, 53,890 in the binned metagenome assembly and 4,343 in the MAG tier. The vast majority of this data has been submitted by the MGnify team.

INSDC

While data in the MAG tier continue to be exchanged across INSDC, discussions are ongoing relating to exchange of data in the primary metagenome assembly and binned metagenome assembly tiers. These will continue during the May 2019 annual technical meeting of the INSDC. Further discussions will be held at this meeting as to how the model for MAG representation might evolve as in most cases, the MAG tier record is a

[4] https://ena-docs.readthedocs.io/en/latest/cli_01.html
[5] https://ena-docs.readthedocs.io/en/latest/prog_01.html
[6] https://ena-docs.readthedocs.io/en/latest/cli_07.html
[7] https://www.ebi.ac.uk/ena/data/warehouse/search
[8] https://www.ebi.ac.uk/ena/portal/api/

redundant replicate of the binned metagenome assembly tier record; ultimately a lighter-weight metadata record indicating that a binned metagenome assembly also has MAG status may be more effective.

8.4. Conclusion and Future Plans

The work described in this deliverable report highlights some of the strengths and weakness of the current standards in the community for assessing genome quality. While the MIMAGs and MISAGs represent an important first step along the pathway, it is clear that there is both a lot of legacy data of varying quality, that MAGs and SAGs are of lower quality, comparing to marker genes is insufficient to estimate quality alone. Significantly, higher levels of completeness MAGs generally represent closer matches to genomes, when there are low levels of contamination. Within the medium quality, we suggest using the QS50 metric, do identify better quality MAGs.

Ideally, other parameters should be included such as assembly length, coverage, number of replicons, number of contigs, presence of 5S, 16S and 23S rRNA and number of tRNAs. Also, a better description of how the assembly, binning and refinement was performed is also needed to evaluate the quality (e.g. co-assembly). However, as noted previously, rRNAs are often not found in MAGs, meaning that MAGs are rarely likely to meet the MIMAGs high quality standard, even when near complete is little or no detectable contamination. The lack of rRNAs is particularly prevalent in diverse communities, such as marine. Of the total 11499 genome entries assessed in the MAR databases, only 738 or 6.5% can be regarded as finished and complete, less than compared to genomes from other sources.

In the near future (before the end of EXCELERATE), we will upload all of the MGnify MAGs to the appropriate levels in ENA, making use of the new submission interfaces. We will report completeness and contamination, and use the MIMAGs checklist. From here, they can be imported in the the MAR databases.

From the analysis of the MAR databases, it is clear that genome assemblies from WGS sequencing give the longest genome assemblies. The genome assemblies generated from MAGs are approximately half of the size and SAGs only a quarter of the size, compared to WGSs. While some of this is undoubtedly experimental error, it is also important to realise that most genomes from Candidate Phyla Radiation are significantly smaller (14), lacking many of the standard housekeeping genes (e.g. amino acid biosynthesis).

The estimated average completeness and contamination varies between the different technologies. The WGSs, on average gives, by far, the highest completeness, while SAGs shows low completeness, but also the lowest degree of contamination. The completeness of the MAGs is on average between the WGSs and SAGs, while contamination in the same range as for WGSs. Although, low completeness of the SAGs, they are very valuable e.g. in increasing the precision in the taxonomic classification of marine samples, but less useful for analysis of pathways and networks. The major concern is, however, the high degree of contamination in some of the genomes present in public databases, which may lead to the incorrect taxonomic classification of reads and assemblies. Many of the genomes, with a low QS and high degree of contamination, should probably be more thoroughly examined and those that are clearly highly contaminated removed in order to increase the precision of taxonomic classification and/or functional assignment.

There is a need to establish an “best practices” for scientists generating microbial genomes to ensure that all required metadata to access the quality is present before the deposition to public databases. The MIMAGs and MISAGs standards will improve this, and the new developments within ENA provide better support for the capture of such information. Furthermore, public databases presenting any form of genome should provide a better description of the parameters used to assess the quality of the genome, thereby allowing users to select genomes of appropriate quality for use in their research.

Furthermore, the community is currently highly dependent of the use of CheckM for the use of assessing completeness and annotation. This uses sets of single copy marker genes based on isolate genomes. This has two potential issues: (i) the use of the mark gene models are from Pfam, while representing the most sensitive form of search today, the models will be bias in their sensitivity; (ii) the gene sets are based on isolate genomes, so as we explore new areas, these sets may not appropriately reflect completeness/contamination. Thus, we would recommend that research routinely also compare to genome reference database using tools such as Mash (15, 16) as an additional method for indicating completeness/contamination. A final problem is that CheckM only deals with bacteria (archaea and eubacteria), so is not suitable for assessing eukaryotic or viral genomes. While BUSCO (15) offers a similar approach as CheckM for eukaryotes, the gene sets are very bias toward the reference databases. Assessment of completeness and contamination for eukaryotes and viruses remains a largely unsolved problem, which needs to solved to evaluate genomic assemblies from the marine microbiota.

8.5. References

1. Land,M., Hauser,L., Jun,S.-R., Nookaew,I., Leuze,M.R., Ahn,T.-H., Karpinets,T., Lund,O., Kora,G., Wassenaar,T., *et al.* (2015) Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics*, **15**, 141–161.

2. Creevey,C.J., Doerks,T., Fitzpatrick,D.A., Raes,J. and Bork,P. (2011) Universally distributed single-copy genes indicate a constant rate of horizontal transfer. *PLoS One*, **6**, e22099.

3. 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.

4. Bowers,R.M., Kyrpides,N.C., Stepanauskas,R., Harmon-Smith,M., Doud,D., Reddy,T.B.K., Schulz,F., Jarett,J., Rivers,A.R., Eloe-Fadrosh,E.A., *et al.* (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.*, **35**, 725–731.

5. Parks,D.H., Rinke,C., Chuvochina,M., Chaumeil,P.-A., Woodcroft,B.J., Evans,P.N., Hugenholtz,P. and Tyson,G.W. (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat*

Microbiol, **2**, 1533–1542.

6. Martijn,J., Vosseberg,J., Guy,L., Offre,P. and Ettema,T.J.G. (2018) Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature*, **557**, 101–105.

7. Gruber-Vodicka,H.R., Seah,B.K.B. and Pruesse,E. (2019) phyloFlash — Rapid SSU rRNA profiling and targeted assembly from metagenomes: Supplementary Information. *Bioinformatics*.

8. 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.

9. Eren,A.M., Esen,Ö.C., Quince,C., Vineis,J.H., Morrison,H.G., Sogin,M.L. and Delmont,T.O. (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, **3**, e1319.

10. Klemetsen,T., Raknes,I.A., Fu,J., Agafonov,A., Balasundaram,S.V., Tartari,G., Robertsen,E. and Willassen,N.P. (2018) The MAR databases: development and implementation of databases specific for marine metagenomics. *Nucleic Acids Res.*, **46**, D692–D699.

11. Haft,D.H., DiCuccio,M., Badretdin,A., Brover,V., Chetvernin,V., O'Neill,K., Li,W., Chitsaz,F., Derbyshire,M.K., Gonzales,N.R., *et al.* (2018) RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.*, **46**, D851–D860.

12. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.

13. Li,M., Baker,B.J., Anantharaman,K., Jain,S., Breier,J.A. and Dick,G.J. (2015) Genomic and transcriptomic evidence for scavenging of diverse organic compounds by widespread deep-sea archaea. *Nat. Commun.*, **6**, 8933.

14. Brown,C.T., Hug,L.A., Thomas,B.C., Sharon,I., Castelle,C.J., Singh,A., Wilkins,M.J., Wrighton,K.C., Williams,K.H. and Banfield,J.F. (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, **523**, 208–211.

15. Simão,F.A., Waterhouse,R.M., Ioannidis,P., Kriventseva,E.V. and Zdobnov,E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

16. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.